IBM SPSS Modeler 16
Modellierungsknoten

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 299 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 16, Release 0, Modifikation 0 von IBM(r) SPSS(r) Modeler und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Modeler CRISP 15, Modeling Nodes,
herausgegeben von International Business Machines Corporation, USA

Herausgegeben von:
TSC Germany
Kst. 2877
Oktober 2013

Inhaltsverzeichnis

Vorwort

IBM® SPSS Modeler ist die auf Unternehmensebene einsetzbare Data-Mining-Workbench von IBM. Mit SPSS Modeler können Unternehmen und Organisationen die Beziehungen zu ihren Kunden bzw. zu den Bürgern durch ein tief greifendes Verständnis der Daten verbessern. Organisationen verwenden die mithilfe von SPSS Modeler gewonnenen Erkenntnisse zur Bindung profitabler Kunden, zur Ermittlung von Cross-Selling-Möglichkeiten, zur Gewinnung neuer Kunden, zur Ermittlung von Betrugsfällen, zur Reduzierung von Risiken und zur Verbesserung der Verfügbarkeit öffentlicher Dienstleistungen.

Die visuelle Benutzerschnittstelle von SPSS Modeler erleichtert Benutzern die Anwendung des spezifischen Fachwissens, was zu leistungsfähigeren Vorhersagemodellen führt und die Zeit bis zur Lösungsfindung verkürzt. SPSS Modeler bietet viele Modellierungsverfahren, wie beispielsweise Algorithmen für Vorhersage, Klassifizierung, Segmentierung und Assoziationserkennung. Nach der Modellerstellung ermöglicht IBM SPSS Modeler Solution Publisher die unternehmensweite Bereitstellung des Modells für Entscheidungsträger oder in einer Datenbank.

Informationen zu IBM Business Analytics

Die Software IBM Business Analytics liefert umfassende, einheitliche und korrekte Informationen, mit denen Entscheidungsträger die Unternehmensleistung verbessern können. Ein umfassendes Portfolio aus Anwendungen für Business Intelligence, Vorhersageanalyse, Finanz- und Strategiemanagement sowie Analysen bietet Ihnen sofort klare und umsetzbare Einblicke in die aktuelle Leistung und gibt Ihnen die Möglichkeit, zukünftige Ergebnisse vorherzusagen. Durch umfassende Branchenlösungen, bewährte Vorgehensweisen und professionellen Service können Unternehmen jeder Größe die Produktivität maximieren, Entscheidungen automatisieren und bessere Ergebnisse erzielen.

Als Teil dieses Portfolios unterstützt IBM SPSS Predictive Analytics-Software Unternehmen dabei, zukünftige Ereignisse vorherzusagen und proaktiv Maßnahmen zu ergreifen, um bessere Geschäftsergebnisse zu erzielen. Kunden aus Wirtschaft, öffentlichem Dienst und staatlichen Lehr- und Forschungseinrichtungen weltweit nutzen IBM SPSS-Technologie als Wettbewerbsvorteil für die Kundengewinnung, Kundenbindung und Erhöhung der Kundenumsätze bei gleichzeitiger Eindämmung der Betrugsmöglichkeiten und Minderung von Risiken. Durch die Einbindung von IBM SPSS-Software in ihre täglichen Operationen wandeln sich Organisationen zu "Predictive Enterprises", die Entscheidungen auf Geschäftsziele ausrichten und automatisieren und einen messbaren Wettbewerbsvorteil erzielen können. Wenn Sie weitere Informationen wünschen oder Kontakt zu einem Mitarbeiter aufnehmen möchten, besuchen Sie die Seite http://www.ibm.com/spss.

Technical Support

Kunden mit Wartungsvertrag können den Technical Support in Anspruch nehmen. Kunden können sich an den Technical Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Produkten oder bei der Installation in einer der unterstützten Hardwareumgebungen benötigen. Zur Kontaktaufnahme mit dem Technical Support besuchen Sie die IBM Website unter http://www.ibm.com/support. Sie müssen bei der Kontaktaufnahme Ihren Namen, Ihre Organisation und Ihre Supportvereinbarung angeben.

Kapitel 1. Informationen zu IBM SPSS Modeler

IBM SPSS Modeler ist ein Set von Data-Mining-Tools, mit dem Sie auf der Grundlage Ihres Fachwissens schnell und einfach Vorhersagemodelle erstellen und zur Erleichterung der Entscheidungsfindung in die Betriebsabläufe einbinden können. Das Produkt IBM SPSS Modeler, das auf der Grundlage des den Industrienormen entsprechenden Modells CRISP-DM entwickelt wurde, unterstützt den gesamten Data Mining-Prozess, von den Daten bis hin zu besseren Geschäftsergebnissen.

IBM SPSS Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen. Mit den in der Modellierungspalette verfügbaren Methoden können Sie aus Ihren Daten neue Informationen ableiten und Vorhersagemodelle erstellen. Jede Methode hat ihre speziellen Stärken und eignet sich besonders für bestimmte Problemtypen.

SPSS Modeler kann als Standalone-Produkt oder als Client in Verbindung mit SPSS Modeler Server erworben werden. Außerdem ist eine Reihe von Zusatzoptionen verfügbar, die in den folgenden Abschnitten kurz zusammengefasst werden. Weitere Informationen finden Sie unter http://www.ibm.com/software/analytics/spss/products/modeler/.

IBM SPSS Modeler-Produkte

Zur IBM SPSS Modeler-Produktfamilie und der zugehörigen Software gehören folgende Elemente.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server-Adapter für IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler ist eine funktionell in sich abgeschlossene Produktversion, die Sie auf Ihrem PC installieren und ausführen können. Sie können SPSS Modeler im lokalen Modus als Standalone-Produkt oder im verteilten Modus zusammen mit IBM SPSS Modeler Server verwenden, um bei Datasets die Leistung zu verbessern.

Mit SPSS Modeler können Sie schnell und intuitiv genaue Vorhersagemodelle erstellen, und das ohne Programmierung. Mithilfe der speziellen visuellen Benutzerschnittstelle können Sie den Data Mining-Prozess auf einfache Weise visualisieren. Mit der Unterstützung der in das Produkt eingebetteten erweiterten Analyseprozesse können Sie zuvor verborgene Muster und Trends in Ihren Daten aufdecken. Sie können Ergebnisse modellieren und Einblick in die Faktoren gewinnen, die Einfluss auf diese Ergebnisse haben, wodurch Sie in die Lage versetzt werden, Geschäftschancen zu nutzen und Risiken zu mindern.

SPSS Modeler ist in zwei Editionen erhältlich: SPSS Modeler Professional und SPSS Modeler Premium. Weitere Informationen finden Sie unter im Thema „IBM SPSS Modeler-Editionen" auf Seite 2.

IBM SPSS Modeler Server

SPSS Modeler verwendet eine Client/Server-Architektur zur Verteilung von Anforderungen für ressourcenintensive Vorgänge an leistungsstarke Serversoftware, wodurch bei größeren Datasets eine höhere Leistung erzielt werden kann.

SPSS Modeler Server ist ein separat lizenziertes Produkt, das durchgehend im verteilten Analysemodus auf einem Server-Host in Verbindung mit einer oder mehreren IBM SPSS Modeler-Installationen ausgeführt wird. Auf diese Weise bietet SPSS Modeler Server eine herausragende Leistung bei großen Datasets, da speicherintensive Vorgänge auf dem Server ausgeführt werden können, ohne Daten auf den Client-Computer herunterladen zu müssen. IBM SPSS Modeler Server bietet außerdem Unterstützung für SQL-Optimierung sowie Möglichkeiten zur Modellierung innerhalb der Datenbank, was weitere Vorteile hinsichtlich Leistung und Automatisierung mit sich bringt.

IBM SPSS Modeler Administration Console

Modeler Administration Console ist eine grafische Anwendung zur Verwaltung einer Vielzahl der SPSS Modeler Server-Konfigurationsoptionen, die auch mithilfe einer Optionsdatei konfiguriert werden können. Die Anwendung bietet eine Konsolenbenutzerschnittstelle zur Überwachung und Konfiguration der SPSS Modeler Server-Installationen und steht aktuellen SPSS Modeler Server-Kunden kostenlos zur Verfügung. Die Anwendung kann nur unter Windows installiert werden. Der von ihr verwaltete Server kann jedoch auf einer beliebigen unterstützten Plattform installiert sein.

IBM SPSS Modeler Batch

Das Data-Mining ist zwar in der Regel ein interaktiver Vorgang, es ist jedoch auch möglich, SPSS Modeler über eine Befehlszeile auszuführen, ohne dass die grafische Benutzerschnittstelle verwendet werden muss. Beispielsweise kann es sinnvoll sein, langwierige oder sich wiederholende Aufgaben ohne Eingreifen des Benutzers durchzuführen. SPSS Modeler Batch ist eine spezielle Version des Produkts, die die vollständigen Analysefunktionen von SPSS Modeler ohne Zugriff auf die reguläre Benutzerschnittstelle bietet. Zur Verwendung von SPSS Modeler Batch ist eine SPSS Modeler Server-Lizenz erforderlich.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher ist ein Tool, mit dem Sie eine gepackte Version eines SPSS Modeler-Streams erstellen können, der durch eine externe Runtime-Engine ausgeführt oder in eine externe Anwendung eingebettet werden kann. Auf diese Weise können Sie vollständige SPSS Modeler-Streams für die Verwendung in Umgebungen veröffentlichen und bereitstellen, in denen SPSS Modeler nicht installiert ist. SPSS Modeler Solution Publisher wird als Teil des Diensts für IBM SPSS Collaboration and Deployment Services - Scoring verteilt, für den eine separate Lizenz erforderlich ist. Mit dieser Lizenz erhalten Sie SPSS Modeler Solution Publisher Runtime, womit Sie die veröffentlichten Streams ausführen können.

IBM SPSS Modeler Server-Adapter für IBM SPSS Collaboration and Deployment Services

Für IBM SPSS Collaboration and Deployment Services ist eine Reihe von Adaptern verfügbar, mit denen SPSS Modeler und SPSS Modeler Server mit einem IBM SPSS Collaboration and Deployment Services-Repository interagieren können. Auf diese Weise kann ein im Repository bereitgestellter SPSS Modeler-Stream von mehreren Benutzern gemeinsam verwendet werden. Auch der Zugriff über die Thin-Client-Anwendung IBM SPSS Modeler Advantage ist möglich. Sie installieren den Adapter auf dem System, das als Host für das Repository fungiert.

IBM SPSS Modeler-Editionen

SPSS Modeler ist in den folgenden Editionen erhältlich.

SPSS Modeler Professional

SPSS Modeler Professional bietet sämtliche Tools, die Sie für die Arbeit mit den meisten Typen von strukturierten Daten benötigen, beispielsweise in CRM-Systemen erfasste Verhaltensweisen und Interaktionen, demografische Daten, Kaufverhalten und Umsatzdaten.

SPSS Modeler Premium

SPSS Modeler Premium ist ein separat lizenziertes Produkt, das SPSS Modeler Professional für die Arbeit mit spezialisierten Daten, wie beispielsweise Daten, die für Entitätsanalysen oder soziale Netze verwendet werden, sowie für die Arbeit mit unstrukturierten Textdaten erweitert. SPSS Modeler Premium umfasst die folgenden Komponenten.

IBM SPSS Modeler Entity Analytics fügt den IBM SPSS Modeler-Vorhersageanalysen eine weitere Dimension hinzu. Während bei Vorhersageanalysen versucht wird, zukünftiges Verhalten aus früheren Daten vorherzusagen, liegt der Schwerpunkt bei der Entitätsanalyse auf der Verbesserung von Kohärenz und Konsistenz der aktuellen Daten, indem Identitätskonflikte innerhalb der Datensätze selbst aufgelöst werden. Bei der Identität kann es sich um die Identität einer Person, einer Organisation, eines Objekts oder einer anderen Entität handeln, bei der Unklarheiten bestehen könnten. Die Identitätsauflösung kann in einer Reihe von Bereichen entscheidend sein, darunter Customer Relationship Management, Betrugserkennung, Bekämpfung der Geldwäsche sowie nationale und internationale Sicherheit.

IBM SPSS Modeler Social Network Analysis transformiert Informationen zu Beziehungen in Felder, die das Sozialverhalten von Einzelpersonen und Gruppen charakterisieren. Durch die Verwendung von Daten, die die Beziehungen beschreiben, die sozialen Netzen zugrunde liegen, ermittelt IBM SPSS Modeler Social Network Analysis Führungskräfte in sozialen Netzen, die das Verhalten anderer Personen im Netz beeinflussen. Außerdem können Sie feststellen, welche Personen am meisten durch andere Teilnehmer im Netz beeinflusst werden. Durch die Kombination dieser Ergebnisse mit anderen Maßen können Sie aussagekräftige Profile für Einzelpersonen erstellen, die Sie als Grundlage für Ihre Vorhersagemodelle verwenden können. Modelle, die diese sozialen Informationen berücksichtigen, sind leistungsstärker als Modelle, die dies nicht tun.

IBM SPSS Modeler Text Analytics verwendet hoch entwickelte linguistische Technologien und die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP), um eine schnelle Verarbeitung einer großen Vielfalt an unstrukturierten Textdaten zu ermöglichen, um die Schlüsselkonzepte zu extrahieren und zu ordnen und um diese Konzepte in Kategorien zusammenzufassen. Extrahierte Konzepte und Kategorien können mit bestehenden strukturierten Daten, beispielsweise demografischen Informationen, kombiniert und mithilfe der vollständigen Suite der Data-Mining-Tools von IBM SPSS Modeler auf die Modellierung angewendet werden, um bessere und fokussiertere Entscheidungen zu ermöglichen.

IBM SPSS Modeler-Dokumentation

Eine Dokumentation im Onlinehilfeformat finden Sie im Hilfemenü von SPSS Modeler. Diese umfasst die Dokumentation für SPSS Modeler, SPSS Modeler Server und SPSS Modeler Solution Publisher sowie das Anwendungshandbuch und weiteres Material zur Unterstützung.

Die vollständige Dokumentation für die einzelnen Produkte (einschließlich Installationsanweisungen) steht im PDF-Format im Ordner *Documentation* auf der jeweiligen Produkt-DVD zur Verfügung. Installationsdokumente können auch aus dem Internet unter http://www-01.ibm.com/support/docview.wss?uid=swg27038316 heruntergeladen werden.

Dokumentation in beiden Formaten steht auch im SPSS Modeler Information Center unter http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/ zur Verfügung.

SPSS Modeler Professional-Dokumentation

Die SPSS Modeler Professional-Dokumentationssuite (ohne Installationsanweisungen) umfasst folgende Dokumente:

- **IBM SPSS Modeler Benutzerhandbuch.** Allgemeine Einführung in die Verwendung von SPSS Modeler, in der u. a. die Erstellung von Datenstreams, der Umgang mit fehlenden Werten, die Erstellung von CLEM-Ausdrücken, die Arbeit mit Projekten und Berichten sowie das Packen von Streams für die Bereitstellung in IBM SPSS Collaboration and Deployment Services, Predictive Applications (Vorhersageanwendungen) oder IBM SPSS Modeler Advantage beschrieben werden.

- **IBM SPSS Modeler Quellen-, Prozess- und Ausgabeknoten.** Beschreibung aller Knoten, die zum Lesen, zum Verarbeiten und zur Ausgabe von Daten in verschiedenen Formaten verwendet werden. Im Grunde sind sie alle Knoten, mit Ausnahme der Modellierungsknoten.
- **IBM SPSS Modeler Modellierungsknoten.** Beschreibungen sämtlicher für die Erstellung von Data-Mining-Modellen verwendeter Knoten. IBM SPSS Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen.
- **IBM SPSS Modeler Algorithms Guide.** Beschreibung der mathematischen Grundlagen der in IBM SPSS Modeler verwendeten Modellierungsmethoden. Dieses Handbuch steht nur im PDF-Format zur Verfügung.
- **IBM SPSS Modeler Anwendungshandbuch.** Die Beispiele in diesem Handbuch bieten eine kurze, gezielte Einführung in bestimmte Modellierungsmethoden und -verfahren. Eine Online-Version dieses Handbuchs kann auch über das Hilfemenü aufgerufen werden. Weitere Informationen finden Sie unter im Thema „Anwendungsbeispiele" auf Seite 5.
- **IBM SPSS Modeler Handbuch für Scripterstellung und Automatisierung.** Informationen zur Automatisierung des Systems über Scripterstellung, einschließlich der Eigenschaften, die zur Bearbeitung von Knoten und Streams verwendet werden können.
- IBM SPSS Modeler Bereitstellungshandbuch. Informationen zum Ausführen von IBM SPSS Modeler-Streams und -Szenarios als Schritte bei der Verarbeitung von Jobs im IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF-Entwicklerhandbuch.** CLEF bietet die Möglichkeit, Drittanbieterprogramme, wie Datenverarbeitungsroutinen oder Modellierungsalgorithmen, als Knoten in IBM SPSS Modeler zu integrieren.
- **IBM SPSS Modeler Datenbankinternes Mining.** Informationen darüber, wie Sie Ihre Datenbank dazu einsetzen, die Leistung zu verbessern, und wie Sie die Palette der Analysefunktionen über Drittanbieteralgorithmen erweitern.
- **IBM SPSS Modeler Server Verwaltungs- und Leistungshandbuch.** Informationen zur Konfiguration und Verwaltung von IBM SPSS Modeler Server.
- **IBM SPSS Modeler Administration Console Benutzerhandbuch.** Informationen zur Installation und Nutzung der Konsolenbenutzerschnittstelle zur Überwachung und Konfiguration von IBM SPSS Modeler Server. Die Konsole ist als Plug-in für die Deployment Manager-Anwendung implementiert.
- IBM SPSS Modeler CRISP-DM Handbuch. Schritt-für-Schritt-Anleitung für das Data Mining mit SPSS Modeler unter Verwendung der CRISP-DM-Methode.
- IBM SPSS Modeler Batch Benutzerhandbuch. Vollständiges Handbuch für die Verwendung von IBM SPSS Modeler im Stapelmodus, einschließlich Details zur Ausführung des Stapelmodus und zu Befehlszeilenargumenten. Dieses Handbuch steht nur im PDF-Format zur Verfügung.

SPSS Modeler Premium-Dokumentation

Die SPSS Modeler Premium-Dokumentationssuite (ohne Installationsanweisungen) umfasst folgende Dokumente:

- IBM SPSS Modeler Entity Analytics Benutzerhandbuch. Informationen zur Verwendung von Entitätsanalysen mit SPSS Modeler, unter Behandlung der Repository-Installation und -Konfiguration, Entity Analytics-Knoten und Verwaltungsaufgaben.
- **IBM SPSS Modeler Social Network Analysis User Guide.** Ein Handbuch zur Durchführung einer sozialen Netzanalyse mit SPSS Modeler, einschließlich einer Gruppenanalyse und Diffusionsanalyse.
- SPSS Modeler Text Analytics Benutzerhandbuch. Informationen zur Verwendung von Textanalysen mit SPSS Modeler, unter Behandlung der Text Mining-Knoten, der interaktiven Workbench sowie von Vorlagen und anderen Ressourcen.
- IBM SPSS Modeler Text Analytics Administration Console Benutzerhandbuch. Informationen zur Installation und Nutzung der Konsolenbenutzerschnittstelle zur Überwachung und Konfiguration von IBM SPSS Modeler Server für die Verwendung mit SPSS Modeler Text Analytics . Die Konsole ist als Plug-in für die Deployment Manager-Anwendung implementiert.

Anwendungsbeispiele

Mit den Data-Mining-Tools in SPSS Modeler kann eine große Bandbreite an geschäfts- und unternehmensbezogenen Problemen gelöst werden; die Anwendungsbeispiele dagegen bieten jeweils eine kurze, gezielte Einführung in spezielle Modellierungsmethoden und -verfahren. Die hier verwendeten Datasets sind viel kleiner als die großen Datenbestände, die von einigen Data-Mining-Experten verwaltet werden müssen, die zugrunde liegenden Konzepte und Methoden sollten sich jedoch auch auf reale Anwendungen übertragen lassen.

Sie können auf die Beispiele zugreifen, indem Sie im Menü "Hilfe" in SPSS Modeler auf die Option **Anwendungsbeispiele** klicken. Die Datendateien und Beispielstreams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Weitere Informationen finden Sie unter im Thema „Ordner "Demos"".

Beispiele für die Datenbankmodellierung. Die Beispiele finden Sie im IBM SPSS Modeler-Handbuch zum datenbankinternen Mining.

Scriptbeispiele. Die Beispiele finden Sie im *IBM SPSS Modeler Handbuch für Scripterstellung und Automatisierung*.

Ordner "Demos"

Die in den Anwendungsbeispielen verwendeten Datendateien und Beispielstreams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Auf diesen Ordner können Sie auch über die Programmgruppe IBM SPSS Modeler im Windows-Startmenü oder durch Klicken auf *Demos* in der Liste der zuletzt angezeigten Verzeichnisse im Dialogfeld "Datei öffnen" zugreifen.

Kapitel 2. Einführung in die Modellierung

Ein Modell ist eine Menge von Regeln, Formeln bzw. Gleichungen, mit der ein Ergebnis auf der Grundlage einer Menge von Eingabefeldern bzw. -variablen vorhergesagt werden kann. Ein Finanzinstitut verwendet z. B. möglicherweise ein Modell, um auf Basis von bekannten Informationen über vorherige Kreditantragsteller vorherzusagen, ob ein Kreditantragsteller ein geringes oder hohes Risiko darstellt.

Die Möglichkeit zur Vorhersage eines Ergebnisses ist das zentrale Ziel von Vorhersageanalysen und ein Verständnis des Modellierungsprozesses ist der Schlüssel für die Verwendung von IBM SPSS Modeler.

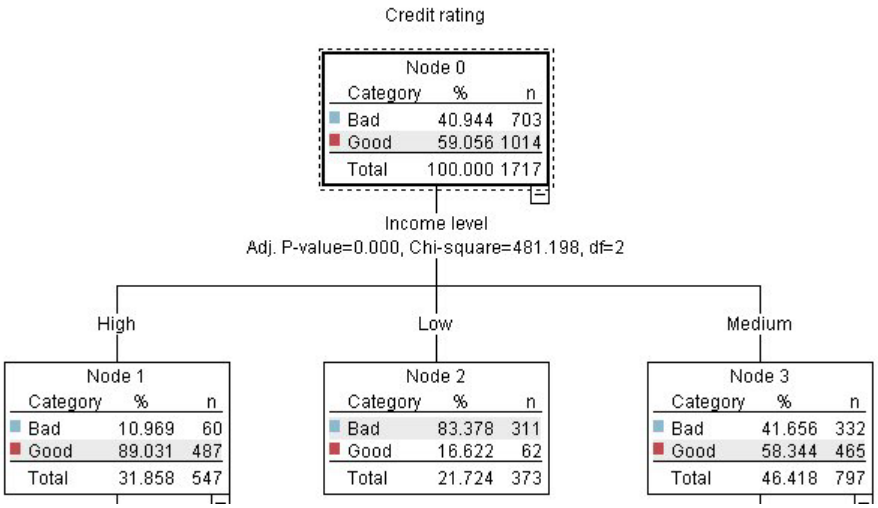


Abbildung 1. Ein einfaches Entscheidungsbaummodell

In diesem Beispiel wird ein **Entscheidungsbaummodell** verwendet, das Datensätze aufzeichnet (und eine Reaktion vorhersagt), wobei eine Reihe von Entscheidungsregeln verwendet wird. Beispiel:

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

In diesem Beispiel wird zwar ein Modell vom Typ "CHAID" (Chi-squared Automatic Interaction Detection) verwendet, es ist jedoch als allgemeine Einführung gedacht und die meisten Konzepte gelten im Wesentlichen auch für andere Modellierungstypen in IBM SPSS Modeler.

Um ein Modell zu verstehen, müssen Sie zunächst ein Verständnis für die darin verwendeten Daten entwickeln. Die Daten in diesem Beispiel enthalten Informationen über die Kunden einer Bank. Es werden folgende Felder verwendet:

Feldname	**Beschreibung**
Credit_rating	Kreditrating: 0 = Schlecht, 1 = Gut, 9 = fehlende Werte
Alter	Alter in Jahren
Einkommen	Einkommen in Kategorien: 1 = Niedrig, 2 = Mittel, 3 = Hoch
Credit_cards	Anzahl der Kreditkarten: 1 = Weniger als fünf, 2 = Fünf oder mehr
Bildung	Bildungsniveau: 1 = Hauptschulabschluss, 2 = Hochschulabschluss
Car_loans	Anzahl der Autokredite: 1 = Keine oder einen, 2 = Mehr als zwei

Die Bank führt eine Datenbank historischer Informationen über Kunden, die bei der Bank Kredite in Anspruch genommen haben, in der auch festgehalten wird, ob ein Kredit zurückgezahlt wurde (Bonität = Gut) oder nicht (Bonität = Schlecht). Mithilfe dieser vorhandenen Daten will die Bank ein Modell erstellen, das vorhersagen kann, mit welcher Wahrscheinlichkeit zukünftige Kreditantragsteller ihren Kreditverpflichtungen nicht nachkommen.

Anhand eines Entscheidungsbaummodells können Sie die Charakteristiken der beiden Kundengruppen analysieren und die Wahrscheinlichkeit von Kreditausfällen vorhersagen.

Für dieses Beispiel wird der Stream *modelingintro.str* verwendet, der im Ordner *Demos* unter dem Unterordner *streams* verfügbar ist. Die Datendatei ist *tree_credit.sav*. Weitere Informationen finden Sie unter im Thema „Ordner "Demos"" auf Seite 5.

Werfen wir nun einen Blick auf den Stream.

1. Wählen Sie im Hauptmenü folgende Menüoption:
   Datei > **Stream öffnen**
2. Klicken Sie auf der Symbolleiste des Dialogfelds "Öffnen" auf das Gold-Nugget-Symbol und wählen Sie den Ordner "Demos" aus.
3. Doppelklicken Sie auf den Ordner *streams*.
4. Doppelklicken Sie auf die Datei *modelingintro.str*.

Erstellen des Streams

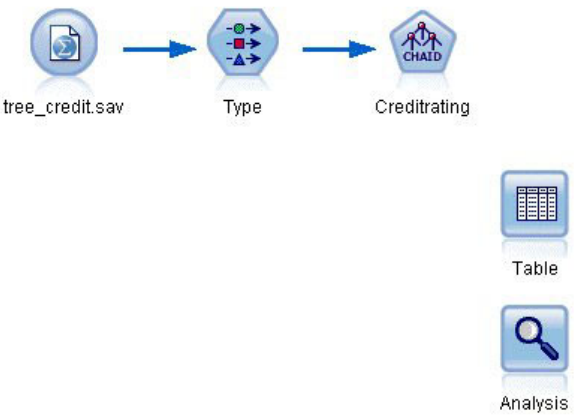


Abbildung 2. Modellierungsstream

Um einen Stream zum Erzeugen eines Modells zu erstellen, sind mindestens die folgenden drei Elemente erforderlich:

- Ein Quellenknoten, der Daten aus einer externen Quelle einliest, in diesem Fall eine IBM SPSS Statistics-Datendatei.
- Ein Quellen- oder Typknoten, der Feldeigenschaften wie das Messniveau (die Daten, die das Feld enthält) und die Rolle der einzelnen Felder als Ziel oder Eingabe in der Modellierung angibt.
- Ein Modellierungsknoten, der bei Ausführung des Streams ein Modellnugget erstellt.

In diesem Beispiel verwenden wir einen CHAID-Modellierungsknoten. CHAID (Chi-squared Automatic Interaction Detection) ist eine Klassifizierungsmethode für die Erstellung von Entscheidungsbäumen mit bestimmten Statistiktypen namens Chi-Quadrat-Statistiken zur Identifizierung der optimalen Splits.

Wenn Messniveaus im Quellenknoten angegeben sind, kann auf den separaten Typknoten verzichtet werden. Hinsichtlich der Funktion ist das Ergebnis dasselbe.

Dieser Stream weist außerdem Tabellen- und Analyseknoten auf, mit denen die Scoring-Ergebnisse angezeigt werden, nachdem das Modellnugget erstellt und in den Stream aufgenommen wurde.

Der Quellenknoten für Statistikdateien liest Daten im IBM SPSS Statistics-Format aus der Datendatei *tree_credit.sav* ein, die im Ordner *Demos* installiert wurde. (Eine spezielle Variable mit der Bezeichnung *$CLEO_DEMOS* dient zur Referenzierung dieses Ordners in der aktuellen IBM SPSS Modeler-Installation. Dadurch wird sichergestellt, dass der Pfad gültig ist, unabhängig vom aktuellen Installationsordner bzw. der jeweiligen Version.)

Abbildung 3. Einlesen von Daten mit einem Quellenknoten für Statistikdateien

Der Typknoten gibt das **Messniveau** für die einzelnen Felder an. Das Messniveau ist eine Kategorie, die den Datentyp für das Feld anzeigt. Unsere Quellendatendatei verwendet drei verschiedene Messniveaus.

Ein Feld des Typs **Stetig** (z. B. das Feld *Alter*) enthält stetige numerische Werte, während ein Feld des Typs **Nominal** (z. B. das Feld *Kreditrating*) zwei oder mehr bestimmte Werte enthält, z. B. *Schlecht*, *Gut* oder *Keine früheren Schulden*. Ein Feld des Typs **Ordinal** (z. B. *Einkommen in Kategorien*) beschreibt Daten mit mehreren unterschiedlichen Werten, die eine natürliche Reihenfolge aufweisen - in diesem Fall *Niedrig*, *Mittel* und *Hoch*.

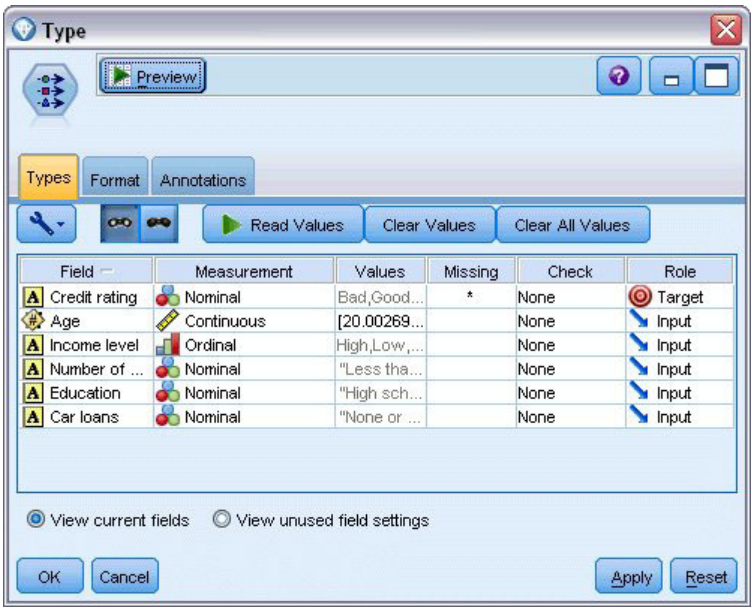


Abbildung 4. Festlegen des Ziels und der Eingabefelder mit dem Typknoten

Der Typknoten legt für jedes Feld außerdem die **Rolle** fest, die jedes Feld bei der Modellierung spielt. Für das Feld *Kreditrating*, das angibt, ob ein bestimmter Kunde seinen Kreditverpflichtungen nicht nachgekommen ist, ist die Rolle als *Ziel* festgelegt. Hierbei handelt es sich also um das **Ziel** oder das Feld, für das wir den Wert vorhersagen möchten.

Für die anderen Felder ist die Rolle auf *Eingabe* eingestellt. Eingabefelder werden manchmal auch als **Prädiktoren** bezeichnet oder als Felder, mit deren Werten der Modellierungsalgorithmus den Wert des Zielfelds vorhersagt.

Der CHAID-Modellierungsknoten generiert das Modell.

Auf der Registerkarte "Felder" im Modellierungsknoten wird die Option **Vordefinierte Rollen verwenden** ausgewählt. Dies bedeutet, dass die im Typknoten angegebenen Ziele und Eingaben verwendet werden sollen. Wir können die Feldrollen hier ändern, doch in diesem Beispiel belassen wir sie unverändert.

1. Klicken Sie auf die Registerkarte "Erstellungsoptionen".

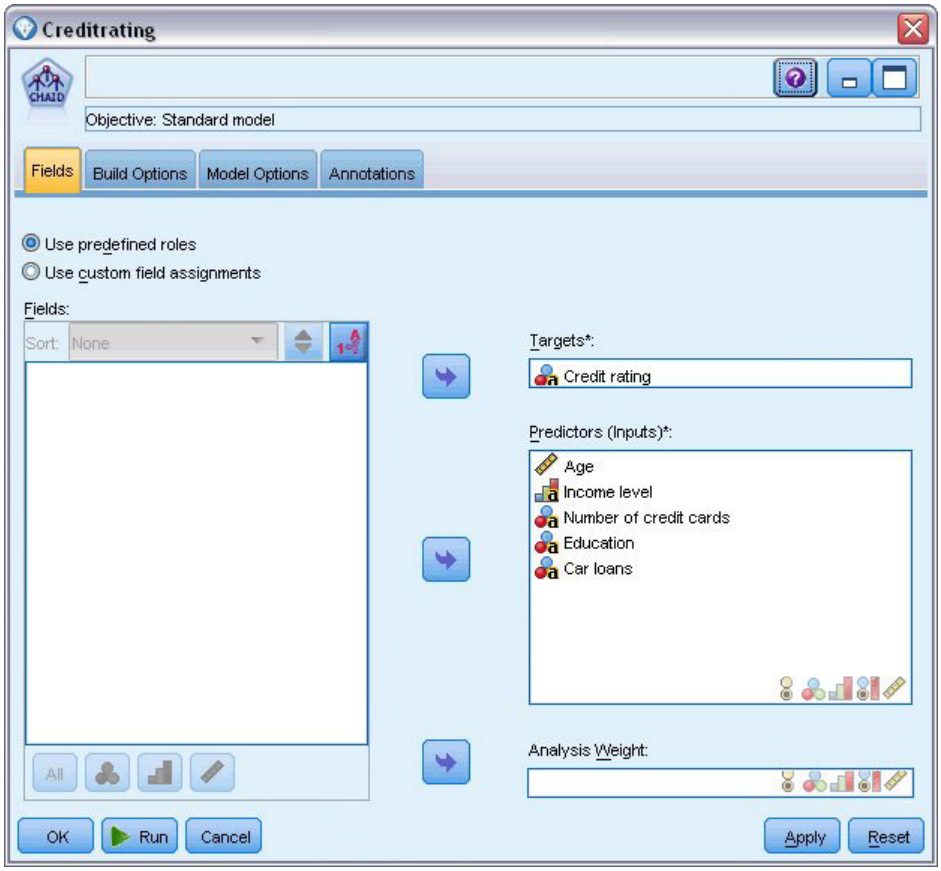


Abbildung 5. CHAID-Modellierungsknoten, Registerkarte "Felder"

Hier finden Sie einige Optionen, über die Sie die Art des aufzubauenden Modells festlegen können.

Da wir ein komplett neues Modell möchten, verwenden wir die Standardoption **Neues Modell aufbauen**.

Außerdem möchten wir nur ein einzelnes Standardentscheidungsbaummodell ohne Erweiterungen, weshalb wir auf die Standardzieloption **Einzelnen Baum aufbauen** zurückgreifen.

Sie können optional eine interaktive Modellierungssitzung starten, mit der Sie eine Feinabstimmung des Modells vornehmen können. Im vorliegenden Beispiel wird jedoch einfach ein Modell mit der Standardmoduseinstellung **Modell erzeugen** generiert.

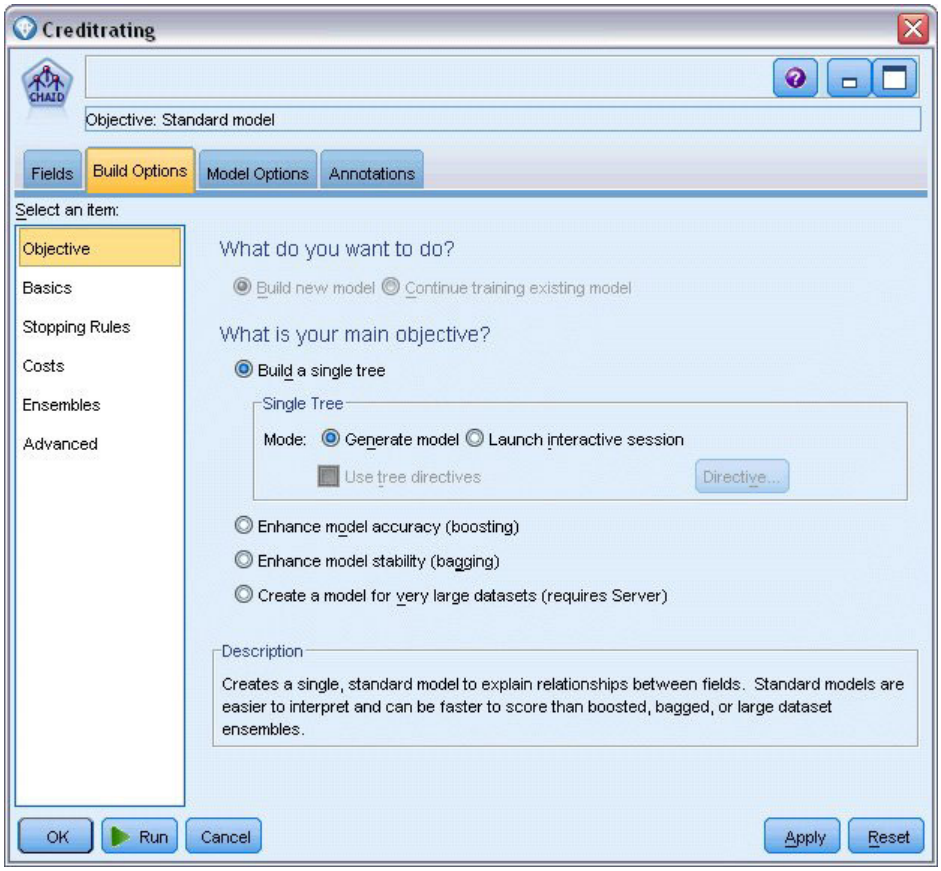


Abbildung 6. CHAID-Modellierungsknoten, Registerkarte "Erstellungsoptionen"

Für dieses Beispiel möchten wir einen einfach strukturierten Baum verwenden und begrenzen deshalb die Baumerweiterung, indem wir die minimale Anzahl der Fälle für über- und untergeordnete Knoten erhöhen.

2. Wählen Sie auf der Registerkarte "Erstellungsoptionen" im linken Navigationsbereich **Stoppregeln** aus.
3. Wählen Sie die Option **Absolutwert verwenden** aus.
4. Legen Sie für **Mindestanzahl der Datensätze in übergeordneter Verzweigung** den Wert 400 fest.
5. Legen Sie für **Mindestanzahl der Datensätze in untergeordneter Verzweigung** den Wert 200 fest.

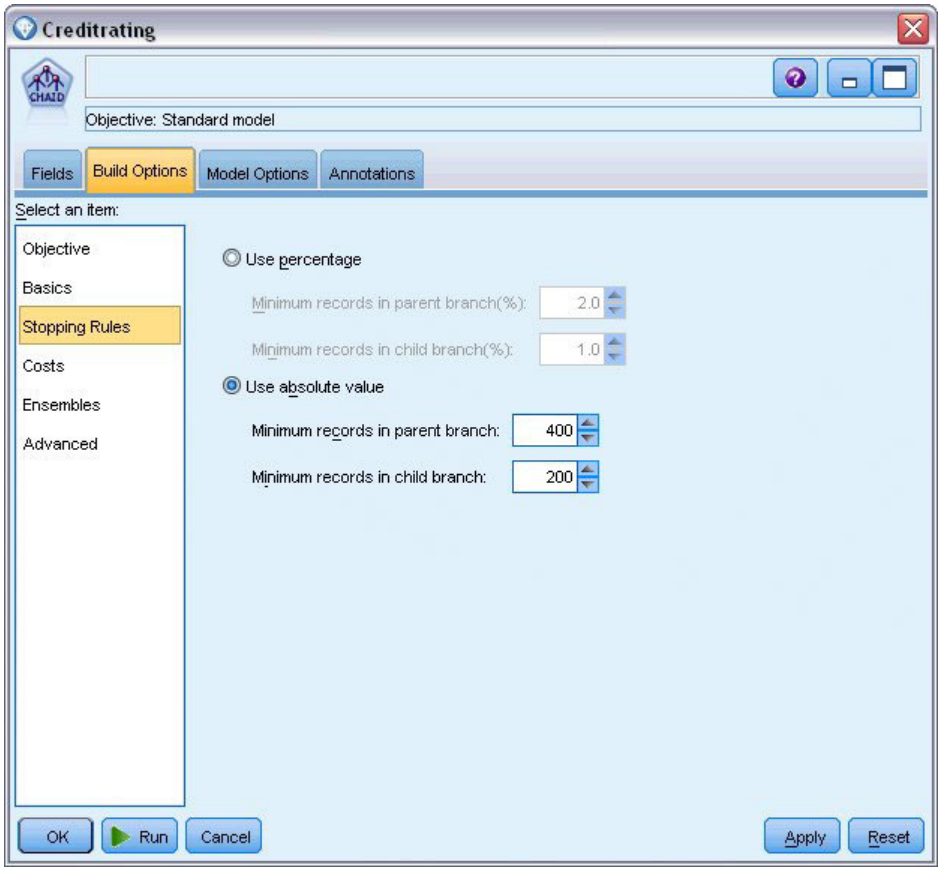


Abbildung 7. Festlegen der Stoppkriterien beim Erstellen von Entscheidungsbäumen

Wir können in diesem Beispiel alle anderen Standardoptionen verwenden. Klicken Sie daher auf **Ausführen**, um das Modell zu erstellen. (Alternativ können Sie mit der rechten Maustaste auf den Knoten klicken und im Kontextmenü **Ausführen** auswählen oder können Sie den Knoten auswählen und **Ausführen** im Menü "Extras" auswählen.)

Durchsuchen des Modells

Nach Abschluss der Ausführung wird das Modellnugget der Modellpalette rechts oben im Anwendungsfenster hinzugefügt. Zusätzlich wird es im Streamerstellungsbereich mit einer Verknüpfung zum Modellierungsknoten angezeigt, von dem aus es erstellt wurde. Um die Modelldetails anzuzeigen, klicken Sie mit der rechten Maustaste auf den generierten Modellknoten und wählen **Durchsuchen** (in der Modellpalette) oder **Bearbeiten** (im Erstellungsbereich).

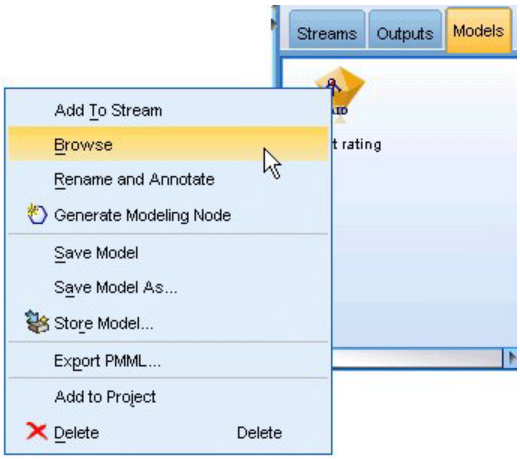


Abbildung 8. Modellpalette

Im Fall des CHAID-Nuggets werden auf der Registerkarte "Modell" die Details in Form eines Regelsets dargestellt. Im Wesentlichen handelt es sich hierbei um eine Reihe von Regeln, die dazu verwendet werden können, einzelne Datensätze untergeordneten Knoten basierend auf den Werten verschiedener Eingabefelder zuzuweisen.

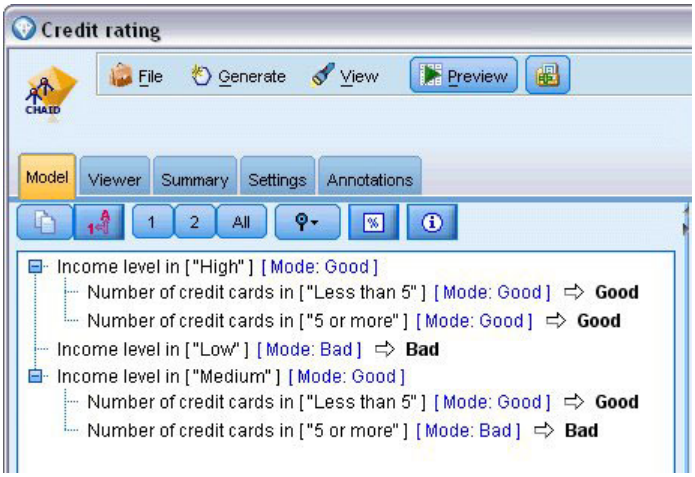


Abbildung 9. CHAID-Modellnugget, Regelset

Für jeden Entscheidungsbaum-Endknoten (also die Baumknoten, die nicht weiter aufgeteilt werden) wird die Vorhersage *Gut* oder *Schlecht* getroffen. In jedem Fall wird die Vorhersage für Datensätze, die unter diesen Knoten fallen, durch den **Modus** bestimmt, also durch die häufigste Antwort.

Rechts neben dem Regelset werden auf der Registerkarte "Modell" das Diagramm "Bedeutsamkeit der Prädiktoren", das die relative Wichtigkeit jedes Prädiktors beim Schätzen des Modells angezeigt. Das zeigt uns, dass *Einkommen in Kategorien* in diesem Fall eindeutig die größte Bedeutung hat, und dass der einzige andere bedeutsame Faktor die *Anzahl der Kreditkarten* ist.

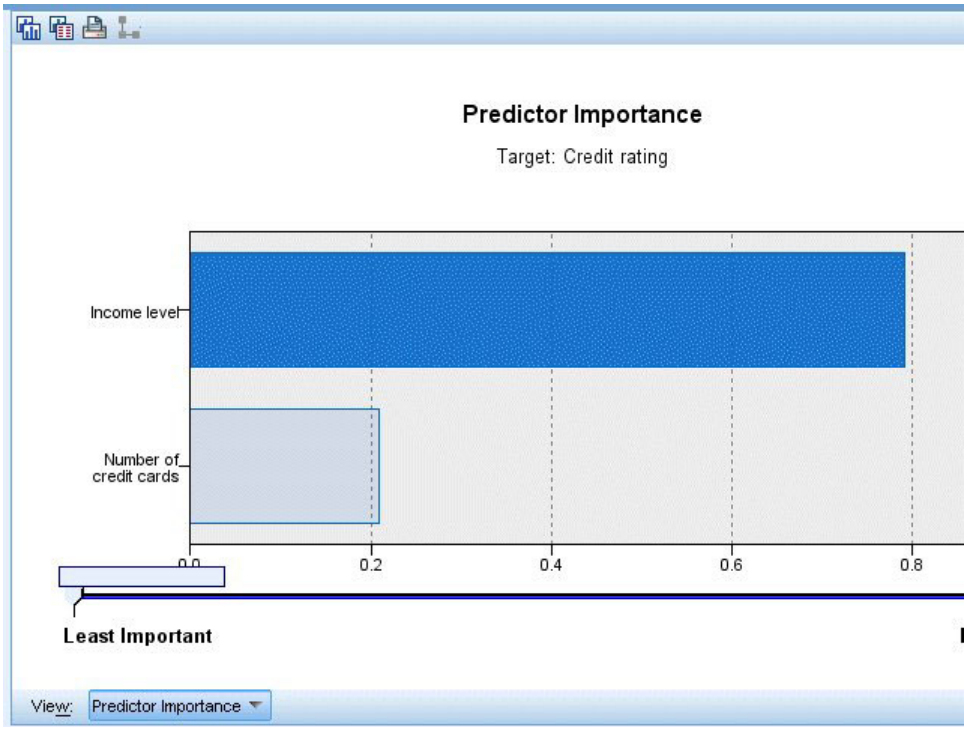


Abbildung 10. Bedeutsamkeit der Prädiktoren - Diagramm

Auf der Registerkarte "Viewer" im Modellnugget wird dasselbe Modell in Form eines Baums angezeigt, mit einem Knoten bei jedem Entscheidungspunkt. Mit den Zoomsteuerelementen auf der Symbolleiste können Sie die Ansicht eines bestimmten Knotens vergrößern bzw. die Ansicht verkleinern, um einen größeren Ausschnitt aus dem Baum anzuzeigen.

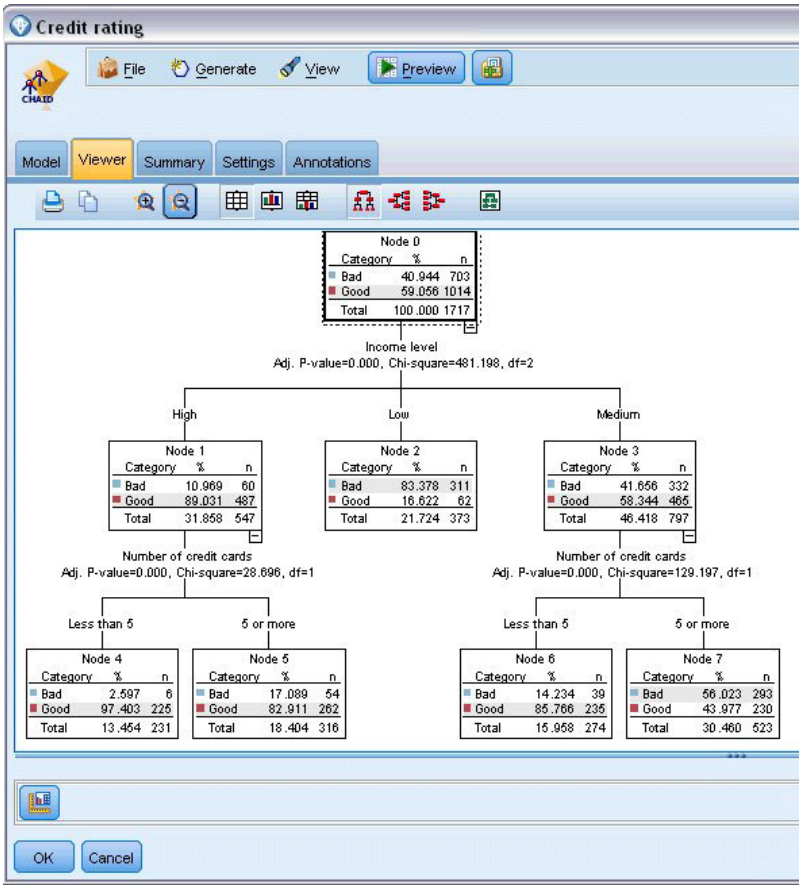


Abbildung 11. Registerkarte "Viewer" im Modellnugget, "Verkleinern" ausgewählt

Im oberen Teil des Baums fasst der erste Knoten (Knoten 0) alle Datensätze im Dataset zusammen. Knapp über 40 % der Fälle im Dataset sind als hoch riskant eingestuft. Da dieser Anteil ziemlich hoch ist, interessiert es uns, ob der Baum Informationen darüber enthält, welche Faktoren hierfür verantwortlich sind.

Wie wir sehen, findet die erste Aufteilung bei *Einkommen in Kategorien* statt. Datensätze, bei denen die Einkommensstufe in der Kategorie *Niedrig* liegt, werden Knoten 2 zugewiesen. Entsprechend enthält diese Kategorie den höchsten Prozentsatz an Kreditausfällen. Die Kreditvergabe an Kunden in dieser Kategorie bringt offensichtlich ein hohes Risiko mit sich.

Bei 16 % der Kunden in dieser Kategorie kam es allerdings *nicht* zum Kreditausfall, die Vorhersage stimmt also nicht in jedem Fall. Kein Modell kann jede Antwort korrekt vorhersagen, aber ein gutes Modell sollte es ermöglichen, die auf der Grundlage der verfügbaren Daten *wahrscheinlichste* Antwort für die einzelnen Datensätze vorherzusagen.

Wenn wir die Kunden mit hohem Einkommen betrachten (Knoten 1), ist das Risiko bei der überwiegenden Mehrheit (89 %) entsprechend gering. Aber mehr als 1 aus 10 dieser Kunden ist ebenfalls seinen Kreditverpflichtungen nicht nachgekommen. Ist es möglich, die Kreditvergabekriterien zu verfeinern, um das Risiko zu minimieren?

Wie Sie sehen, hat das Modell diese Kunden auf Basis der Anzahl ihrer Kreditkarten in zwei Unterkategorien (Knoten 4 und 5) aufgeteilt. Wenn wir Kredite nur an Kunden mit hohem Einkommen vergeben, die weniger als 5 Kreditkarten besitzen, können wir unsere Erfolgsquote von 89 % auf 97 % erhöhen und somit ein noch zufrieden stellenderes Ergebnis erzielen.

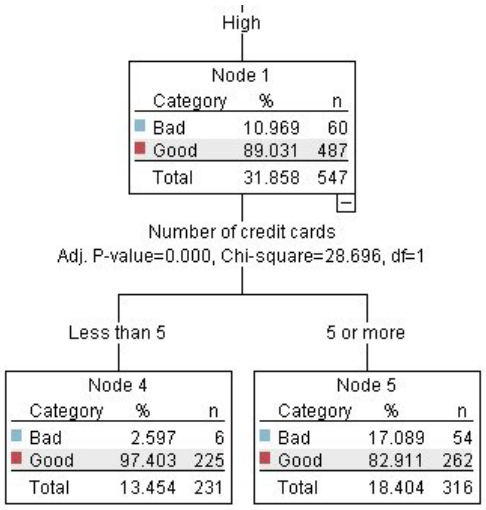


Abbildung 12. Baumansicht der Kunden mit hohem Einkommen

Aber was ist mit den Kunden in der Kategorie mit mittlerem Einkommen (Knoten 3)? Die Verteilung auf gute und schlechte Bonität fällt bei ihnen viel gleichmäßiger aus.

Auch hier sind wieder die Unterkategorien (in diesem Fall Knoten 6 und 7) sehr hilfreich. Wenn wir Kredite nur an Kunden mit mittlerem Einkommen vergeben, die weniger als 5 Kreditkarten besitzen, können wir unsere Erfolgsquote wieder von 58 % auf 85 % erhöhen und somit ein noch zufrieden stellenderes Ergebnis erzielen.

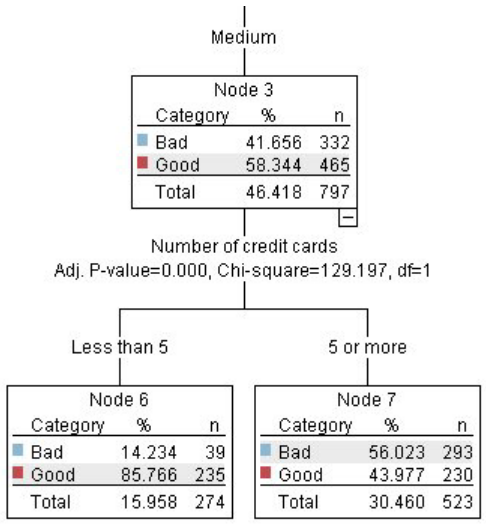


Abbildung 13. Baumansicht der Kunden mit mittlerem Einkommen

Wir haben gesehen, dass jeder Datensatz, der in diesem Modell verarbeitet wird, einem spezifischen Knoten und der Vorhersage *Gut* oder *Schlecht* zugewiesen wird, je nachdem, welche die häufigste Antwort für den jeweiligen Knoten ist.

Dieser Vorgang der Zuweisung von Vorhersagen zu einzelnen Datensätzen wird als **Scoring** bezeichnet. Indem wir die Datensätze scoren, die auch zur Schätzung des Modells verwendet wurden, können wir evaluieren, mit welcher Genauigkeit das Modell für die Trainingsdaten (die Daten, für die das Ergebnis berechnet werden soll) funktioniert. Sehen wir uns an, wie das funktioniert.

Bewerten des Modells

Wir haben das Modell durchsucht, um zu verstehen, wie das Scoren funktioniert. Aber um zu evaluieren, *mit welcher Genauigkeit* es funktioniert, müssen wir einige Datensätze scoren und die vom Modell vorhergesagten Ergebnisse mit den tatsächlichen Ergebnissen vergleichen. Nun werden wir dieselben Datensätze bewerten, die zum Schätzen des Modells verwendet wurden, und können damit die beobachteten und vorhergesagten Antworten vergleichen.

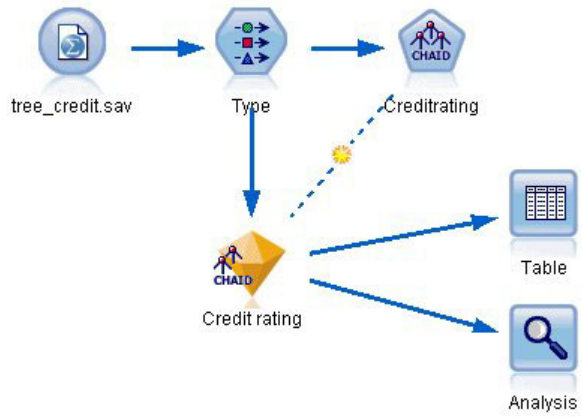
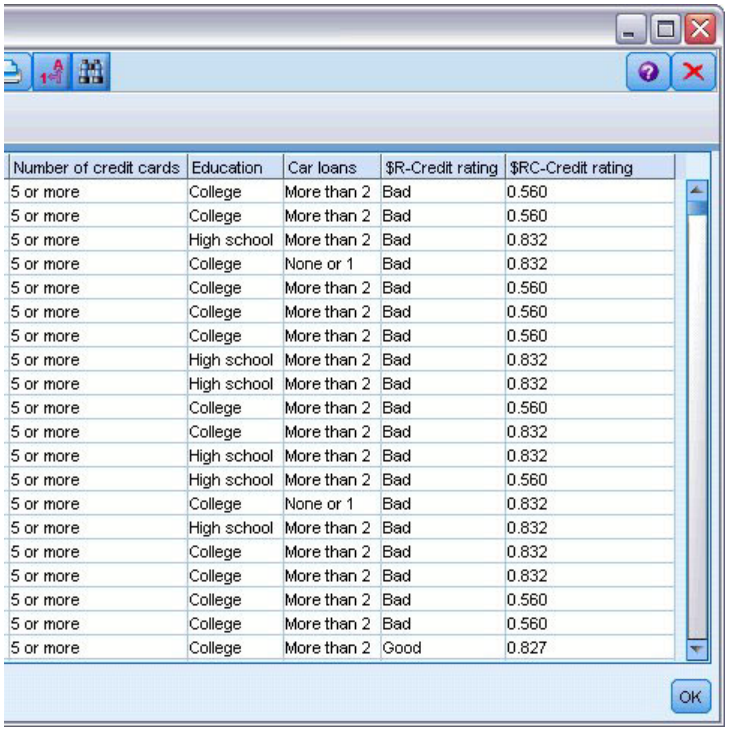


Abbildung 14. Anfügen des Modellnuggets an Ausgabeknoten zur Modellevaluierung

1. Fügen Sie zum Anzeigen der Scores oder Vorhersagen den Tabellenknoten dem Modellnugget hinzu, doppelklicken Sie auf den Tabellenknoten und klicken Sie auf **Ausführen**.

   In der Tabelle werden die vorhergesagten Scores unter einem Feldnamen (*$R-Credit rating*) angezeigt, der vom Modell erstellt wurde. Wir können diese Werte mit dem ursprünglichen Feld *Kreditrating* vergleichen, das die tatsächlichen Antworten enthält.

   Gemäß der Konvention beruhen die Namen der während des Scorens generierten Felder auf dem Zielfeld, weisen jedoch ein Standardpräfix auf, wie beispielsweise *$R-* für Vorhersagen oder *$RC-* für Konfidenzwerte. Verschiedene Modelltypen verwenden verschiedene Präfixsets. Ein **Konfidenzwert** ist die Schätzung des Modells (auf einer Skala von 0,0 bis 1,0) bezüglich der Genauigkeit der einzelnen vorhergesagten Werte.



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	College	None or 1	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.560
5 or more	College	None or 1	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Good	0.827

OK

Abbildung 15. Tabelle mit generierten Scores und Konfidenzwerten

Erwartungsgemäß stimmt der vorhergesagte Wert bei vielen - nicht jedoch bei allen - Datensätzen mit dem tatsächlichen Ergebnis überein. Der Grund hierfür besteht darin, dass jeder CHAID-Endknoten eine Mischung von Ergebnissen aufweist. Die Vorhersage stimmt mit dem *häufigsten* überein, ist jedoch für alle anderen im Knoten falsch. (Wir erinnern uns an die Minderheit von 16 % der Kunden mit niedrigem Einkommen, die Ihren Kredit zurückgezahlt haben.)

Um dies zu vermeiden, könnten wir damit fortfahren, den Baum in immer kleinere Verzweigungen aufzuspalten, bis jeder Knoten 100%ig einheitlich wäre - nur *Gut* oder nur *Schlecht*, ohne gemischte Antworten. Ein derartiges Modell wäre jedoch extrem kompliziert und ließe sich vermutlich nicht gut auf andere Datasets verallgemeinern.

Um herauszufinden, wie viele der Vorhersagen genau zutreffen, könnten wir die Tabelle durchlesen und die Datensätze zählen, bei denen der Wert im vorhergesagten Feld *$R-Credit rating* dem Wert im Feld *Credit rating* entspricht. Zum Glück gibt es eine viel einfachere Methode: Wir können einen Analyseknoten verwenden, der dies automatisch erledigt.

2. Verbinden Sie das Modellnugget mit dem Analyseknoten.
3. Doppelklicken Sie auf den Analyseknoten und klicken Sie auf **Ausführen**.

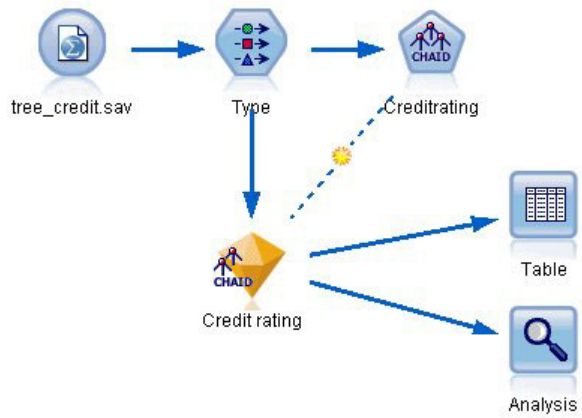


Abbildung 16. Einfügen eines Analyseknotens

Die Analyse zeigt, dass für 1899 von 2464 Datensätzen (über 77 %) der vom Modell vorhergesagte Wert mit der tatsächlichen Antwort übereinstimmte.

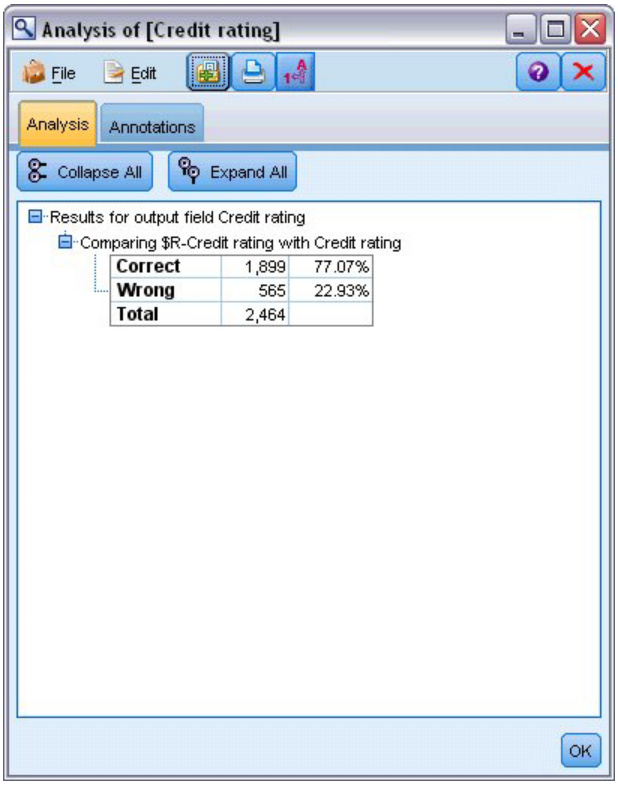


Abbildung 17. Analyseergebnisse für den Vergleich zwischen den beobachteten und vorhergesagten Ergebnissen

Das Ergebnis wird durch die Tatsache eingeschränkt, dass die gescorten Datensätze dieselben sind, die zur Schätzung des Modells verwendet werden. In einer realen Situation könnten Sie einen Partitionsknoten verwenden, um die Daten in separate Stichproben für Training und Evaluierung aufzuteilen.

Durch Verwendung einer Stichprobenpartition zur Generierung des Modells und einer weiteren Stichprobenpartition zum Testen des Modells können Sie einen wesentlich besseren Anhaltspunkt dafür erhalten, wie gut sich das Modell für andere Datasets verallgemeinern lässt.

Mit dem Analyseknoten können wir das Modell an Datensätzen testen, bei denen wir das tatsächliche Ergebnis bereits kennen. Im nächsten Schritt wird gezeigt, wie wir mit dem Modell Datensätze scoren können, deren Ergebnis wir noch nicht kennen. Es könnten z. B. Personen miteinbezogen werden, die noch keine Kunden der Bank sind, die aber potenzielle Ziele für Werberundschreiben sind.

Scoren von Datensätzen

Zuvor haben wir dieselben Datensätze gescort, die zur Schätzung des Modells verwendet wurden, um zu evaluieren, wie genau das Modell war. Jetzt werden wir sehen, wie wir einen anderen Datensatz verwenden als den zur Erstellung des Modells. Dies ist das Ziel der Modellierung mit einem Zielfeld: Untersuchung von Datensätzen, bei denen das Ergebnis bekannt ist, um Muster zu ermitteln, mit denen sich Ergebnisse vorhersagen lassen, die noch nicht bekannt sind.

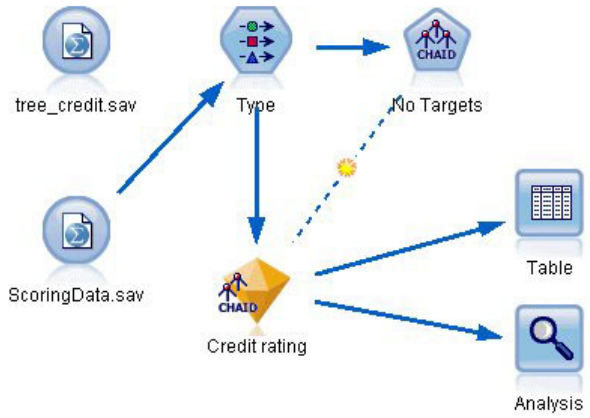


Abbildung 18. Anfügen neuer Daten zum Scoring

Sie können den Quellenknoten für Statistikdateien so aktualisieren, dass er auf eine andere Datendatei verweist, oder Sie können einen neuen Quellenknoten hinzufügen, der die zu scorenden Daten einliest. In beiden Fällen muss das neue Dataset dieselben Eingabefelder enthalten wie das Modell (*Age* (Alter), *Income level* (Einkommenskategorie), *Education* (Bildung) usw.), nicht jedoch das Zielfeld *Credit Rating* (Kreditrating).

Alternativ können Sie das Modellnugget einem beliebigen Stream hinzufügen, der die erwarteten Eingabefelder enthält. Es ist egal, ob die Daten aus einer Datei oder einer Datenbank eingelesen wurden; der Quellentyp ist unerheblich, solange die Feldnamen und -typen mit denen im Modell verwendeten übereinstimmen.

Sie können das Modellnugget auch als separate Datei speichern, das Modell im PMML-Format exportieren, um es in anderen Anwendungen zu nutzen, die dieses Format unterstützen, oder das Modell in IBM SPSS Collaboration and Deployment Services speichern, was Bereitstellung, Scoring und Verwaltung der Modelle im gesamten Unternehmen ermöglicht.

Unabhängig von der verwendeten Infrastruktur funktioniert das Modell auf dieselbe Weise.

Zusammenfassung

In diesem Beispiel werden die grundlegenden Schritte für Erstellung, Evaluation und Scoring eines Modells erläutert.

- Der Modellierungsknoten schätzt das Modell durch Untersuchung von Datensätzen, deren Ergebnis bekannt ist, und erstellt ein Modellnugget. Dieser Vorgang wird auch als Trainieren des Modells bezeichnet.
- Das Modellnugget kann jedem Stream mit den erwarteten Feldern hinzugefügt werden, um Datensätze zu scoren. Durch Scoren der Datensätze, deren Ergebnis Sie bereits kennen (z. B. bestehende Kunden), können Sie die Leistung des Modells evaluieren.
- Sobald Sie mit der Leistungsfähigkeit des Modells zufrieden sind, können Sie neue Daten (beispielsweise potenzielle Kunden) scoren, um vorherzusagen, wie diese reagieren.
- Die zum Trainieren bzw. Schätzen des Modells verwendeten Daten können auch als analytische oder historische Daten bezeichnet werden; die Scoring-Daten können auch als operationale Daten bezeichnet werden.

Kapitel 3. Übersicht über die Modellbildung

Modellierungsknoten - Übersicht

IBM SPSS Modeler bietet eine Vielzahl von Modellbildungsmethoden, die aus dem maschinellen Lernen, der künstlichen Intelligenz und der Statistik stammen. Mit den in der Modellierungspalette verfügbaren Methoden können Sie aus Ihren Daten neue Informationen ableiten und Vorhersagemodelle erstellen. Jede Methode besitzt ihre Stärken und eignet sich besonders für bestimmte Problemtypen.

Im *IBM SPSS Modeler-Anwendungshandbuch* finden Sie Beispiele für viele dieser Methoden sowie eine allgemeine Einführung in den Modellierungsprozess. Dieses Handbuch ist als Online-Lernprogramm und im PDF-Format verfügbar. Weitere Informationen finden Sie im Thema „Anwendungsbeispiele" auf Seite 5.

Modellierungsmethoden werden in drei Kategorien unterteilt:

- Klassifikation
- Assoziation
- Segmentierung

Klassifizierungsmodelle

Klassifikationsmodelle verwenden den Wert mindestens eines **Eingabefeldes**, um den Wert mindestens eines Ausgabe- oder **Zielfeldes** vorherzusagen. Einige Beispiele dieser Verfahren sind: Entscheidungsbäume (C&R-Baum-, QUEST-, CHAID- und C5.0-Algorithmen), Regression (lineare, logistische, verallgemeinert lineare und Cox-Regressionsalgorithmen), neuronale Netze, Support Vector Machines und Bayes-Netze.

Klassifizierungsmodelle können Unternehmen ein bekanntes Ergebnis vorhersagen. Beispielsweise, ob ein Kunde kaufen wird oder nicht, oder ob eine Transaktion mit einem bekannten Betrugsmuster übereinstimmt. Zu den Modellierungstechniken gehören Maschinenlernen, Regelinduktion, Identifikation von Untergruppen, statistische Methoden und die Erzeugung mehrerer Modelle.

Klassifikationsknoten

Mit dem Knoten "Autom. Klassifikationsmerkmal" können Sie eine Reihe verschiedener Modelle für binäre Ergebnisse ("Ja" oder "Nein", "Abwanderung" oder "Keine Abwanderung" usw.) erstellen und vergleichen, um den besten Ansatz für die jeweilige Analyse auszuwählen. Es wird eine Reihe von Modellierungsalgorithmen unterstützt, sodass Sie die gewünschten Methoden, die spezifischen Optionen für die jeweilige Methode und die Kriterien zum Vergleich der Ergebnisse auswählen können. Der Knoten generiert eine Gruppe von Modellen, die auf den angegebenen Optionen beruhen, und erstellt anhand der von Ihnen angegebenen Kriterien eine Rangordnung der besten Kandidaten.

Der Knoten "Autonumerisch" schätzt und vergleicht mit einer Reihe verschiedener Methoden Modelle für die Ergebnisse stetiger numerischer Bereiche. Der Knoten arbeitet auf dieselbe Weise wie der Knoten "Autom. Klassifikationsmerkmal": Sie können die zu verwendenden Algorithmen auswählen und in einem Modellierungsdurchlauf mit mehreren Optionskombinationen experimentieren. Folgende Algorithmen werden unterstützt: Neuronale Netze, C&R-Baum, CHAID, lineare Regression, verallgemeinerte lineare Regression nd Support Vector Machines (SVM). Modelle können anhand von Korrelation, relativem Fehler bzw. Anzahl der verwendeten Variablen verglichen werden.

Der Knoten für Klassifizierungs- und Regressionsbäume (C&R-Bäume) erstellt einen Entscheidungsbaum, mit dem Sie zukünftige Beobachtungen vorhersagen oder klassifizieren können. Bei dieser Methode wird eine rekursive Partitionierung verwendet, um die Trainingsdatensätze in Segmente aufzuteilen. Dabei wird bei jedem Schritt die Unreinheit verringert. Ein Knoten im Baum wird als "rein" betrachtet, wenn 100 % der Fälle im Knoten in eine bestimmte Kategorie des Zielfelds fallen. Ziel- und Eingabefelder können numerische Bereiche oder kategorial (nominal, ordinal oder Flags) sein. Alle Aufteilungen sind binär (nur zwei Untergruppen).

Der QUEST-Knoten bietet eine binäre Klassifizierungsmethode zum Erstellen von Entscheidungsbäumen, die dazu dient, die für Analysen von großen C&R-Bäumen erforderliche Verarbeitungszeit zu verkürzen. Gleichzeitig soll die in den Klassifizierungsbaummethoden festgestellte Tendenz verringert werden, die darin besteht, dass Eingaben bevorzugt werden, die mehr Aufteilungen erlauben. Eingabefelder können stetig (numerische Bereiche) sein, das Zielfeld muss aber kategorial sein. Alle Aufteilungen sind binär.

Der CHAID-Knoten erzeugt Entscheidungsbäume unter Verwendung von Chi-Quadrat-Statistiken zur Ermittlung optimaler Aufteilungen. Im Gegensatz zu den Knoten vom Typ "C&R-Baum" und "QUEST" kann CHAID nicht binäre Bäume generieren, d. h. Bäume mit Aufteilungen mit mehr als zwei Verzweigungen. Ziel- und Eingabefelder können in einem numerischen Bereich (stetig) oder kategorial sein. Exhaustive CHAID ist eine Änderung von CHAID, die noch gründlicher vorgeht, indem sie alle möglichen Aufteilungen untersucht, allerdings mehr Rechenzeit beansprucht.

Der C5.0-Knoten erstellt entweder einen Entscheidungsbaum oder ein Regelset. Das Modell teilt die Stichprobe auf der Basis des Felds auf, das auf der jeweiligen Ebene den maximalen Informationsgewinn liefert. Das Zielfeld muss kategorial sein. Es sind mehrere Aufteilungen in mehr als zwei Untergruppen zulässig.

Der Knoten "Entscheidungsliste" kennzeichnet Untergruppen bzw. Segmente, die eine höhere oder geringere Wahrscheinlichkeit für ein bestimmtes binäres Ergebnis aufweisen als die Gesamtpopulation. Sie könnten beispielsweise nach Kunden suchen, deren Abwanderung unwahrscheinlich ist oder die mit großer Wahrscheinlichkeit positiv auf eine Kampagne reagieren. Sie können Ihr Fachwissen in das Modell integrieren, indem Sie eigene, benutzerdefinierte Segmente hinzufügen und eine Vorschau anzeigen, in der alternative Modelle nebeneinander angezeigt werden, um die Ergebnisse zu vergleichen. Entscheidungslistenmodelle bestehen aus einer Liste von Regeln, bei denen jede Regel eine Bedingung und ein Ergebnis aufweist. Regeln werden in der vorgegebenen Reihenfolge angewendet und die erste Regel, die zutrifft, bestimmt das Ergebnis.

Bei linearen Regressionsmodellen wird ein stetiges Ziel auf der Basis linearer Beziehungen zwischen dem Ziel und einem oder mehreren Prädiktoren vorhergesagt.

Der Faktor/PCA-Knoten bietet leistungsstarke Datenreduktionsverfahren zur Verringerung der Komplexität der Daten. Die Hauptkomponentenanalyse (PCA) findet lineare Kombinationen der Eingabefelder, die die Varianz im gesamten Set der Felder am besten erfassen, wenn die Komponenten orthogonal (senkrecht) zueinander sind. Mit der Faktorenanalyse wird versucht, die zugrunde liegenden Faktoren zu bestimmen, die die Korrelationsmuster innerhalb eines Sets beobachteter Felder erklären. Bei beiden Ansätzen besteht das Ziel darin, eine kleine Zahl abgeleiteter Felder zu finden, mit denen die Informationen im ursprünglichen Set der Felder effektiv zusammengefasst werden können.

Der Merkmalauswahlknoten sichtet die Eingabefelder, um auf der Grundlage einer Reihe von Kriterien (z. B. dem Prozentsatz der fehlenden Werte) zu entscheiden, ob diese entfernt werden sollen. Anschließend erstellt er eine Wichtigkeitsrangfolge der verbleibenden Eingaben in Bezug auf ein angegebenes Ziel. Beispiel: Angenommen, Sie haben ein Dataset mit Hunderten potenzieller Eingaben. Welche davon sind voraussichtlich für die Modellierung von medizinischen Behandlungsergebnissen von Bedeutung?

Bei der Diskriminanzanalyse werden strengere Annahmen als bei der logistischen Regression verwendet, sie kann jedoch eine wertvolle Alternative oder Ergänzung zu einer logistischen Regressionsanalyse sein, wenn diese Annahmen erfüllt sind.

Die logistische Regression ist ein statistisches Verfahren zur Klassifizierung von Datensätzen auf der Grundlage der Werte von Eingabefeldern. Sie ist analog zur linearen Regression, außer dass statt eines numerischen Bereichs ein kategoriales Zielfeld verwendet wird.

Das verallgemeinerte lineare Modell erweitert das allgemeine lineare Modell so, dass die abhängige Variable über eine angegebene Verknüpfungsfunktion in linearem Zusammenhang zu den Faktoren und Kovariaten steht. Außerdem ist es mit diesem Modell möglich, dass die abhängige Variable eine von der Normalverteilung abweichende Verteilung aufweist. Es deckt die Funktionen einer großen Bandbreite an Statistikmodellen ab, darunter lineare Regression, logistische Regression, loglineare Modelle für Häufigkeitsdaten und Überlebensmodelle mit Intervallzensierung.

Verallgemeinerte lineare gemischte Modelle (GLMM - Generalized Linear Mixed Models) erweitern lineare Modelle so, dass das Ziel nicht normalverteilt zu sein braucht und über eine angegebene Verknüpfungsfunktion in einer linearen Beziehung zu den Faktoren und Kovariaten steht und die Beobachtungen korreliert werden können. Verallgemeinerte lineare gemischte Modelle decken eine breite Palette verschiedener Modelle ab, von einfacher linearer Regression bis hin zu komplexen Mehrebenenmodellen für nicht normalverteilte Longitudinaldaten.

Der Knoten vom Typ "Cox-Regression" ermöglicht Ihnen auch bei zensierten Datensätzen die Erstellung eines Überlebensmodells für Daten über die Zeit bis zum Eintreten des Ereignisses. Das Modell erstellt eine Überlebensfunktion, die die Wahrscheinlichkeit vorhersagt, dass das untersuchte Ereignis für bestimmte Werte der Eingabevariablen zu einem bestimmten Zeitpunkt (*t*) eingetreten ist.

Der Knoten "Support Vector Machine" (SVM) ermöglicht die Klassifizierung von Daten in eine von zwei Gruppen ohne Überanpassung. SVM eignet sich gut für umfangreiche Datasets, beispielsweise solche mit einer großen Anzahl an Eingabefeldern.

Mithilfe des Bayes-Netzknotens können Sie ein Wahrscheinlichkeitsmodell erstellen, idem Sie beobachtete und aufgezeichnete Hinweise mit Weltwissen kombinieren, um die Wahrscheinlichkeit ihres Vorkommens zu ermitteln. Der Knoten ist speziell für Netze vom Typ "Tree Augmented Naïve Bayes" (TAN) und "Markov-Decke" gedacht, die in erster Linie zur Klassifizierung verwendet werden.

Mithilfe des Knotens für das lernfähige Antwortmodell (Self-Learning Response Model, SLRM) können Sie ein Modell erstellen, in dem das Modell anhand eines einzelnen neuen Falls oder einer kleinen Anzahl neuer Fälle neu eingeschätzt werden kann, ohne dass das Modell mit allen Daten neu trainiert werden muss.

Der Zeitreihenknoten berechnet Schätzungen für die exponenielle Glättung sowie univariate und multivariate ARIMA-Modelle (ARIMA steht für Autoregressive Integrated Moving Average (autoregressiver integrierter gleitender Durchschnitt)) für Zeitreihendaten und erstellt Vorhersagen über die zukünftige Leistung. Einem Zeitreihenknoten muss stets ein Zeitintervallknoten vorangehen.

Der Knoten "*k*-Nächste Nachbarn" (KNN) verknüpft einen neuen Fall mit der Kategorie oder dem Wert der *k* Objekte, die ihm im Prädiktorraum am nächsten liegen, wobei *k* eine Ganzzahl ist. Ähnliche Fälle liegen nah beieinander und Fälle mit geringer Ähnlichkeit sind weit voneinander entfernt.

Assoziationsmodelle

Assoziationsmodelle finden Muster in Ihren Daten, bei denen mindestens eine Entität (wie Ereignisse, Einkäufe oder Attribute) mindestens einer anderen Entität zugeordnet sind. Die Modelle erstellen Regelsets, die diese Beziehungen definieren. Hier können die Felder innerhalb der Daten sowohl als Eingabe- als auch als Zielfelder fungieren. Sie könnten diese Assoziationen manuell finden, doch mithilfe von Assoziationsregelalgorithmen ist die Suche wesentlich schneller und es können komplexere Muster untersucht werden. Apriori- und Carma-Modelle sind Beispiele für die Verwendung solcher Algorithmen. Ein weiterer Typ eines Assoziationsmodells ist ein Sequenzerkennungsmodell, das sequenzielle Muster in zeitstrukturierten Daten findet.

Assoziationsmodelle sind bei der Vorhersage mehrerer Ergebnisse am nützlichsten, beispielsweise Kunden, die Produkt X gekauft haben, kauften auch Produkt Y und Z. Assoziationsmodelle ordnen einem Set von Bedingungen eine bestimmte Schlussfolgerung zu (wie zum Beispiel die Entscheidung, etwas zu kaufen). Der Vorteil von Algorithmen für Assoziationsregeln im Vergleich zu Algorithmen für Standardentscheidungsbäume (C5.0 und C&R-Baum) liegt darin, dass Zuordnungen zwischen beliebigen Attributen bestehen können. Ein Entscheidungsbaumalgorithmus erstellt Regeln mit nur einer Schlussfolgerung, während Assoziationsalgorithmen viele Regeln zu finden versuchen, von denen jede zu einer anderen Schlussfolgerung kommen kann.

Assoziationsknoten

Der Apriori-Knoten extrahiert ein Regelset aus den Daten und daraus die Regeln mit dem höchsten Informationsgehalt. Apriori bietet fünf verschiedene Methoden zur Auswahl von Regeln und verwendet ein ausgereiftes Indizierungsschema zur effizienten Verarbeitung großer Datasets. Bei großen Problemen ist Apriori in der Regel schneller zu trainieren, es gibt keine willkürliche Begrenzung für die Anzahl der Regeln, die beibehalten werden können, und es können Regeln mit bis zu 32 Vorbedingungen verarbeitet werden. Bei Apriori müssen alle Ein- und Ausgabefelder kategorial sein; dafür bietet es jedoch eine bessere Leistung, da es für diesen Datentyp optimiert ist.

Beim CARMA-Modell wird ein Regelset aus den Daten extrahiert, ohne dass Sie Eingabe- oder Zielfelder angeben müssen. Im Gegensatz zu Apriori bietet der CARMA-Knoten Einstellungen für Regelunterstützung (Unterstützung für Antezedens und Sukzedens) und nicht nur Unterstützung für Antezedens. Die erstellten Regeln können somit für eine größere Palette von Anwendungen verwendet werden, beispielsweise um eine Liste mit Produkten oder Services (Antezedenzien) zu suchen, deren Sukzedens das Element ist, das Sie in dieser Ferienzeit bewerben möchten.

Der Sequenzknoten erkennt Assoziationsregeln in sequenziellen oder zeitorientierten Daten. Eine Sequenz ist eine Liste mit Elementsets, die in einer vorhersagbaren Reihenfolge auftreten. Beispiel: Ein Kunde, der einen Rasierer und After-Shave-Lotion kauft, kauft möglicherweise beim nächsten Einkauf Rasiercreme. Der Sequenzknoten basiert auf dem CARMA-Assoziationsregelalgorithmus, der eine effiziente bidirektionale Methode zum Suchen von Sequenzen verwendet.

Segmentierungsmodelle

Segmentierungsmodelle teilen die Daten in Segmente, oder Cluster, von Datensätzen auf, die ähnliche Muster von Eingabefeldern aufweisen. Da sie nur an den Eingabefeldern interessiert sind, verfügen Segmentierungsmodelle nicht über die Konzepte der Ausgabe- oder Zielfelder. Beispiele für Segmentierungsmodelle sind Kohonen-Netze, K-Means-Clustering, TwoStep-Clustering und Anomalieerkennung.

Segmentierungsmodelle (auch "Clustering-Modelle") sind dann nützlich, wenn das genaue Ergebnis unbekannt ist (zum Beispiel, beim Ermitteln neuer Betrugsmuster oder von Interessengruppen in Ihrem Kundenstamm). Clustermodelle konzentrieren sich auf die Ermittlung ähnlicher Datensätze und Beschriftung der Datensätze anhand der Gruppe, in die sie gehören. Dies erfolgt ohne den Vorteil bereits zuvor vorhandener Kenntnisse der Gruppen und der zugehörigen Merkmale. Dies unterscheidet Clustermodelle von anderen Modellierungsverfahren: Es gibt kein zuvor definiertes Ausgabe- oder Zielfeld für das vorherzusagende Modell. Für diese Modelle gibt es keine richtigen oder falschen Antworten. Ihr Wert wird durch die Möglichkeit bestimmt, interessante Gruppierungen in den Daten zu erfassen und sinnvolle Beschreibungen dieser Gruppierungen zu liefern. Clustermodelle werden häufig verwendet, um Cluster oder Segmente zu erstellen, die dann als Eingaben in nachfolgenden Analysen verwendet werden (z. B. die Segmentierung potenzieller Kunden in homogene Untergruppen).

Segmentierungsknoten

Mit dem Knoten "Autom. Cluster" können Sie Clustering-Modelle, die Gruppen und Datensätze mit ähnlichen Merkmalen identifizieren, schätzen und vergleichen. Die Funktionsweise des Knotens gleicht der von anderen Knoten für automatisierte Modellierung, und Sie können in einem einzigen Modellierungsdurchgang mit mehreren Optionskombinationen experimentieren. Modelle können mithilfe grundlegender Messwerte für Filterung und Rangfolge der Nützlichkeit von Clustermodellen verglichen werden, um ein Maß auf der Basis der Wichtigkeit von bestimmten Feldern zu liefern.

Der K-Means-Knoten teilt das Dataset in unterschiedliche Gruppen (oder Cluster) auf. Bei dieser Methode wird eine festgelegte Anzahl von Clustern definiert, den Clustern werden iterativ Datensätze zugewiesen und die Clusterzentren werden angepasst, bis eine weitere Verfeinerung keine wesentliche Verbesserung des Modells mehr darstellen würde. Statt zu versuchen, ein Ergebnis vorherzusagen, versucht *K*-Means mithilfe eines als "nicht überwachtes Lernen" bezeichneten Prozesses Muster im Set der Eingabefelder zu entdecken.

Der Kohonen-Knoten erstellt eine Art von neuronalem Netz, das verwendet werden kann, um ein Clustering des Datasets in einzelne Gruppen vorzunehmen. Wenn das Netz voll trainiert ist, sollten ähnliche Datensätze auf der Ausgabekarte eng nebeneinander stehen, während Datensätze, die sich unterscheiden, weit voneinander entfernt sein sollten. Die Zahl der von jeder Einheit im Modellnugget erfassten Beobachtungen gibt Aufschluss über die starken Einheiten. Dadurch wird ein Eindruck von der ungefähren Zahl der Cluster vermittelt.

Der TwoStep-Knoten verwendet eine aus zwei Schritten bestehende Clusterbildungsmethode. Im ersten Schritt wird ein einzelner Durchlauf durch die Daten vorgenommen, bei dem die Eingangsrohdaten zu einem verwaltbaren Set von Subclustern komprimiert werden. Im zweiten Schritt werden die Subcluster mithilfe einer hierarchischen Methode zur Clusterbildung nach und nach in immer größere Cluster zusammengeführt. TwoStep hat den Vorteil, dass die optimale Anzahl an Clustern für die Trainingsdaten automatisch geschätzt wird. Mit dem Verfahren können gemischte Feldtypen und große Datasets effizient verarbeitet werden.

Der Knoten "Anomalieerkennung" ermittelt ungewöhnliche Fälle oder Ausreißer, die nicht den Mustern von "normalen" Daten entsprechen. Mit diesem Knoten können Ausreißer ermittelt werden, selbst wenn sie keinem bereits bekannten Muster entsprechen und selbst wenn Sie nicht genau wissen, wonach Sie suchen.

Modelle für datenbankinternes Mining

IBM SPSS Modeler unterstützt die Integration in Data-Mining- und Modellierungstools von Datenbankanbietern wie Oracle Data Miner, IBM DB2 InfoSphere Warehouse und Microsoft Analysis Services. Sie können Modelle erstellen, scoren und in der Datenbank speichern, ohne dazu die IBM SPSS Modeler-Anwendung verlassen zu müssen. Vollständige Informationen finden Sie im IBM SPSS Modeler-Handbuch zum datenbankinternen Mining, das sich auf der Produkt-DVD befindet.

IBM SPSS Statistics-Modelle

Wenn auf Ihrem Computer eine Kopie von IBM SPSS Statistics installiert und lizenziert ist, können Sie auf bestimmte IBM SPSS Statistics-Routinen in IBM SPSS Modeler zugreifen und diese ausführen, um Modelle zu erstellen und zu scoren.

Weitere Informationen

Zu den Modellierungsalgorithmen ist außerdem eine detaillierte Dokumentation verfügbar. Weitere Informationen finden Sie im Algorithmushandbuch zu IBM SPSS Modeler auf der Produkt-DVD.

Erstellen von aufgeteilten Modellen

Die Erstellung aufgeteilter Modelle ermöglicht es Ihnen, einen einzelnen Stream für die Erstellung getrennter Modelle für jeden möglichen Wert eines Flagfelds oder eines nominalen oder stetigen Eingabefelds zu verwenden, wobei auf alle daraus resultierenden Modelle von einem einzelnen Modellnugget aus zugegriffen werden kann. Die möglichen Werte für die Eingabefelder könnten sehr unterschiedliche Effekte auf das Modell haben. Durch aufgeteilte Modellierung können Sie ganz einfach das am besten geeignete Modell für jeden möglichen Feldwert mit einer einzigen Ausführung des Streams erstellen.

Bitte beachten Sie, dass die Aufteilungsfunktion in interaktiven Modellierungssitzungen nicht verwendet werden kann. Bei der interaktiven Modellierung geben Sie jedes Modell einzeln an, weswegen die Verwendung der Aufteilungsfunktion, über die mehrere Modelle automatisch erstellt werden, nicht von Vorteil wäre.

Die aufgeteilte Modellierung wird angewendet, indem man ein bestimmtes Eingabefeld als Aufteilungsfeld angibt. Dies ist möglich, indem Sie die Feldrolle in der Typspezifikation auf **Aufteilen** setzen.

Sie können nur Felder mit einem Messniveau **Flag**, **Nominal**, **Ordinal** oder **Stetig** als Aufteilungsfelder festlegen.

Sie können mehr als ein Eingabefeld als Aufteilungsfeld festlegen. Dadurch kann jedoch die Anzahl erstellter Modelle erheblich gesteigert werden. Für jede mögliche Kombination der Werte der ausgewählten

Aufteilungsfelder wird ein Modell erstellt. Wenn beispielsweise drei Eingabefelder mit je drei möglichen Werten als Aufteilungsfelder festgelegt werden, führt dies zu einer Erstellung von 27 unterschiedlichen Modellen.

Selbst nachdem Sie mindestens ein Feld als Aufteilungsfeld zugeordnet haben, können Sie mithilfe einer Einstellung für ein Kontrollkästchen im Dialogfeld "Modellierungsknoten" immer noch auswählen, ob aufgeteilte Modelle oder ein einzelnes Modell erstellt werden sollen.

Wenn Aufteilungsfelder definiert sind, das Kontrollkästchen aber nicht aktiviert ist, wird nur ein einzelnes Modell generiert. Ebenso wird, wenn das Kontrollkästchen aktiviert, aber kein Aufteilungsfeld definiert ist, die Aufteilung ignoriert und ein einzelnes Modell generiert.

Wenn Sie einen Stream ausführen, werden für jeden möglichen Wert der Aufteilungsfelder im Hintergrund separate Modelle generiert, aber es wird nur ein einzelnes Modellnugget in der Modellpalette und im Streamerstellungsbereich erstellt. Ein Nugget für aufgeteilte Modelle wird durch das Aufteilungssymbol gekennzeichnet. Dieses besteht aus zwei grauen Rechtecken, die das Nuggetbild überlagern.

Wenn Sie das Nugget für aufgeteilte Modelle durchsuchen, wird eine Liste aller separaten Modelle angezeigt, die erstellt wurden.

Sie können ein individuelles Modell aus einer Liste untersuchen, indem Sie im Viewer auf sein Nuggetsymbol doppelklicken. Damit wird ein Standardbrowserfenster für das individuelle Modell geöffnet. Wenn sich das Nugget im Erstellungsbereich befindet, wird durch Doppelklicken auf ein Piktogramm ein Diagramm in voller Größe geöffnet. Weitere Informationen finden Sie im Thema „Modellviewer aufteilen” auf Seite 47.

Nachdem ein Modell als aufgeteiltes Modell erstellt wurde, können Sie den Aufteilungsprozess nicht mehr rückgängig machen; auch weiter abwärts vorgenommene Aufteilungen können für Aufteilungsmodellierungsknoten oder nuggets nicht rückgängig gemacht werden.

Beispiel. Ein national operierender Einzelhändler möchte Schätzungen der Verkäufe nach Produktkategorie für jedes seiner Geschäfte im ganzen Land vornehmen. Unter Verwendung von Aufteilungsmodellierung legt er das Speicherfeld für seine Eingabedaten als Aufteilungsfeld fest und kann so für jede Kategorie in jedem Geschäft mithilfe eines einzigen Vorgangs separate Modelle erstellen. Durch die so gewonnenen Informationen kann er die Lagerbestände viel genauer kontrollieren als es anhand eines einzelnen Modells möglich wäre.

Aufteilung und Partitionierung

Aufteilung und Partitionierung haben einige gemeinsame Eigenschaften, werden aber auf vollkommen unterschiedliche Arten verwendet.

Die Partitionierung teilt das Dataset zufällig in zwei oder drei Teile auf: Training, Testen und (optional) Validierung. Sie wird verwendet, um die Leistung eines einzelnen Modells zu testen.

Die **Aufteilung** unterteilt das Dataset in so viele Teile, wie es mögliche Werte für ein Aufteilungsfeld gibt, und wird verwendet, um mehrere Modelle zu erstellen.

Partitionierung und Aufteilung sind Vorgänge, die vollkommen unabhängig voneinander sind. In einem Modellierungsknoten können Sie einen von ihnen, beide oder keinen auswählen.

Modellierungsknoten zur Unterstützung aufgeteilter Modelle

Eine Reihe von Modellierungsknoten können aufgeteilte Modelle erstellen. Die Ausnahmen sind: Autom. Cluster, Zeitreihe, Faktor/PCA, Merkmalauswahl, SLRM, die Assoziationsmodelle (Apriori, Carma und Sequence), die Clustering-Modelle (K-Means, Kohonen, Two Step und Anomaly), das Statistics-Modell und die Knoten zur Modellierung innerhalb der Datenbank.

Die folgenden Modellierungsknoten unterstützen aufgeteilte Modellierung:


C&R-Baum


Bayes-Netz


QUEST


GenLin


CHAID


KNN


C5.0


Cox


Netz


Automatisches Klassifikationsmerkmal


Entscheidungsliste


Autonumerisch


Regression


Logistisch


Diskriminanz


SVM

Von der Aufteilung betroffene Merkmale

Die Verwendung von aufgeteilten Modellen beeinflusst eine Reihe von IBM SPSS Modeler-Merkmalen auf mehrere Arten. Dieser Abschnitt bietet Richtlinien zur Nutzung von aufgeteilten Modellen im Zusammenhang mit anderen Knoten in einem Stream.

Knoten "Datensatzoperationen"

Beim Verwenden von aufgeteilten Modellen in einem Stream, der einen **Stichproben**-Knoten enthält, schichten Sie Datensätze nach dem Aufteilungsfeld, um gleichmäßige Stichproben von Datensätzen zu erhalten. Diese Option ist verfügbar, wenn Sie "Komplex" als Stichprobenmethode wählen.

Wenn der Stream einen **Balancierungsknoten** enthält, beachten Sie, dass die Balancierung für das vollständige Set der Eingabedatensätze gilt, nicht für ein Subset von Datensätzen innerhalb einer Aufteilung.

Beim Aggregieren von Datensätzen mithilfe eines **Aggregatknotens** legen Sie die aufgeteilten Felder als Schlüsselfelder fest, wenn Sie Aggregate für jede Aufteilung berechnen möchten.

Feldoperationsknoten

Im Knoten **Typ** geben Sie an, welches Feld bzw. welche Felder als aufgeteilte Felder dienen sollen.

Beachten Sie: Der Knoten **Ensemble** wird zur Kombination von zwei oder mehr Modellnuggets verwendet, lässt sich jedoch nicht verwenden, um die Aktion des Aufteilens umzukehren, da sich die aufgeteilten Modelle in einem einzigen Modellnugget befinden.

Modellierungsknoten

Aufgeteilte Modelle unterstützen keine Berechnung des Prädiktoreinflusses (der relativen Wichtigkeit der Prädiktoreingabefelder bei der Modellschätzung). Die Einstellungen des Prädiktoreinflusses werden bei der Erstellung von aufgeteilten Modellen ignoriert.

Der Knoten **KNN** (Nächster Nachbar) unterstützt nur dann aufgeteilte Modelle, wenn er auf die Vorhersage eines Zielfelds eingestellt ist. Die alternative Einstellung (nur nächste Nachbarn identifizieren) erzeugt kein Modell. Wenn die Option "Automatisch k wählen" aktiviert ist, kann jedes aufgeteilte Modell über eine unterschiedliche Anzahl nächster Nachbarn verfügen. Das Gesamtmodell verfügt so über eine Reihe von generierten Spalten gleich der größten Anzahl nächster Nachbarn, die in allen aufgeteilten Modellen gefunden wird. Für die aufgeteilten Modelle, deren Anzahl nächster Nachbarn dieses Maximum unterschreitet, gibt es eine entsprechende Anzahl von Spalten, die mit `$null`-Werten gefüllt sind. Weitere Informationen finden Sie im Thema „KNN-Knoten” auf Seite 289.

Datenbankmodellierungsknoten

Die Knoten für Modellierung innerhalb der Datenbank unterstützen keine aufgeteilten Modelle.

Modellnuggets

PMML exportieren ist aus einem aufgeteilten Modellnugget nicht möglich, da das Nugget mehrere Modelle enthält und PMML eine solche Zusammenfassung nicht unterstützt. Das Exportieren von Text oder HTML ist jedoch möglich.

Feldoptionen der Modellierungsknoten

Alle Modellierungsknoten besitzen die Registerkarte "Felder", auf der Sie die Felder festlegen können, die beim Erstellen des Modells verwendet werden.

Bevor Sie ein Modell erstellen können, müssen Sie festlegen, welche Felder als Ziele und als Eingaben verwendet werden sollen. Von wenigen Ausnahmen abgesehen, verwenden alle Modellierungsknoten die Feldinformationen des oberhalb liegenden Typknotens. Wenn Sie einen Typknoten verwenden, um Eingabe- und Zielfelder auszuwählen, brauchen Sie auf dieser Registerkarte keine Änderungen vorzunehmen. (Eine Ausnahme bilden der Sequenzknoten und der Textextraktionsknoten, deren Feldeinstellungen im Modellierungsknoten angegeben sein müssen.)

Typknoteneinstellungen verwenden. Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

Benutzerdefinierte Einstellungen verwenden. Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option wie erforderlich die unten stehenden Felder an.

Hinweis: Nicht alle Felder werden für alle Knoten angezeigt.

- **Transaktionsformat verwenden (nur Apriori-, CARMA-, MS-Assoziationsregeln und Oracle Apriori-Knoten).** Aktivieren Sie dieses Kontrollkästchen, wenn die Quellendaten im **Transaktionsformat** vorliegen. Datensätze in diesem Format enthalten zwei Felder, eines für eine ID und eines für den Inhalt. Jeder Datensatz steht für ein einzelnes Element. Zugeordnete Elemente werden verknüpft, indem sie dieselbe ID erhalten. Inaktivieren Sie dieses Feld, wenn die Daten im **Tabellenformat** vorliegen, in dem Elemente durch separate Flags repräsentiert werden, wobei jedes Flagfeld für das Vorhandensein oder

die Abwesenheit eines bestimmten Elements steht und jeder Datensatz ein vollständiges Set an zugehörigen Elementen repräsentiert. Weitere Informationen finden Sie im Thema „Tabellendaten im Vergleich zu Transaktionsdaten" auf Seite 232.

  - **ID.** Wählen Sie für Transaktionsdaten ein ID-Feld aus der Liste aus. Als ID-Feld können numerische oder symbolische Felder verwendet werden. Jeder eindeutige Wert in diesem Feld sollte eine bestimmte Analyseeinheit darstellen. Bei einer Warenkorbanwendung könnte z. B. jede ID einen einzelnen Kunden darstellen. Für eine Webprotokoll-Analyseanwendung könnte jede ID einen Computer (nach IP-Adresse) oder einen Benutzer (nach Anmeldedaten) darstellen.
  - **IDs sind zusammenhängend.** (Nur Apriori- und CARMA-Knoten) Wenn Ihre Daten vorsortiert sind, sodass alle Datensätze mit derselben ID im Datenstream zusammengefasst sind, wählen Sie diese Option, um die Verarbeitung zu beschleunigen. Wenn Ihre Daten nicht vorsortiert sind (oder Sie nicht sicher sind), lassen Sie diese Option inaktiviert. Die Daten werden dann vom Knoten automatisch sortiert.

    Hinweis: Wenn Ihre Daten nicht sortiert sind und Sie diese Option auswählen, erhalten Sie möglicherweise ungültige Ergebnisse in Ihrem Modell.
  - **Inhalt.** Geben Sie die Inhaltsfelder für das Modell an. Diese Felder enthalten die Elemente, die für die Assoziationsmodellierung interessant sind. Sie können mehrere Flagfelder angeben (falls die Daten in tabellarischer Form vorliegen) oder ein einzeln nominales Feld (falls die Daten im Transaktionsformat vorliegen).

- **Ziel.** Wählen Sie die Zielfelder für Modelle aus, die eines oder mehrere Zielfelder benötigen. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Ziel* festlegen.
- **Auswertung.** (Nur für Modelle vom Typ "Autom. Cluster".) Für Clustermodelle ist kein Ziel angegeben. Sie können jedoch ein Evaluierungsfeld auswählen, um das Wichtigkeitsniveau zu ermitteln. Darüber hinaus können Sie evaluieren, wie gut die Cluster Werte dieses Felds differenzieren. Dies wiederum gibt an, ob die Cluster zur Vorhersage dieses Felds verwendet werden können.
  - **Eingaben.** Wählen Sie das/die Eingabefeld(er) aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.
  - **Partition.** In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer anderen Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln. Wenn mehrere Partitionsfelder mithilfe von Typ- oder Partitionsknoten erstellt wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.) Beachten Sie außerdem, dass die Partitionierung auch auf der Registerkarte "Modelloptionen" des Knotens aktiviert sein muss, wenn die ausgewählte Partitionierung in Ihrer Analyse angewendet werden soll. (Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung ohne Änderung der Feldeinstellungen inaktiviert werden.)
- **Aufteilungen.** Wählen Sie für Aufteilungsmodelle das Aufteilungsfeld bzw. die Aufteilungsfelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Aufteilung* festlegen. Sie können nur Felder mit einem Messniveau **Flag**, **Nominal**, **Ordinal** oder **Stetig** als Aufteilungsfelder festlegen. Als Aufteilungsfelder gewählte Felder können nicht als Ziel-, Prädiktor-, Partitions-, Häufigkeits oder Gewichtungsfelder verwendet werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.
- **Häufigkeitsfeld verwenden.** Mit dieser Option können Sie ein Feld als Häufigkeitsgewichtung auswählen. Dies sollten Sie tun, wenn die in Ihren Trainingsdaten enthaltenen Datensätze jeweils mehr als eine Einheit darstellen - wenn Sie beispielsweise aggregierte Daten verwenden. Die Feldwerte sollten die Anzahl der Einheiten sein, die von jedem Datensatz repräsentiert werden. Weitere Informationen finden Sie im Thema „Verwenden von Häufigkeits- und Gewichtungsfeldern" auf Seite 33.

Hinweis: Wenn die Fehlernachricht **Metadaten (in Eingabe-/Ausgabefeldern) nicht gültig** angezeigt wird, stellen Sie sicher, dass Sie alle erforderlichen Felder angegeben haben, wie beispielsweise das Häufigkeitsfeld.

- **Gewichtungsfeld verwenden.** Mit dieser Option können Sie ein Feld als Fallgewichtung auswählen. Fallgewichtungen werden verwendet, um Differenzen in der Varianz zwischen den Ebenen des Ausgabefelds zu berücksichtigen. Weitere Informationen finden Sie im Thema „Verwenden von Häufigkeits- und Gewichtungsfeldern".
- **Sukzedenzien.** Wählen Sie für Regelinduktionsknoten (Apriori) die Felder aus, die im resultierenden Regelset als Sukzedenzien verwendet werden sollen. (Dies entspricht den in einem Typknoten vorhandenen Feldern mit der Rolle *Ziel* oder *Beides*.)
- **Antezedenzien.** Wählen Sie bei Regelinduktionsknoten (Apriori) die Felder aus, die im resultierenden Regelset als Antezedenzien verwendet werden sollen. (Dies entspricht den in einem Typknoten vorhandenen Feldern mit der Rolle *Eingabe* oder *Beides*.)

Bei einigen Modellen weicht die Registerkarte "Felder" von den in diesem Abschnitt beschriebenen ab.

- Weitere Informationen finden Sie im Thema „Feldoptionen für den Sequenzknoten" auf Seite 248.
- Weitere Informationen finden Sie im Thema „Feldoptionen für den CARMA-Knoten" auf Seite 236.

Verwenden von Häufigkeits- und Gewichtungsfeldern

Häufigkeits- und Gewichtungsfelder dienen dazu, einigen Datensätzen eine größere Bedeutsamkeit zu verleihen als anderen, beispielsweise, weil Sie wissen, dass ein Bevölkerungsteil in den Trainingsdaten unterrepräsentiert ist (Gewichtung) oder weil ein Datensatz für eine Reihe identischer Fälle steht (Häufigkeit).

- Werte für ein Häufigkeitsfeld sollten positive ganze Zahlen sein. Datensätze mit einer Häufigkeitsgewichtung kleiner oder gleich 0 werden von der Analyse ausgeschlossen. Nicht als ganze Zahlen angegebene Häufigkeitsgewichtungen werden auf die nächstliegende ganze Zahl gerundet.
- Fallgewichtungswerte müssen als positive Zahlen angegeben werden, müssen aber keine ganzen Zahlen sein. Datensätze mit einer Fallgewichtung kleiner oder gleich 0 werden von der Analyse ausgeschlossen.

Scoren von Häufigkeits- und Gewichtungsfeldern

Häufigkeits- und Gewichtungsfelder werden beim Trainieren von Modellen verwendet, nicht jedoch beim Scoren, da der Score für die einzelnen Datensätze auf den jeweiligen Merkmalen des Datensatzes beruht, unabhängig davon, wie viele Fälle er umfasst. Nehmen Sie beispielsweise an, dass die Daten in der folgenden Tabelle vorliegen.

Tabelle 1. Datenbeispiel

Verheiratet	**Antwort**
Ja	Ja
Ja	Ja
Ja	Ja
Ja	Nein
Nein	Ja
Nein	Nein
Nein	Nein

Auf dieser Grundlage schließen Sie, dass drei von vier verheirateten Personen auf die Werbeaktion antworten und zwei von drei unverheirateten Personen nicht geantwortet haben. Daher scoren Sie alle neuen Datensätze entsprechend wie in der folgenden Tabelle dargestellt.

Tabelle 2. Beispiel für gescorte Datensätze

Verheiratet	**$-Antwort**	**$RP-Antwort**
Ja	Ja	0,75 (drei/vier)
Nein	Nein	0,67 (zwei/drei)

Alternativ könnten Sie Ihre Trainingsdaten mithilfe eines Häufigkeitsfelds in kompakterer Form speichern, wie in der folgenden Tabelle dargestellt.

Tabelle 3. Alternativbeispiel für gescorte Datensätze

Verheiratet	**Antwort**	**Häufigkeit**
Ja	Ja	3
Ja	Nein	E
Nein	Ja	E
Nein	Nein	Z

Da dies für genau dasselbe Dataset steht, erstellen Sie damit dasselbe Modell und sagen die Antworten ausschließlich auf der Grundlage des Ehestandes voraus. Wenn Ihre Scoring-Daten zehn verheiratete Personen enthalten, sagen Sie für alle jeweils *Ja* voraus, unabhängig davon, ob sie als zehn separate Datensätze vorgelegt werden oder als ein einziger Datensatz mit der Häufigkeit 10. Die Gewichtung ist zwar im Allgemeinen keine ganze Zahl, aber zeigt dennoch in ähnlicher Weise die Bedeutsamkeit eines Datensatzes an. Daher werden Häufigkeits- und Gewichtungsfelder beim Scoren von Datensätzen nicht verwendet.

Evaluation und Vergleich von Modellen

Einige Modelltypen unterstützen Häufigkeitsfelder, einige Gewichtungsfelder und einige beide Arten von Feldern. In allen Fällen, in denen sie zulässig sind, werden sie jedoch ausschließlich für die Modellerstellung verwendet und bei der Evaluation von Modellen mithilfe eines Evaluierungs- oder Analyseknotens nicht verwendet. Ebenso wenig verwendet werden sie bei der Rangeinteilung von Modellen mithilfe der meisten von den Knoten vom Typ "Autom. Klassifikationsmerkmal" und "Autonumerisch" unterstützten Methoden.

- Beim Vergleichen von Modellen (beispielsweise mithilfe von Evaluierungsdiagrammen) werden Häufigkeits- und Gewichtungswerte ignoriert. Hierdurch wird zwar ein Niveauvergleich zwischen Modellen, die diese Felder verwenden, und Modellen ohne diese Felder möglich, es bedeutet jedoch auch, dass für eine genaue Evaluierung ein Dataset verwendet werden muss, das die Grundgesamtheit genau darstellt, ohne dass dafür auf ein Häufigkeits- oder Gewichtungsfeld zurückgegriffen werden muss. In der Praxis können Sie dies tun, indem Sie sicherstellen, dass die Modelle mithilfe einer Teststichprobe evaluiert werden, in der der Wert des Häufigkeitsfelds immer null oder 1 ist. (Diese Einschränkung gilt nur bei der Evaluation von Modellen; wenn die Häufigkeits- bzw. Gewichtungswerte sowohl für die Trainings- als auch für die Teststichprobe stets 1 wären, gäbe es keinen Grund, diese Felder überhaupt zu verwenden.)
- Bei Verwendung von "Autom. Klassifikationsmerkmal" kann die Häufigkeit berücksichtigt werden, wenn die Modelle auf der Grundlage des Profits in Ränge eingeteilt werden. In diesem Fall wird also diese Methode empfohlen.
- Falls erforderlich, können Sie die Daten mithilfe eines Partitionsknotens in Trainings- und Teststichproben aufspalten.

Analyseoptionen bei Modellierungsknoten

Zahlreiche Modellierungsknoten enthalten die Registerkarte "Analysieren", mit der Sie Informationen zum Prädiktoreinfluss sowie Raw- und Adjusted-Propensity-Scores abrufen können.

Modellevaluation

Prädiktoreinfluss berechnen. Bei Modellen, die zu einem angemessenen Maß an Bedeutsamkeit führen, können Sie ein Diagramm anzeigen, in dem der relative Einfluss der einzelnen Prädiktoren bei der Modellschätzung angegeben wird. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und diejenigen Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass die Berechnung des Prädiktoreinflusses bei einigen Modellen längere Zeit in Anspruch nehmen kann, insbesondere bei der Arbeit mit großen Datasets, und daher bei einigen Modellen standardmäßig inaktiviert ist. Der Prädiktoreinfluss ist für Entscheidungslistenmodelle nicht verfügbar. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Propensity-Scores

Propensity-Scores können im Modellierungsknoten oder auf der Registerkarte "Einstellungen" im Modellnugget aktiviert werden. Diese Funktionalität ist nur verfügbar, wenn das ausgewählte Ziel ein Flagfeld ist. Weitere Informationen finden Sie im Thema „Propensity-Scores" auf Seite 36.

Raw-Propensity-Scores berechnen. Raw-Propensity-Scores werden ausschließlich auf der Grundlage der Trainingsdaten aus dem Modell abgeleitet. Wenn das Modell den Wert *wahr* (wird antworten) vorhersagt, ist die Neigung mit P identisch. Dabei ist P die Wahrscheinlichkeit der Vorhersage. Wenn das Modell den Wert "falsch" vorhersagt, wird die Neigung als (1 - P) berechnet.

- Wenn Sie bei der Modellerstellung diese Option auswählen, werden standardmäßig Propensity-Scores im Modellnugget aktiviert. Sie können jedoch immer festlegen, dass Raw-Propensity-Scores im Modellnugget aktiviert werden sollen, unabhängig davon, ob Sie sie im Modellierungsknoten auswählen.
- Beim Scoren des Modells werden Raw-Propensity-Scores in einem Feld hinzugefügt, bei dem die Buchstaben *RP* an das Standardpräfix angehängt sind. Wenn sich die Vorhersagen beispielsweise im Feld *$R-Abwanderung* befinden, lautet der Name des Felds für den Propensity-Score *$RRP-Abwanderung*.

Adjusted-Propensity-Scores berechnen. Raw Propensitys basieren ausschließlich auf vom Modell angegebenen Schätzungen. Beim Modell kann jedoch eine Überanpassung vorliegen, was zu übermäßig optimistischen Schätzungen für die Neigung führt. Adjusted Propensitys versuchen, dies zu kompensieren, indem untersucht wird, wie leistungsfähig das Modell bei den Test- bzw. Validierungspartitionen ist, und die Neigungen entsprechend angepasst werden, um eine bessere Schätzung zu erzielen.

- Diese Einstellung ist nur möglich, wenn ein gültiges Partitionsfeld im Stream vorhanden ist.
- Anders als rohe Konfidenzscores müssen Adjusted-Propensity-Scores bei der Erstellung des Modells berechnet werden; andernfalls sind sie beim Scoren des Modellnuggets nicht verfügbar.
- Beim Scoren des Modells werden Adjusted-Propensity-Scores in einem Feld hinzugefügt, bei dem die Buchstaben *AP* an das Standardpräfix angehängt sind. Wenn sich die Vorhersagen beispielsweise im Feld *$R-Abwanderung* befinden, lautet der Name des Felds für den Propensity-Score *$RAP-Abwanderung*. Adjusted-Propensity-Scores sind bei logistischen Regressionsmodellen nicht verfügbar.
- Bei der Berechnung der Adjusted-Propensity-Scores darf die für die Berechnung verwendete Test- bzw. Validierungspartition nicht ausbalanciert worden sein. Um dies zu vermeiden, müssen Sie darauf achten, dass in etwaigen weiter oben im Stream befindlichen Balancierungsknoten die Option **Balancierung nur für Trainingsdaten durchführen** ausgewählt wurde. Zusätzlich gilt: Wenn eine komplexe Stichprobe gezogen wurde, werden dadurch die Adjusted-Propensity-Scores ungültig.
- Adjusted-Propensity-Scores sind bei verstärkten Baum- und Regelsetmodellen nicht verfügbar. Weitere Informationen finden Sie im Thema „Verbesserte C5.0-Modelle" auf Seite 113.

Basierend auf. Um Adjusted-Propensity-Scores berechnen zu können, muss im Stream ein Partitionsfeld vorhanden sein. Sie können angeben, ob die Test- bzw. Validierungspartition für diese Berechnung verwendet werden soll. Um bestmögliche Ergebnisse zu erzielen, sollte die Test- bzw. Validierungspartition mindestens so viele Datensätze enthalten wie die Partition, die zum Trainieren des ursprünglichen Modells verwendet wurde.

Propensity-Scores

Bei Modellen, die eine Vorhersage mit den Werten *Ja* und *Nein* ergeben, können Sie neben den standardmäßigen Vorhersage- und Konfidenzwerten auch Propensity-Scores anfordern. Propensity-Scores geben die Wahrscheinlichkeit eines bestimmten Ergebnisses bzw. einer bestimmten Antwort an. Die folgende Tabelle enthält ein Beispiel.

Tabelle 4. Propensity-Scores

Kunde	**Neigung, zu antworten**
Karl Schmidt	35 %
Marion Schmidt	15 %

Propensity-Scores sind nur bei Modellen mit Flagzielen verfügbar und geben die Wahrscheinlichkeit des für das Feld definierten *Wahr*-Werts an, der in einem Quellen- oder Typknoten angegeben wurde.

Propensity-Scores im Vergleich zu Konfidenzscores

Propensity-Scores unterscheiden sich von Konfidenzscores, die auf die aktuelle Vorhersage angewendet werden, unabhängig, ob der Wert *Ja* oder *Nein* ist. In Fällen, in denen die Vorhersage *Nein* ist, entspricht eine hohe Konfidenz beispielsweise einer hohen Wahrscheinlichkeit *nicht* zu antworten. Propensity-Scores umgehen diese Einschränkung, um einen einfacheren Vergleich zwischen allen Datensätzen zu ermöglichen. So wird eine *Nein*-Vorhersage mit der Konfidenz *0,85* zu einer Raw Propensity von *0,15* (d. h. *1 minus 0,85*).

Tabelle 5. Konfidenzscores

Kunde	**Vorhersage**	**Konfidenz**
Karl Schmidt	Wird antworten	,35
Marion Schmidt	Wird nicht antworten	,85

Propensity-Scores

- Propensity-Scores können auf der Registerkarte "Analysieren" im Modellierungsknoten oder auf der Registerkarte "Einstellungen" im Modellnugget aktiviert werden. Diese Funktionalität ist nur verfügbar, wenn das ausgewählte Ziel ein Flagfeld ist. Weitere Informationen finden Sie im Thema „Analyseoptionen bei Modellierungsknoten” auf Seite 35.
- Propensity-Scores können je nach der verwendeten Ensemble-Methode auch vom Ensemble-Knoten berechnet werden.

Berechnen der Adjusted-Propensity-Scores

Adjusted-Propensity-Scores werden im Rahmen der Modellerstellung berechnet und sind ansonsten nicht verfügbar. Nach der Erstellung des Modells wird es mithilfe von Daten aus der Test- oder Validierungspartition gescort und es wird ein neues Modell erstellt, dass die angepassten Propensity-Scores bereitstellen soll. Dazu wird die Leistung des ursprünglichen Modells auf dieser Partition analysiert. Je nach Modelltyp kann eine von zwei Methoden zur Berechnung der Adjusted-Propensity-Scores verwendet werden.

- Bei Regelset- und Baummodellen werden Adjusted-Propensity-Scores durch erneute Berechnung der Häufigkeit der einzelnen Kategorien an jedem Baumknoten (bei Baummodellen) bzw. der Unterstüt-

zung und Konfidenz der einzelnen Regeln (bei Regelsetmodellen) generiert. Die führt zu einem neuen Regelset- bzw. Baummodell, das zusammen mit dem ursprünglichen Modell gespeichert wird, um immer dann verwendet zu werden, wenn Adjusted-Propensity-Scores angefordert werden. Jedes Mal, wenn das ursprüngliche Modell auf neue Daten angewendet wird, kann das neue Modell anschließend auf die Raw-Propensity-Scores angewendet werden, um die korrigierten Scores zu generieren.

- Bei anderen Modellen werden die durch Scoren des ursprünglichen Modells anhand der Test- bzw. Validierungspartition erstellten Datensätze anschließend nach ihrem Raw-Propensity-Score klassiert. Als Nächstes wird ein neuronales Netzmodell trainiert, das eine nicht lineare Funktion definiert, das eine Zuordnung zwischen der Raw Propensity in den einzelnen Klassen und der mittleren beobachteten Neigung in derselben Klasse erstellt. Wie zuvor für Baummodelle angemerkt, wird das resultierende neuronale Netzmodell zusammen mit dem ursprünglichen Modell gespeichert und kann jedes Mal auf die Raw-Propensity-Scores angewendet werden, wenn Adjusted-Propensity-Scores angefordert werden.

Vorsicht bei fehlenden Werten in der Testpartition. Die Verarbeitung fehlender Eingabewerte in der Test-/Validierungspartition variiert je nach Modell (Details hierzu finden Sie in den einzelnen Algorithmen für das Modellscoring). Das C5-Modell kann Adjusted Propensitys bei fehlenden Eingaben nicht berechnen.

Modellnuggets

Abbildung 19. Modellnugget

Ein Modelnugget ist ein Behälter für ein Modell, d. h. das Set von Regeln, Formeln oder Gleichungen, die das Ergebnis Ihrer Operationen zur Modellerstellung in IBM SPSS Modeler repräsentieren. Ein Nugget dient hauptsächlich zum Scoren von Daten, zum Generieren von Vorhersagen oder zum Ermöglichen einer weiteren Analyse der Modelleigenschaften. Durch Öffnen eines Modellnuggets am Bildschirm können Sie verschiedene Details zum Modell wie z. B. die relative Wichtigkeit der Eingabefelder beim Erstellen des Modells sehen. Zur Anzeige der Vorhersagen müssen Sie einen weiteren Prozess oder Ausgabeknoten anfügen und ausführen. Weitere Informationen finden Sie im Thema „Verwendung von Modellnuggets in Streams" auf Seite 48.

Abbildung 20. Modellverknüpfung vom Modellierungsknoten zum Modellnugget

Wenn Sie einen Modellierungsknoten erfolgreich ausführen, wird ein entsprechendes Modellnugget auf den Streamerstellungsbereich platziert, wo es durch ein goldfarbenes Rautensymbol repräsentiert wird (daher der Name "Nugget"). Im Streamerstellungsbereich wird das Nugget mit einer Verbindung (einer durchgehenden Linie) zum nächsten passenden Knoten vor dem Modellierungsknoten sowie einer Verknüpfung (einer gepunkteten Linie) zum Modellierungsknoten angezeigt.

Das Nugget wird auch in die Modellpalette in der rechten oberen Ecke des IBM SPSS Modeler-Fensters platziert. An beiden Positionen können Nuggets ausgewählt und durchsucht werden, um Details des Modells anzuzeigen.

Nuggets werden stets in der Modellpalette platziert, nachdem ein Modellierungsknoten erfolgreich ausgeführt wurde. Sie können eine Benutzeroption festlegen, die steuert, ob das Nugget zusätzlich im Streamerstellungsbereich platziert wird.

In den folgenden Themenabschnitten finden Sie Informationen zur Verwendung von Modellnuggets in IBM SPSS Modeler. Wenn Sie ein tieferes Verständnis der verwendeten Algorithmen wünschen, lesen Sie im *IBM SPSS Modeler-Algorithmushandbuch* nach, das Sie im Ordner *\Documentation* auf der DVD für IBM SPSS Modeler finden.

Modellverknüpfungen

Standardmäßig wird ein Nugget im Erstellungsbereich mit einer Verknüpfung zu dem Modellierungsknoten angezeigt, der das Nugget erstellt hat. Dies ist vor allem in komplexen Streams mit mehreren Nuggets nützlich und ermöglicht Ihnen das Nugget zu identifizieren, das von jedem Modellierungsknoten aktualisiert wird. Jede Verknüpfung enthält ein Symbol, um anzuzeigen, ob das Modell beim Ausführen des Modellierungsknotens ersetzt wird. Weitere Informationen finden Sie im Thema „Ersetzen eines Modells" auf Seite 39.

Definieren und Entfernen von Modellverknüpfungen

Sie können Verknüpfungen im Erstellungsbereich manuell definieren und entfernen. Wenn Sie eine neue Verknüpfung definieren, ändert der Cursor seine Form zum Verknüpfungscursor.

Abbildung 21. Verknüpfungscursor

Definieren einer neuen Verknüpfung (Kontextmenü)

1. Klicken Sie mit der rechten Maustaste auf den Modellierungsknoten, von dem die Verknüpfung ausgehen soll.
2. Wählen Sie **Modellverknüpfung definieren** aus dem Kontextmenü aus.
3. Klicken Sie auf das Nugget, der das Ende der Verknüpfung darstellen soll.

Definieren einer neuen Verknüpfung (Hauptmenü)

1. Klicken Sie auf den Modellierungsknoten, von dem die Verknüpfung ausgehen soll.
2. Wählen Sie im Hauptmenü Folgendes aus:

   Bearbeiten > **Knoten** > **Modellverknüpfung definieren**
3. Klicken Sie auf das Nugget, der das Ende der Verknüpfung darstellen soll.

Entfernen einer vorhandenen Verknüpfung (Kontextmenü)

1. Klicken Sie mit der rechten Maustaste auf das Nugget am Ende der Verknüpfung.
2. Wählen Sie **Modellverknüpfung entfernen** aus dem Kontextmenü aus.

Alternative:

1. Klicken Sie mit der rechten Maustaste auf das Symbol in der Mitte der Verknüpfung.
2. Wählen Sie **Verknüpfung entfernen** aus dem Kontextmenü aus.

Entfernen einer vorhandenen Verknüpfung (Hauptmenü)

1. Klicken Sie auf den Modellierungsknoten oder das Nugget, von dem Sie die Verknüpfung entfernen möchten.
2. Wählen Sie im Hauptmenü Folgendes aus:

   Bearbeiten > **Knoten** > **Modellverknüpfung entfernen**

Kopieren und Einfügen von Modellverknüpfungen

Wenn Sie ein verknüpftes Nugget ohne seinen Modellierungsknoten kopieren und in denselben Stream einfügen, wird das Nugget mit einer Verknüpfung zum Modellierungsknoten eingefügt. Die neue Verknüpfung weist denselben Modellersetzungsstatus (siehe „Ersetzen eines Modells" auf Seite 39) wie die

ursprüngliche Verknüpfung auf.

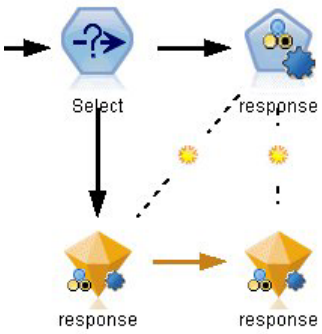


Abbildung 22. Kopieren und Einfügen eines verknüpften Nuggets

Wenn Sie ein Nugget zusammen mit seinem verknüpften Modellierungsknoten kopieren und einfügen, wird die Verknüpfung beibehalten, unabhängig davon, ob die Objekte in denselben oder in einen neuen Stream eingefügt werden.

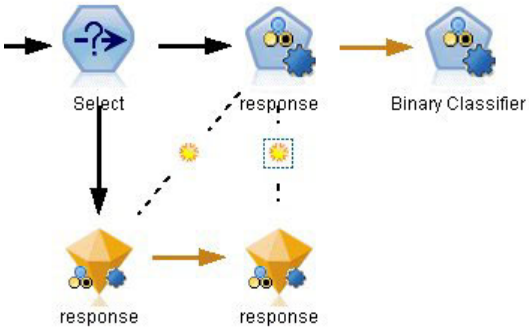


Abbildung 23. Kopieren und Einfügen eines verknüpften Nuggets

Hinweis: Wenn Sie ein verknüpftes Nugget ohne den zugehörigen Modellierungsknoten kopieren und in einen neuen Stream einfügen (oder in einen Superknoten, der den Modellierungsknoten nicht enthält), wird die Verknüpfung getrennt und nur das Nugget wird eingefügt.

Modellverknüpfungen und Superknoten

Wenn Sie einen Superknoten so definieren, dass er entweder den Modellierungsknoten oder das Modellnugget eines verknüpften Modells (jedoch nicht beide) enthält, wird die Verknüpfung unterbrochen. Ein Erweitern des Superknotens stellt die Verknüpfung nicht wieder her. Dies ist nur möglich, indem Sie die Erstellung des Superknotens widerrufen.

Ersetzen eines Modells

Sie können wählen, ob ein bestehendes Nugget bei der Neuausführung des Modellierungsknotens, der das Nugget erstellt hat, ersetzt (d. h. aktualisiert) wird. Wenn Sie die Ersetzungsoption inaktivieren, wird bei der erneuten Ausführung des Modellierungsknotens ein neues Nugget angelegt.

Hinweis: Das Ersetzen eines Modells unterscheidet sich vom Aktualisieren eines Modells, das sich auf das Aktualisieren eines Modells in einem Szenario bezieht.

Jede Verknüpfung vom Modellierungsknoten zum Nugget enthält ein Symbol, um anzuzeigen, ob das Modell beim erneuten Ausführen des Modellierungsknotens ersetzt wird.

Abbildung 24. Modellverknüpfung mit aktivierter Modellersetzung

Die Verknüpfung wird anfangs mit aktivierter Modellersetzung gezeigt, dargestellt durch das kleine Sonnensymbol in der Verknüpfung. In diesem Status wird bei erneuter Ausführung des Modellierungsknotens an einem Ende der Verknüpfung einfach das Nugget am anderen Ende aktualisiert.

Abbildung 25. Modellverknüpfung mit inaktivierter Modellersetzung

Wenn die Modellersetzung inaktiviert ist, wird das Verknüpfungssymbol durch einen grauen Punkt ersetzt. In diesem Status wird bei erneuter Ausführung des Modellierungsknotens an einem Ende der Verknüpfung des Erstellungsbereichs eine neue, aktualisierte Version des Nuggets hinzugefügt.

In beiden Fällen wird das vorhandene Nugget in der Modellpalette aktualisiert oder ein neues Nugget hinzugefügt, abhängig von der Systemoption **Vorheriges Modell ersetzen**.

Reihenfolge der Ausführung

Wenn Sie einen Stream mit mehreren Verzweigungen ausführen, die Modellnuggets enthalten, wird der Stream zunächst evaluiert, um sicherzustellen, dass eine Verzweigung mit aktivierter Modellersetzung vor einer Verzweigung ausgeführt wird, die das resultierende Modellnugget verwendet.

Bei komplexeren Anforderungen können Sie die Ausführungsreihenfolge manuell durch Erstellung eines Scripts festlegen.

Ändern der Modellersetzungseinstellung

So ändern Sie die Einstellung für Modellersetzung:

1. Klicken Sie mit der rechten Maustaste auf das Symbol in der Verknüpfung.
2. Wählen Sie wie gewünscht **Modellersetzung aktivieren (inaktivieren)**.

Hinweis: Die Einstellung zum Ersetzen eines Modells auf einer Modellverknüpfung überschreibt die Einstellung auf der Registerkarte "Benachrichtigungen" des Dialogfelds "Benutzeroptionen" (Tools > Optionen > Benutzeroptionen).

Modellpalette

Mithilfe der Modellpalette (auf der Registerkarte "Modelle" im Managerfenster) können Sie Modellnuggets auf verschiedene Weise verwenden, prüfen und ändern.

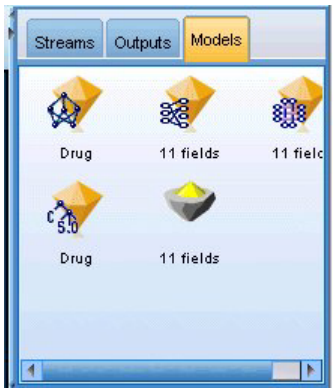


Abbildung 26. Modellpalette

Wenn Sie mit der rechten Maustaste auf ein Modellnugget in der Modellpalette klicken, wird ein Kontextmenü mit folgenden Optionen geöffnet:

- **Zu Stream hinzufügen.** Fügt das generierte Modellnugget zum derzeit aktiven Stream hinzu. Wenn im Stream ein Knoten ausgewählt wurde, wird das Modellnugget mit dem ausgewählten Knoten verbunden, wenn eine solche Verbindung möglich ist. Andernfalls erfolgt die Verbindung zum nächsten möglichen Knoten. Das Nugget wird mit einer Verknüpfung zum Modellierungsknoten gezeigt, von dem aus das Modell erstellt wurde, sofern sich der Knoten noch im Stream befindet.
- **Durchsuchen.** Öffnet den Modellbrowser für das Nugget.
- **Umbenennen und mit Anmerkung versehen.** Ermöglicht das Umbenennen des Modellnuggets und/oder die Bearbeitung der Anmerkung für das Nugget.
- **Modellierungsknoten generieren.** Wenn Sie ein Modellnugget ändern oder aktualisieren möchten und der zum Erstellen des Modells verwendete Stream nicht verfügbar ist, können Sie mithilfe dieser Option erneut einen Modellierungsknoten mit denselben Optionen erzeugen, die zum Erstellen des ursprünglichen Modells verwendet wurden.
- **Modell speichern, Modell speichern als.** Speichert das Modellnugget in einer externen Binärdatei für generierte Modelle (.gm).
- **Modell speichern.** Speichert das Modellnugget in IBM SPSS Collaboration and Deployment Services Repository.
- **PMML exportieren.** Exportiert das Modellnugget als Predictive Model Markup Language (PMML), die zum Scoring neuer Daten außerhalb von IBM SPSS Modeler verwendet werden kann. **PMML exportieren** ist für alle generierten Modellknoten verfügbar. *Hinweis*: Für die Verwendung dieser Funktion ist eine IBM SPSS Modeler Server-Lizenz erforderlich.
- **Zu Projekt hinzufügen.** Speichert das generierte Modellnugget und fügt es zum aktuellen Projekt hinzu. Auf der Registerkarte "Klassen" wird das Nugget zum Ordner "Generierte Modelle" hinzugefügt. Auf der Registerkarte "CRISP-DM" wird es zur Standardprojektphase hinzugefügt.
- **Löschen.** Löscht das Modellnugget aus der Palette.

Wenn Sie mit der rechten Maustaste auf einen nicht belegten Bereich in der Modellpalette klicken, wird ein Kontextmenü mit folgenden Optionen geöffnet:

- **Modell öffnen.** Lädt ein Modellnugget, das zuvor in IBM SPSS Modeler erstellt wurde.
- **Modell abrufen.** Ruft ein gespeichertes Modell aus einem IBM SPSS Collaboration and Deployment Services-Repository ab.
- **Palette laden.** Lädt eine gespeicherte Modellpalette aus einer externen Datei.
- **Palette abrufen.** Ruft eine gespeicherte Modellpalette aus einem IBM SPSS Collaboration and Deployment Services-Repository ab.
- **Palette speichern.** Speichert den gesamten Inhalt der Modellpalette in einer externen Datei für generierte Modellpaletten (.gen).

- **Palette temporär speichern.** Speichert den gesamten Inhalt der Modellpalette in einem IBM SPSS Collaboration and Deployment Services-Repository.
- **Palette löschen.** Löscht alle Nuggets aus der Palette.
- **Palette zu Projekt hinzufügen.** Speichert die Modellpalette und fügt sie dem aktuellen Projekt hinzu. Auf der Registerkarte "Klassen" wird das Nugget zum Ordner "Generierte Modelle" hinzugefügt. Auf der Registerkarte "CRISP-DM" wird es zur Standardprojektphase hinzugefügt.
- **PMML importieren.** Lädt ein Modell aus einer externen Datei. Sie können PMML-Modelle öffnen, durchsuchen und scoren, die von IBM SPSS Statistics oder anderen Anwendungen, die dieses Format unterstützen, erstellt wurden. Weitere Informationen finden Sie im Thema „Importieren und Exportieren von Modellen als PMML" auf Seite 49.

Durchsuchen von Modellnuggets

Mit den Browsern für Modellnuggets können Sie die Ergebnisse Ihrer Modelle prüfen und verwenden. Über den Browser können Sie das generierte Modell speichern, drucken oder exportieren, die Modellübersicht überprüfen und Anmerkungen für das Modell anzeigen oder bearbeiten. Bei einigen Typen von Modellnuggets können Sie auch neue Knoten generieren, beispielsweise Filterknoten oder Regelsetknoten. Bei einigen Modellen können Sie außerdem Modellparameter, wie Regeln oder Clusterzentren, anzeigen. Bei einigen Modelltypen (baumbasierten Modellen und Clustermodellen) können Sie eine grafische Darstellung der Struktur des Modells anzeigen. Die Steuerelemente für die Verwendung der Browser für Modellnuggets sind unten beschrieben.

Menüs

Menü "Datei". Alle Modellnuggets weisen ein Dateimenü auf, das ein Subset der folgenden Optionen enthält:

- **Knoten speichern.** Speichert das Modellnugget in einer Knotendatei (.nod).
- **Knoten temporär speichern.** Speichert das Modellnugget in einem IBM SPSS Collaboration and Deployment Services-Repository.
- **Kopf-/Fußzeile.** Erlaubt die Bearbeitung der Kopf- und Fußzeile der Seite zum Drucken über das Nugget.
- **Seite einrichten.** Erlaubt die Bearbeitung der Seiteneinrichtung zum Drucken über das Nugget.
- **Druckvorschau.** Zeigt als Vorschau an, wie das Nugget beim Ausdruck aussieht. Wählen Sie im Untermenü die Informationen aus, die in der Vorschau angezeigt werden sollen.
- **Drucken.** Druckt den Inhalt des Nuggets. Wählen Sie im Untermenü die Informationen aus, die gedruckt werden sollen.
- **Druckansicht.** Druckt die aktuelle Ansicht oder alle Ansichten.
- **Text exportieren.** Exportiert den Inhalt des Nuggets in eine Textdatei. Wählen Sie im Untermenü die Informationen aus, die exportiert werden sollen.
- **HTML generieren.** Exportiert den Inhalt des Nuggets in eine HTML-Datei. Wählen Sie im Untermenü die Informationen aus, die exportiert werden sollen.
- **PMML exportieren.** Exportiert das Modell als Predictive Model Markup Language (PMML), die mit anderen PMML-kompatiblen Softwareprodukten verwendet werden kann. Weitere Informationen finden Sie im Thema „Importieren und Exportieren von Modellen als PMML" auf Seite 49. *Hinweis*: Für die Verwendung dieser Funktion ist eine IBM SPSS Modeler Server-Lizenz erforderlich.
- **SQL exportieren.** Exportiert das Modell als SQL (Structured Query Language), die mit anderen Datenbanken bearbeitet und verwendet werden kann.

  Hinweis: SQL-Export ist nur über die folgenden Modelle verfügbar: C5, C&R-Baum, CHAID, QUEST, Lineare Regression, Logistische Regression, Neuronales Netz, Faktor und Entscheidungsliste.

- **Für Server-Scoring-Adapter veröffentlichen.** Veröffentlicht das Modell für eine Datenbank mit Scoring-Adapter, sodass das Scoring des Modells innerhalb der Datenbank ausgeführt werden kann. Weitere Informationen finden Sie im Thema „Veröffentlichen von Modellen für einen Scoring-Adapter" auf Seite 51.

Menü generieren. Die meisten Modellnuggets verfügen auch über das Menü "Generieren", mit dem Sie neue Knoten basierend auf dem Modellnugget generieren können. Die über dieses Menü verfügbaren Optionen hängen vom durchsuchten Modelltyp ab. Einzelheiten zu den Elementen, die aus einem bestimmten Modell generiert werden können, finden Sie in den Informationen zum jeweils generierten Modellnuggettyp.

Menü "Ansicht". Auf der Registerkarte "Modell" eines Nugget ermöglicht dieses Menü das Ein- bzw. Ausblenden der verschiedenen Visualisierungs-Symbolleisten, die im aktuellen Modus verfügbar sind. Um alle Symbolleisten verfügbar zu machen, wählen Sie in der Symbolleiste "Allgemein" die Option "Bearbeitungsmodus" (das Pinselsymbol).

Schaltfläche "Vorschau". Einige Modellnuggets verfügen über die Schaltfläche "Vorschau". Sie ermöglicht die Anzeige eines Auszugs der Modelldaten, inklusive der vom Modellierungsprozess erstellten Zusatzfelder. Die Standardanzahl der angezeigten Zeilen ist 10. Sie können diese Einstellung jedoch in den Streameigenschaften ändern.

Schaltfläche "Zu aktuellem Projekt hinzufügen". Speichert das generierte Modellnugget und fügt es zum aktuellen Projekt hinzu. Auf der Registerkarte "Klassen" wird das Nugget zum Ordner "Generierte Modelle" hinzugefügt. Auf der Registerkarte "CRISP-DM" wird es zur Standardprojektphase hinzugefügt.

Modellnuggets - Übersicht/Informationen

Auf der Registerkarte "Übersicht" oder der Informationsansicht für ein Modellnugget werden Informationen über die Felder, die Aufbaueinstellungen und die Modellschätzung angezeigt. Die Ergebnisse werden in einer Baumansicht dargestellt, die durch Klicken auf bestimmte Elemente erweitert bzw. reduziert werden kann.

Analyse. Zeigt Informationen zum Modell an. Die konkreten Details variieren nach Modelltyp und werden jeweils im Abschnitt zu den einzelnen Modellnuggets behandelt. Wenn Sie einen Analyseknoten ausgeführt haben, der an diesen Modellierungsknoten angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt.

Felder. Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden. Listet bei aufgeteilten Modellen außerdem die Felder auf, welche die Aufteilung bestimmen.

Erstellungseinstellungen /-optionen. Enthält Informationen zu den beim Aufbau des Modells verwendeten Einstellungen.

Trainingsübersicht. In diesem Abschnitt werden folgende Informationen angezeigt: der Typ des Modells, der für seine Erstellung verwendete Stream, der Benutzer, der ihn erstellt hat, der Erstellungszeitpunkt und die für den Aufbau des Modells benötigte Zeit.

Bedeutung des Prädiktors

In der Regel konzentriert man sich bei der Modellerstellung auf die Prädiktorfelder, die am wichtigsten sind, und vernachlässigt jene, die weniger wichtig sind. Dabei unterstützt Sie das Wichtigkeitsdiagramm für die Prädiktoren, da es die relative Wichtigkeit der einzelnen Prädiktoren für das Modell angibt. Da die Werte relativ sind, beträgt die Summe der Werte aller Prädiktoren im Diagramm 1,0. Die Wichtigkeit der Prädiktoren steht in keinem Bezug zur Genauigkeit des Modells. Sie bezieht sich lediglich auf die Wichtigkeit der einzelnen Prädiktoren für eine Vorhersage und nicht auf die Genauigkeit der Vorhersage.

Der Prädiktoreinfluss ist für Modelle verfügbar, die ein angemessenes statistisches Maß an Wichtigkeit erzeugen, darunter neuronale Netze, Entscheidungsbäume (C&R-Baum, C5.0, CHAID und QUEST), Bayes-Netze, Diskriminanz, SVM- und SLRM-Modelle, lineare und logistische Regression, verallgemeinerte lineare Modelle und Nächste-Nachbarn-Modelle (KNN). Für die meisten dieser Modelle kann der Prädiktoreinfluss auf der Registerkarte "Analysieren" im Modellierungsknoten aktiviert werden. Weitere Informationen finden Sie im Thema „Analyseoptionen bei Modellierungsknoten" auf Seite 35. Informationen zu KNN-Modellen finden Sie unter „Nachbarn" auf Seite 291.

Hinweis: Der Prädiktoreinfluss wird für Aufteilungsmodelle nicht unterstützt. Die Einstellungen für den Prädiktoreinfluss werden bei der Erstellung von aufgeteilten Modellen ignoriert. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Die Berechnung des Prädiktoreinflusses kann erheblich länger dauern als die Modellerstellung, insbesondere bei großen Datasets. Die Berechnung dauert bei SVM und logistischer Regression länger als bei anderen Modellen und ist daher für diese Modelle standardmäßig inaktiviert. Bei Verwendung eines Datasets mit einer großen Anzahl an Prädiktoren kann ein anfängliches Screening mit einem Merkmalauswahlknoten zu schnelleren Ergebnissen führen (siehe unten).

- Der Prädiktoreinfluss wird aus der Testpartition berechnet, sofern verfügbar. Andernfalls werden die Trainingsdaten verwendet.
- Bei SLRM-Modellen ist der Prädiktoreinfluss verfügbar, sie wird jedoch durch den SLRM-Algorithmus berechnet. Weitere Informationen finden Sie im Thema „SLRM-Modellnuggets" auf Seite 280.
- Mit den Diagrammtools von IBM SPSS Modeler können Sie in das Diagramm eingreifen, es bearbeiten und speichern.
- Optional können Sie anhand der Informationen im Diagramm für den Prädiktoreinfluss einen Filterknoten generieren. Weitere Informationen finden Sie im Thema „Filtern von Variablen auf der Grundlage der Bedeutsamkeit".

Prädiktoreinfluss und Merkmalauswahl

In einigen Fällen kann es so aussehen, als ob das in einem Modellnugget angezeigte Diagramm für den Prädiktoreinfluss zu ähnlichen Ergebnissen führt wie der Merkmalauswahlknoten. Während bei der Merkmalauswahl die einzelnen Eingabefelder hinsichtlich der Stärke ihrer Beziehung zum angegebenen Ziel unabhängig von anderen Eingaben in Ränge eingeteilt werden, gibt das Diagramm für den Prädiktoreinfluss die relative Wichtigkeit der einzelnen Eingaben für *dieses* konkrete Modell an. Daher führt die Merkmalauswahl beim Screening der Eingaben zu einem konservativeren Ergebnis. Wenn beispielsweise sowohl *Berufsbezeichnung* als auch *Berufskategorie* in einem starken Zusammenhang zum Gehalt stehen, zeigt die Merkmalauswahl an, dass beide von Bedeutung sind. Bei der Modellierung werden jedoch auch Interaktionen und Korrelationen berücksichtigt. Dies kann dazu führen, dass nur eine von zwei Eingaben verwendet wird, wenn beide größtenteils dieselben Informationen bieten. In der Praxis ist die Merkmalauswahl am nützlichsten für ein erstes Screening, insbesondere beim Umgang mit großen Datasets mit zahlreichen Variablen, während der Prädiktoreinfluxx bei der Feinabstimmung des Modells nützlicher ist.

Filtern von Variablen auf der Grundlage der Bedeutsamkeit

Optional können Sie anhand der Informationen im Diagramm für den Prädiktoreinfluss einen Filterknoten generieren.

Markieren Sie gegebenenfalls die Prädiktoren, die Sie in das Diagramm aufnehmen möchten, und wählen Sie folgende Optionen aus den Menüs aus:

Generieren > **Filterknoten (Prädiktoreinfluss)**

ODER

> **Feldauswahl (Prädiktoreinfluss)**

Wichtigste Zahl der Variablen. Die wichtigsten Prädiktoren bis zur festgelegten Zahl werden aufgenommen oder ausgeschlossen.

Bedeutsamkeit größer als. Alle Prädiktoren mit einem relativen Einfluss, die größer ist als der festgelegte Wert, werden aufgenommen oder ausgeschlossen.

Ensemble-Viewer

Modelle für Ensembles

Das Modell für ein Ensemble bietet Informationen zu den Komponentenmodellen im Ensemble und zur Leistung des Ensembles als Ganzes.

In der (von der Ansicht unabhängigen) Hauptsymbolleiste können Sie auswählen, ob das Ensemble oder ein Referenzmodell für die Bewertung (Scoring) verwendet werden soll. Wenn das Ensemble für das Scoring verwendet wird, können Sie auch die Kombinationsregel auswählen. Für diese Änderungen ist keine erneute Ausführung des Modells erforderlich. Die getroffene Auswahl wird jedoch zur Bewertung (Scoring) und/oder zur nachfolgenden Modellevaluierung im Modell (Nugget) gespeichert. Außerdem wirkt sie sich auf die aus dem Ensemble-Viewer exportierte PMML aus.

Kombinationsregel. Beim Scoren eines Ensembles wird diese Regel angewendet, um die vorhergesagten Werte aus den Basismodellen für die Berechnung des Scorewerts für das Ensemble zu kombinieren.

- Ensemble-Vorhersagewerte für **kategoriale** Ziele können unter Verwendung von Voting, der höchsten Wahrscheinlichkeit oder der höchsten mittleren Wahrscheinlichkeit kombiniert werden. Mit **Voting**wird die Kategorie ausgewählt, die in der Menge aller Basismodelle am häufigsten die höchste Wahrscheinlichkeit aufweist. Mit **Höchste Wahrscheinlichkeit**wird die Kategorie ausgewählt, die in der Menge aller Basismodelle die höchste Wahrscheinlichkeit überhaupt aufweist. Mit **Höchste mittlere Wahrscheinlichkeit** wird die Kategorie mit dem höchsten Wert ausgewählt, wenn der Mittelwert der Kategoriewahrscheinlichkeiten aus der Menge aller Basismodelle berechnet wird.
- Ensemble-Vorhersagewerte für **stetige** Ziele können unter Verwendung des Mittelwerts oder Medians der Vorhersagewerte aus den Basismodellen kombiniert werden.

Die Standardeinstellung beruht auf den während der Modellerstellung angegebenen Spezifikationen. Durch Ändern der Kombinationsregel wird die Modellgenauigkeit neu berechnet und alle Ansichten der Modellgenauigkeit werden aktualisiert. Das Diagramm für den Prädiktoreinflussdiagramm wird ebenfalls aktualisiert. Dieses Steuerelement ist inaktiviert, wenn das Referenzmodell für die Bewertung (Scoring) ausgewählt wurde.

Alle Kombinationsregeln anzeigen. Bei Auswahl dieser Option werden Ergebnisse für alle verfügbaren Kombinationsregeln im Diagramm zur Modellqualität angezeigt. Das Diagramm" Komponentenmodellgenauigkeit" wird ebenfalls aktualisiert und zeigt nun Bezugslinien für die einzelnen Voting-Methoden.

Modellzusammenfassung: Mit der Ansicht "Modellzusammenfassung" erhalten Sie eine momentane, übersichtliche Zusammenfassung über Ensemble-Qualität und -Diversität.

Qualität. Dieses Diagramm zeigt die Genauigkeit des endgültigen Modells im Vergleich mit einem Referenzmodell und einem naiven Modell. Die Genauigkeit wird nach dem Prinzip "größer ist besser" dargestellt. Das "beste" Modell hat die höchste Genauigkeit. Bei kategorialen Zielen ist die Genauigkeit einfach der Prozentsatz der Datensätze, für den der vorhergesagte Wert mit dem beobachteten Wert übereinstimmt. Bei stetigen Zielen ist die Genauigkeit 1 minus dem Verhältnis zwischen dem mittleren absoluten Fehler bei der Vorhersage (Durchschnitt der Absolutwerte der vorhergesagten Werte minus beobachtete Werte) und dem Bereich der vorhergesagten Werte (größter vorhergesagter Wert minus kleinster vorhergesagter Wert).

Bei Bagging-Ensembles ist das Referenzmodell ein Standardmodell, das auf der gesamten Trainingspartition beruht. Bei Boosting-Ensembles ist das Referenzmodell das Modell der ersten Komponente.

Das naive Modell stellt die Genauigkeit dar, die bestünde, wenn kein Modell erstellt würde, und weist alle Datensätze der Modalkategorie zu. Für stetige Ziele wird das naive Modell nicht berechnet.

Diversität. Das Diagramm zeigt die "Meinungsdiversität" unter den zum Erstellen des Ensembles verwendeten Komponentenmodellen an, dargestellt nach dem Prinzip "größer ist besser". Es handelt sich hierbei um die Messung der Variation der Vorhersagen zwischen den verschiedenen Basismodellen. Für verstärkte Ensemble-Modelle steht die Option "Diversität" nicht zur Verfügung und sie wird auch nicht für stetige Ziele angezeigt.

Prädiktoreinfluss: In der Regel konzentriert man sich bei der Modellerstellung auf die Prädiktorfelder, die am wichtigsten sind, und vernachlässigt jene, die weniger wichtig sind. Dabei unterstützt Sie das Wichtigkeitsdiagramm für die Prädiktoren, da es die relative Wichtigkeit der einzelnen Prädiktoren für das Modell angibt. Da die Werte relativ sind, beträgt die Summe der Werte aller Prädiktoren im Diagramm 1,0. Die Wichtigkeit der Prädiktoren steht in keinem Bezug zur Genauigkeit des Modells. Sie bezieht sich lediglich auf die Wichtigkeit der einzelnen Prädiktoren für eine Vorhersage und nicht auf die Genauigkeit der Vorhersage.

Der Prädiktoreinfluss steht nicht für alle Ensemble-Modelle zur Verfügung Die Menge der Prädiktoren kann zwischen verschiedenen Komponentenmodellen variieren, die Wichtigkeit kann jedoch für Prädiktoren berechnet werden, die in mindestens einem Komponentenmodell verwendet werden.

Prädiktorhäufigkeit: Die Menge der Prädiktoren kann aufgrund der Auswahl der verwendeten Modellierungsmethode bzw. der Auswahl der Prädiktoren zwischen verschiedenen Komponentenmodellen variieren. Das Diagramm "Prädiktorhäufigkeit" ist ein Punktdiagramm, das die Verteilung der Prädiktoren in den verschiedenen Komponentenmodellen im Ensemble zeigt. Jeder Punkt steht für eine oder mehrere Komponenten, die den Prädiktor enthält/enthalten. Prädiktoren werden auf der y-Achse dargestellt und in absteigender Reihenfolge ihrer Häufigkeit sortiert. Somit ist der oberste Prädiktor derjenige, der in der größten Anzahl an Komponentenmodellen verwendet wurde, und der unterste derjenige, der in den wenigsten Modellen verwendet wurde. Die ersten 10 Prädiktoren werden angezeigt.

Die am häufigsten vorkommenden Prädiktoren sind normalerweise auch die wichtigsten. Dieses Diagramm ist nicht brauchbar bei Methoden, bei denen die Menge der Prädiktoren zwischen den verschiedenen Komponentenmodellen variieren kann.

Komponentenmodellgenauigkeit: Bei diesem Diagramm handelt es sich um ein Punktdiagramm der Vorhersagegenauigkeit für Komponentenmodelle. Jeder Punkt steht für ein oder mehrere Komponentenmodelle, wobei der Genauigkeitsgrad auf der y-Achse dargestellt wird. Fahren Sie mit der Maus über einen Punkt, um Informationen zum zugehörigen Einzelkomponentenmodell abzurufen.

Bezugslinien. In diesem Diagramm werden farbcodierte Linien für das Ensemble sowie für das Referenzmodell und die naiven Modelle angezeigt. Neben der Linie, die zu dem für die Bewertung (Scoring) verwendeten Modell gehört, wird ein Häkchen angezeigt.

Interaktivität. Das Diagramm wird aktualisiert, wenn Sie die Kombinationsregel ändern.

Verstärkte Ensembles. Für verstärkte Ensembles wird ein Liniendiagramm angezeigt.

Komponentenmodelldetails: In der Tabelle werden Informationen zu Komponentenmodellen, nach Zeile aufgelistet, angezeigt. Standardmäßig werden die Komponentenmodelle in aufsteigender Reihenfolge nach der Modellnummer sortiert. Sie können die Zeilen in aufsteigender oder absteigender Reihenfolge nach den Werten jeder beliebigen Spalte sortieren.

Modell. Eine Nummer, die die Reihenfolge angibt, in der das Komponentenmodell erstellt wurde.

Genauigkeit. Als Prozentwert angegebene Gesamtgenauigkeit.

Methode. Die Modellierungsmethode.

Prädiktoren. Die Anzahl der im Komponentenmodell verwendeten Prädiktoren.

Modellgröße. Die Modellgröße hängt von der Modellierungsmethode ab: Bei Bäumen handelt es sich um die Anzahl der Knoten im Baum, bei linearen Modellen um die Anzahl der Koeffizienten, bei neuronalen Netzen um die Anzahl der Synapsen.

Datensätze. Die gewichtete Anzahl an Eingabedatensätzen in der Trainingsstichprobe.

Automatische Datenaufbereitung:

Diese Ansicht zeigt Informationen darüber an, welche Felder ausgeschlossen wurden und wie transformierte Felder im Schritt "automatische Datenaufbereitung" (ADP) abgeleitet wurden. Für jedes transformierte oder ausgeschlossene Feld listet die Tabelle den Feldnamen, die Rolle in der Analyse und die im ADP-Schritt vorgenommene Aktion auf. Die Felder werden in aufsteigender alphabetischer Reihenfolge der Feldnamen sortiert.

Wenn die Aktion **Ausreißer trimmen** angezeigt wird, bedeutet dies, dass Werte stetiger Prädiktoren, die über einem Trennwert liegen (drei Standardabweichungen vom Mittelwert), auf den Trennwert gesetzt wurden.

Modellnuggets für aufgeteilte Modelle

Das Nugget für ein aufgeteiltes Modell bietet Zugriff auf alle einzelnen Modelle, die durch die Aufteilungen entstanden sind.

Ein Nugget für aufgeteilte Modelle enthält:

- eine Liste der erstellten aufgeteilten Modelle zusammen mit einem Set von Statistikdaten zu jedem Modell
- Informationen zum Gesamtmodell

In der Liste der aufgeteilten Modelle können Sie einzelne Modelle öffnen, um sie weiter zu untersuchen.

Modellviewer aufteilen

Auf der Registerkarte "Modell" sind alle Modelle aufgelistet, die in einem Nugget enthalten sind, und sie enthält zahlreiche Statistikdaten über die aufgeteilten Modelle. Je nach Modellierungsknoten hat sie zwei allgemeine Formen.

Sortieren nach. Verwenden Sie diese Liste, um die Reihenfolge der aufgeführten Modelle zu wählen. Sie können die Liste auf der Basis der Werte in einer der angezeigten Spalten aufsteigend oder absteigend sortieren. Alternativ können Sie auf die Überschrift einer Spalte klicken, um eine Sortierung nach dieser Spalte vorzunehmen. Standard ist absteigende Sortierung mit Gesamtgenauigkeit.

Menü "Spalten anzeigen/ausblenden". Klicken Sie auf diese Schaltfläche, um ein Menü zu öffnen, in dem Sie einzelne Spalten zum Anzeigen oder Ausblenden wählen können.

Ansicht. Wenn Sie Partitionierung verwenden, können Sie die Ergebnisse für die Trainings- bzw. Testdaten anzeigen lassen.

Für jede Aufteilung werden die folgenden Daten angezeigt:

Diagramm. Ein Piktogramm, das die Datenverteilung für dieses Modell anzeigt. Wenn sich das Nugget im Erstellungsbereich befindet, wird das Diagramm durch Doppelklicken auf das Piktogramm in voller Größe geöffnet.

Modell. Ein Symbol des Modelltyps. Doppelklicken Sie auf das Symbol, um das Modellnugget für diese bestimmte Aufteilung zu öffnen.

Aufteilungsfelder. Die Felder, die im Modellierungsknoten als Aufteilungsfelder festgelegt sind, sowie Ihre möglichen Werte.

Anzahl der Datensätze in der Aufteilung. Die Anzahl von Datensätzen, die an dieser bestimmten Aufteilung beteiligt sind.

Anzahl verwendeter Felder. Teilt aufgeteilte Modelle auf der Grundlage der verwendeten Eingabefelder in Ränge ein.

Gesamtgenauigkeit (%). Der Prozentsatz der Datensätze, der korrekt vom aufgeteilten Modell vorhergesagt wird, im Verhältnis zur Gesamtzahl der Datensätze in dieser Aufteilung.

Aufteilen. Die Spaltenüberschrift zeigt das Feld bzw. die Felder, die zum Erstellen der Aufteilungen verwendet wurden, und die Zellen enthalten die Aufteilungswerte. Doppelklicken Sie auf eine Aufteilung, um einen Modellviewer für das für die betreffende Aufteilung erstellte Modell anzuzeigen.

Genauigkeit. Als Prozentwert angegebene Gesamtgenauigkeit.

Modellgröße. Die Modellgröße hängt von der Modellierungsmethode ab: Bei Bäumen handelt es sich um die Anzahl der Knoten im Baum, bei linearen Modellen um die Anzahl der Koeffizienten, bei neuronalen Netzen um die Anzahl der Synapsen.

Datensätze. Die gewichtete Anzahl an Eingabedatensätzen in der Trainingsstichprobe.

Verwendung von Modellnuggets in Streams

Modellnuggets werden in Streams platziert, damit Sie neue Daten scoren und neue Knoten generieren können. Beim **Scoring** von Daten können Sie die aus der Modellerstellung gewonnenen Informationen verwenden, um Vorhersagen für neue Datensätze zu erstellen. Zur Anzeige der Scoring-Ergebnisse müssen Sie dem Nugget einen Endknoten hinzufügen (d. h. einen Verarbeitungs- oder Ausgabeknoten) und den Endknoten ausführen.

Bei einigen Modellen bieten Modellnuggets auch weitere Informationen über die Qualität der Vorhersage, beispielsweise Konfidenzwerte oder Entfernungen von den Clusterzentren. Durch das Generieren neuer Knoten können Sie ganz einfach neue Knoten basierend auf der Struktur des generierten Modells erstellen. So ermöglichen beispielsweise die meisten Modelle, die eine Eingabefeldauswahl durchführen, die Erstellung von Filterknoten, die nur Eingabefelder übergeben, die das Modell als wichtig ermittelt hat.

So verwenden Sie ein Modellnugget zum Scoren von Daten:

1. Verbinden Sie das Modellnugget mit einer Datenquelle oder einem Stream, der ihm Daten übergeben soll.
2. Fügen Sie einen oder mehrere Verarbeitungs- oder Ausgabeknoten (beispielsweise einen Tabellen- oder Analyseknoten) zum Modellnugget hinzu oder verbinden Sie sie damit.
3. Führen Sie einen der Knoten unterhalb des Modellnuggets aus.

Hinweis: Sie können den Knoten "Nicht verfeinerte Regel" nicht zum Scoren von Daten verwenden. Um Daten basierend auf einem Assoziationsregelmodell zu scoren, verwenden Sie den Knoten "Nicht verfei-

nerte Regel", um ein Regelsetnugget zu generieren, und verwenden Sie das Regelsetnugget zum Scoren. Weitere Informationen finden Sie im Thema „Generieren eines Regelsets aus einem Assoziationsmodellnugget" auf Seite 243.

So verwenden Sie ein Modellnugget zum Generieren von Verarbeitungsknoten:

1. Durchsuchen Sie auf der Palette das Modell oder bearbeiten Sie es im Streamerstellungsbereich.
2. Wählen Sie den gewünschten Knotentyp im Generierungsmenü des Browsers für Modellnuggets aus. Die verfügbaren Optionen hängen vom Typ des Modellnuggets ab. Einzelheiten zu den Elementen, die aus einem bestimmten Modell generiert werden können, finden Sie in den Informationen zum jeweils generierten Modellnuggettyp.

Erneutes Erzeugen eines Modellierungsknotens

Wenn Sie ein Modellnugget ändern oder aktualisieren möchten und der zum Erstellen des Modells verwendete Stream nicht verfügbar ist, können Sie erneut einen Modellierungsknoten mit denselben Optionen erzeugen, die zum Erstellen des ursprünglichen Modells verwendet wurden.

Um ein Modell erneut zu erstellen, klicken Sie in der Modellpalette mit der rechten Maustaste auf das gewünschte Modell und wählen Sie die Option **Modellierungsknoten erzeugen**.

Alternativ können Sie beim Durchsuchen eines Modells im Menü "Generieren" die Option **Modellierungsknoten erzeugen** auswählen.

Der erneut erzeugte Modellierungsknoten sollte in den meisten Fällen von der Funktionsweise her identisch mit dem Knoten sein, mit dem das ursprüngliche Modell erstellt wurde.

- Bei Entscheidungsbaummodellen können zusammen mit dem Knoten auch weitere Einstellungen gespeichert werden, die während der interaktiven Sitzung angegeben wurden, und die Option **Interaktiv erstellte Direktiven verwenden** ist in dem neu erzeugten Modellierungsknoten aktiviert.
- Bei Entscheidungslistenmodellen ist die Option **Gespeicherte Informationen aus interaktiver Sitzung verwenden** aktiviert. Weitere Informationen finden Sie im Thema „Entscheidungslistenmodell - Optionen" auf Seite 142.
- Bei Zeitreihenmodellen ist die Option **Schätzung unter Verwendung bestehender Modelle fortsetzen** aktiviert. Diese Einstellung gestattet die erneute Erstellung des vorherigen Modells mit aktuellen Daten. Weitere Informationen finden Sie im Thema „Zeitreihenmodelle - Optionen" auf Seite 265.

Importieren und Exportieren von Modellen als PMML

PMML (Predictive Model Markup Language) ist ein XML-Format zur Beschreibung von Data-Mining-Modellen und statistischen Modellen, einschließlich der Eingaben zu den Modellen, der zur Vorbereitung der Daten für das Data-Mining verwendeten Transformationen sowie der Parameter, die die Modelle selbst definieren. IBM SPSS Modeler kann PMML importieren und exportieren, wodurch es möglich ist, Modelle mit anderen Anwendungen zu teilen, die dieses Format unterstützen, wie beispielsweise IBM SPSS Statistics.

Weitere Informationen zu PMML finden Sie auf der Website der Data Mining Group (*http://www.dmg.org*).

So exportieren Sie ein Modell:

PMML-Export wird für die meisten der in IBM SPSS Modeler erstellten Modelltypen unterstützt. Weitere Informationen finden Sie im Thema „Modelltypen, die PMML unterstützen" auf Seite 50.

1. Klicken Sie mit der rechten Maustaste auf ein Modellnugget in der Modellpalette. (Alternativ können Sie auf ein Modellnugget im Erstellungsbereich klicken und das Menü "Datei" auswählen.)
2. Klicken Sie im Menü auf **PMML exportieren**.

3. Geben Sie im Dialogfeld "Exportieren" (oder "Speichern") ein Zielverzeichnis und einen eindeutigen Namen für das Modell an.

Hinweis: Im Dialogfeld "Benutzeroptionen" können Sie Optionen für den PMML-Export ändern. Klicken Sie im Hauptmenü auf Folgendes:

Tools > **Optionen** > **Benutzeroptionen**

Klicken Sie auf die Registerkarte "PMML".

So importieren Sie ein als PMML gespeichertes Modell:

Modelle, die aus IBM SPSS Modeler oder einer anderen Anwendung als PMML exportiert wurden, können in die Modellpalette importiert werden. Weitere Informationen finden Sie im Thema „Modelltypen, die PMML unterstützen".

1. Klicken Sie in der Modellpalette mit der rechten Maustaste auf die Palette und wählen Sie aus dem Menü die Option **PMML importieren**.
2. Wählen Sie die zu importierende Datei aus und geben Sie nach Bedarf Optionen für Variablenbeschriftungen an.
3. Klicken Sie auf **Öffnen**.

Variablenbeschriftungen verwenden, sofern im Modell vorhanden. Im PMML-Code können sowohl die Variablennamen als auch die Variablenbeschriftungen (beispielsweise "Referrer ID" für *RefID*) für Variablen im Datenwörterbuch gefunden. Wählen Sie diese Option aus, um Variablenbeschriftungen zu verwenden, wenn diese im ursprünglich exportierten PMML-Code vorhanden sind.

Wenn Sie die Optionen zur Variablenbeschriftung ausgewählt haben, im PMML-Code jedoch keine Variablenbeschriftungen vorhanden sind, werden die Variablennamen wie üblich verwendet.

Modelltypen, die PMML unterstützen

PMML-Export

IBM SPSS Modeler-Modelle. Die folgenden in IBM SPSS Modeler erstellten Modelle können als PMML 4.0 exportiert werden:

- C&R-Baum
- QUEST
- CHAID
- Lineare Regression
- Netz
- C5.0
- Logistische Regression
- Genlin
- SVM
- Apriori
- Carma
- K-Means
- Kohonen
- Two Step
- GLMM (nur GLMM-Modelle mit festem Effekt werden unterstützt)
- Entscheidungsliste
- Cox

- Sequenz (Scoring für PMML-Sequenzmodelle wird nicht unterstützt)
- Statistics-Modell

Datenbankeigene Modelle. Bei Modellen, die mithilfe von datenbankeigenen Algorithmen generiert wurden, ist der PMML-Export nur bei IBM InfoSphere Warehouse-Modellen verfügbar. Modelle, die mithilfe von Analysis Services von Microsoft oder mit Oracle Data Miner erstellt wurden, können nicht exportiert werden. Beachten Sie außerdem, dass als PMML exportierte IBM Modelle danach nicht wieder in IBM SPSS Modeler importiert werden können.

PMML-Import

IBM SPSS Modeler kann PMML-Modelle importieren und scoren, die von aktuellen Versionen allerIBM SPSS Statistics-Produkte erstellt wurden, darunter Modelle, die aus IBM SPSS Modeler exportiert wurden, sowie Modell- bzw. Transformations-PMML, die von IBM SPSS Statistics 17.0 oder höher generiert wurde. Dies gilt also im Grunde für jegliche PMML, die die Scoring-Engine scoren kann - mit folgenden Ausnahmen:

- Apriori- und CARMA-Modelle sowie Anomalieerkennungs- und Sequenzmodelle können nicht importiert werden.
- PMML-Modelle können nach dem Import in IBM SPSS Modeler nicht durchsucht werden, obwohl sie für das Scoring verwendet werden können. (Dies gilt auch für Modelle, die ursprünglich aus IBM SPSS Modeler exportiert wurden. Um diese Einschränkung zu vermeiden, sollten Sie das betreffende Modell als generierte Modelldatei (**.gm*) und nicht als PMML exportieren.
- Als PMML exportierte IBM InfoSphere Warehouse-Modelle können nicht importiert werden.
- Eine eingeschränkte Validierung findet beim Import statt, aber die vollständige Validierung erfolgt beim Versuch, das Modell zu scoren. Daher kann der Import erfolgreich durchgeführt werden, das Scoring aber fehlschlagen oder falsche Ergebnisse erzeugen.

Veröffentlichen von Modellen für einen Scoring-Adapter

Sie können ein Modell für eine Datenbank mit Scoring-Adapter veröffentlichen. Ein Scoring-Adapter ermöglicht das Scoren von Modellen innerhalb der Datenbank, indem die benutzerdefinierte Funktion (UDF, User-defined Function) der Datenbank genutzt wird. Durch das Scoren innerhalb der Datenbank müssen vor dem Scoren keine Daten mehr extrahiert werden. Beim Veröffentlichen für einen Scoring-Adapter wird auch SQL-Beispielcode zum Ausführen der benutzerdefinierten Funktion generiert.

Veröffentlichen für einen Scoring-Adapter

1. Doppelklicken Sie auf das Modellnugget, um es zu öffnen.
2. Wählen Sie im Menü des Modellnuggets Folgendes aus:
   Datei > **Für Server-Scoring-Adapter veröffentlichen**
3. Füllen Sie die relevanten Felder im Dialogfeld aus und klicken Sie auf **OK**.

Datenbankverbindung. Die Verbindungsdetails für die für das Modell zu verwendende Datenbank.

Veröffentlichungs-ID. (Nur DB2 unter z/OS-Datenbanken) Eine ID für das Modell. Wenn Sie das gleiche Modell erneut erstellen und die gleiche Veröffentlichungs-ID verwenden, bleibt der generierte SQL-Code gleich. Somit kann ein Modell neu erstellt werden, ohne dass die Anwendung, die den vorher generierten SQL-Code verwendet hat, verändert werden muss. (Bei anderen Datenbanken ist der generierte SQL-Code einzig für das jeweilige Modell verwendbar.)

Beispiel-SQL generieren. Wenn die Option ausgewählt ist, wird der SQL-Beispielcode in der im Feld **Datei** angegebenen Datei erzeugt.

Nicht verfeinerte Modelle

Ein nicht verfeinertes Modell enthält Informationen, die aus den Daten extrahiert wurden, die jedoch nicht zum direkten Generieren von Vorhersagen gedacht sind. Daher kann es nicht zu Streams hinzugefügt werden. Nicht verfeinerte Modelle werden als "Rohdiamanten" in der generierten Modellpalette angezeigt.

Abbildung 27. Symbol für nicht verfeinerte Modelle

Informationen zum nicht verfeinerten Regelmodell erhalten Sie, wenn Sie mit der rechten Maustaste auf das Modell klicken und im Kontextmenü die Option **Durchsuchen** auswählen. Wie bei anderen in IBM SPSS Modeler generierten Modellen bieten die verschiedenen Registerkarten Übersichts- und Regelinformationen zum erstellten Modell.

Generieren von Knoten. Im Menü "Generieren" können Sie anhand der Regeln neue Knoten erstellen.

- **Auswahlknoten.** Generiert einen Auswahlknoten zur Auswahl von Datensätzen, für die die ausgewählte Regel gilt. Diese Option ist inaktiviert, wenn keine Regel ausgewählt wurde.
- **Regelset.** Generiert einen Regelsetknoten zur Vorhersage der Werte für ein einzelnes Zielfeld. Weitere Informationen finden Sie im Thema „Generieren eines Regelsets aus einem Assoziationsmodellnugget" auf Seite 243.

Kapitel 4. Screening von Modellen

Screening von Feldern und Datensätzen

In den vorgelagerten Phasen einer Analyse können mehrere Modellierungsknoten verwendet werden, um Felder und Datensätze zu finden, die voraussichtlich bei der Modellierung relevant sind. Sie können den Merkmalauswahlknoten verwenden, um Felder per Screening zu untersuchen und nach Wichtigkeit zu ordnen, und den Anomalieerkennungsknoten, um ungewöhnliche Datensätze zu finden, die nicht den bekannten Mustern "normaler" Daten entsprechen.

Der Merkmalauswahlknoten sichtet die Eingabefelder, um auf der Grundlage einer Reihe von Kriterien (z. B. dem Prozentsatz der fehlenden Werte) zu entscheiden, ob diese entfernt werden sollen. Anschließend erstellt er eine Wichtigkeitsrangfolge der verbleibenden Eingaben in Bezug auf ein angegebenes Ziel. Beispiel: Angenommen, Sie haben ein Dataset mit Hunderten potenzieller Eingaben. Welche davon sind voraussichtlich für die Modellierung von medizinischen Behandlungsergebnissen von Bedeutung?

Der Knoten "Anomalieerkennung" ermittelt ungewöhnliche Fälle oder Ausreißer, die nicht den Mustern von "normalen" Daten entsprechen. Mit diesem Knoten können Ausreißer ermittelt werden, selbst wenn sie keinem bereits bekannten Muster entsprechen und selbst wenn Sie nicht genau wissen, wonach Sie suchen.

Beachten Sie, dass bei der Anomalieerkennung ungewöhnliche Datensätze oder Fälle mithilfe einer Clusteranalyse ermittelt werden, die auf der im Modell ausgewählten Menge an Feldern beruht - ohne Berücksichtigung eines speziellen Zielfelds (abhängigen Felds) und unabhängig davon, ob diese Felder für das Muster relevant sind, das Sie vorherzusagen versuchen. Aus diesem Grund sollten Sie die Anomalieerkennung in Kombination mit der Merkmalauswahl oder einem anderen Verfahren für Screening und Rangordnung von Feldern verwenden. Beispielsweise können Sie mithilfe der Merkmalauswahl die wichtigsten Felder in Bezug auf ein bestimmtes Ziel ermitteln und anschließend mit der Anomalieerkennung die Datensätze finden, die in Bezug auf diese Felder besonders ungewöhnlich sind. (Eine alternative Vorgehensweise besteht darin, ein Entscheidungsbaummodell zu erstellen und anschließend alle falsch klassifizierten Datensätze als potenzielle Anomalien zu untersuchen. Diese Methode lässt sich jedoch nicht so leicht reproduzieren bzw. in größerem Maßstab automatisieren.)

Merkmalauswahlknoten

Ein Problem beim Data Mining kann darin bestehen, dass Hunderte oder sogar Tausende Felder potenziell als Eingaben in Frage kommen. Als Folge davon muss aufwendig untersucht werden, welche Felder bzw. Variablen in das Modell aufgenommen werden sollen. Um die Auswahlmöglichkeiten einzugrenzen, können mithilfe des Merkmalauswahlalgorithmus die Felder ermittelt werden, die für eine bestimmte Analyse am wichtigsten sind. Wenn Sie beispielsweise versuchen, die Ergebnisse medizinischer Behandlungen anhand einer Reihe von Faktoren vorherzusagen, welche Faktoren sind dann vermutlich am wichtigsten?

Die Merkmalauswahl besteht aus drei Schritten:

- **Screening.** Eliminiert unwichtige und problematische Eingaben und Datensätze bzw. Fälle, beispielsweise Eingabefelder mit zu vielen fehlenden Werten oder Eingaben, die eine so starke oder geringe Variation aufweisen, dass sie nicht brauchbar sind.
- **Ränge verwenden.** Sortiert die verbleibenden Eingaben und weist ihnen Ränge je nach Wichtigkeit zu.
- **Auswahl.** Ermittelt das Subset von Merkmalen, die in den nachfolgenden Modellen verwendet werden sollen, beispielsweise indem nur die wichtigsten Eingaben beibehalten und alle anderen gefiltert oder ausgeschlossen werden.

In einer Zeit, in der viele Unternehmen mit einer gewaltigen Datenflut umgehen müssen, können die Vorteile, die die Merkmalauswahl für die Vereinfachung und Beschleunigung des Modellierungsprozesses bietet, erheblich sein. Indem die Aufmerksamkeit schnell auf die wichtigsten Felder gelenkt wird, lässt sich der Berechnungsaufwand verringern, schwache, aber wichtige Beziehungen, die ansonsten leicht übersehen werden, können einfacher aufgespürt werden und schließlich erhalten Sie einfachere, genauere und leichter erklärbare Modelle. Wenn Sie die Anzahl der im Modell verwendeten Felder verringern, stellen Sie möglicherweise fest, dass Sie die Scoring-Zeiten verkürzen sowie die bei zukünftigen Wiederholungen zu sammelnde Datenmenge reduzieren können.

Beispiel. Eine Telefongesellschaft verfügt über ein Data Warehouse, das Informationen zu Reaktionen auf eine spezielle Werbeaktion enthält, die an 5.000 Kunden des Unternehmens gerichtet war. Die Daten enthalten eine Vielzahl von Feldern, darunter das Alter der Kunden, Ihr Beschäftigungsverhältnis, ihr Einkommen und statistische Daten zu ihrer Telefonnutzung. Drei Zielfelder zeigen jeweils an, ob der Kunde auf die drei Angebote reagierte oder nicht. Das Unternehmen möchte anhand dieser Daten vorhersagen, welche Kunden mit der größten Wahrscheinlichkeit in Zukunft auf ähnliche Angebote antworten.

Anforderungen. Ein einzelnes Zielfeld (mit der Rolle *Ziel*) sowie mehrere Eingabefelder, die in Bezug auf das Ziel untersucht bzw. nach Rang geordnet werden sollen. Sowohl Ziel- als auch Eingabefelder können das Messniveau *Stetig* (numerischer Bereich) oder *Kategorial* aufweisen.

Einstellungen für das Merkmalauswahlmodell

Die Einstellungen auf der Registerkarte "Modell" umfassen die standardmäßigen Modelloptionen sowie Einstellungen, mit denen Sie die Kriterien für das Screening von Eingabefeldern optimieren können.

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Screening von Eingabefeldern

Zum Screening gehört das Entfernen von Eingaben bzw. Fällen, die hinsichtlich der Beziehung zwischen Eingabe und Ziel keine nützlichen Informationen hinzufügen. Die Screeningoptionen beruhen auf Attributen des betreffenden Felds ohne Berücksichtigung der Vorhersagekraft in Bezug auf das ausgewählte Zielfeld. Die untersuchten Felder werden aus den Berechnungen für die Rangordnung der Eingaben ausgenommen und können optional gefiltert bzw. aus den bei der Modellierung verwendeten Daten entfernt werden.

Ein Screening der Felder kann auf folgenden Kriterien beruhen:

- **Maximaler Prozentsatz fehlender Werte.** Überprüft Felder mit zu vielen fehlenden Werten, ausgedrückt als Prozentsatz der Gesamtanzahl an Datensätzen. Felder mit einem großen Prozentsatz an fehlenden Werten bieten wenig Informationen für die Vorhersage.
- **Maximaler Prozentsatz der Datensätze in einer einzelnen Kategorie.** Untersucht Felder, bei denen zu viele Datensätze (im Verhältnis zur Gesamtzahl der Datensätze) in dieselbe Kategorie fallen. Wenn beispielsweise 95 % der Kunden in der Datenbank denselben Autotyp fahren, ist die Aufnahme dieser Information für die Unterscheidung der einzelnen Kunden untereinander nicht hilfreich. Alle Felder, die das angegebene Maximum überschreiten, werden im Screening untersucht. Diese Option gilt nur für kategoriale Felder.
- **Maximale Anzahl von Kategorien als Prozentsatz der Datensätze.** Untersucht Felder mit zu vielen Kategorien im Verhältnis zur Gesamtzahl der Datensätze. Wenn ein hoher Prozentsatz der Kategorien nur einen einzelnen Fall enthält, ist das Feld voraussichtlich von begrenztem Nutzen. Beispiel: Wenn jeder Kunde einen anderen Hut trägt, ist diese Information für die Modellierung von Verhaltensmustern mit großer Wahrscheinlichkeit unbrauchbar. Diese Option gilt nur für kategoriale Felder.
- **Minimaler Variationskoeffizient.** Überprüft Felder mit einem Varianzkoeffizient kleiner oder gleich dem angegebenen Minimum. Dieses Maß ist der Quotient aus der Standardabweichung des Eingabe-

felds und dem Mittelwert des Eingabefelds. Wenn dieser Wert nahe bei null liegt, liegt nur eine geringe Variabilität in den Werten für die betreffende Variable vor. Diese Option gilt nur für stetige Felder (numerischer Bereich).

- **Minimale Standardabweichung.** Überprüft Felder mit einer Standardabweichung kleiner oder gleich dem angegebenen Minimum. Diese Option gilt nur für stetige Felder (numerischer Bereich).

Datensätze mit fehlenden Daten. Datensätze oder Fälle mit fehlenden Werten für das Zielfeld oder fehlenden Werten für alle Eingaben werden automatisch aus allen Berechnungen für die Rangfolge ausgeschlossen.

Merkmalauswahloption

Auf der Registerkarte "Optionen" können Sie die Standardeinstellungen für die Auswahl bzw. den Ausschluss von Eingabefeldern im Modellnugget angeben. Anschließend können Sie das Modell zu einem Stream hinzufügen, um das Subset der Felder auszuwählen, die in nachfolgenden Modellerstellungsvorgängen verwendet werden sollen. Alternativ können Sie diese Einstellungen nach der Modellgeneration durch die Auswahl bzw. das Aufheben der Auswahl weiterer Felder im Modellbrowser überschreiben. Die Standardeinstellungen ermöglichen es jedoch, das Modellnugget ohne weitere Änderungen anzuwenden, was insbesondere für die Scripterstellung nützlich sein kann.

Weitere Informationen finden Sie im Thema „Ergebnisse des Merkmalauswahlmodells” auf Seite 56.

Die folgenden Optionen sind verfügbar:

Alle Felder bewertet als. Wählt die Felder auf der Grundlage ihres Ranges (*bedeutsam*, *marginal* oder *unbedeutend*) aus. Sie können die Beschriftung für jeden Rang bearbeiten sowie die Trennwerte ändern, die verwendet werden um Datensätze einem bestimmten Rang zuzuweisen.

Obere Anzahl an Feldern. Wählt die obersten *n* Felder nach Wichtigkeit aus.

Bedeutsamkeit größer als. Wählt alle Felder aus, deren Wichtigkeit den angegebenen Wert übersteigt.

Das Zielfeld bleibt unabhängig von der Auswahl immer erhalten.

Optionen für die Rangeinteilung nach Wichtigkeit

Ausschließlich kategoriale Werte. Wenn alle Eingaben und das Ziel kategorial sind, sind für die Rangeinteilung nach Wichtigkeit vier Maße verfügbar:

- **Pearson-Chi-Quadrat.** Testet auf Unabhängigkeit von Ziel und Eingabe ohne Angabe der Stärke oder Verwendung (Richtung) einer bestehenden Beziehung.
- **Likelihood-Quotienten-Chi-Quadrat.** Ähnlich dem Pearson-Chi-Quadrat; testet jedoch außerdem auf Ziel-Eingabe-Unabhängigkeit.
- **Cramer-V.** Ein Zusammenhangsmaß auf der Basis der Pearson-Chi-Quadrat-Statistik. Die Werte reichen von 0 (keine Assoziation) bis 1 (vollkommene Assoziation).
- **Lambda.** Ein Zusammenhangsmaß, das die proportionale Fehlerverringerung angibt, wenn die Variable zum Vorhersagen des Zielwerts verwendet wird. Der Wert 1 gibt an, dass das Eingabefeld das Ziel perfekt vorhersagt, während der Wert 0 bedeutet, dass die Eingabe keine nützlichen Informationen über das Ziel bietet.

Teilweise kategoriale Werte. Wenn einige, jedoch nicht alle Eingaben kategorial sind und das Ziel ebenfalls kategorial ist, lässt sich die Rangordnung nach Wichtigkeit basierend auf dem Pearson-Chi-Quadrat oder dem Likelihood-Quotienten-Chi-Quadrat ermitteln. (Cramer-*V* und Lambda sind nur verfügbar, wenn alle Eingaben kategorial sind.)

Kategorial im Vergleich zu stetig. Bei der Rangbewertung einer kategorialen Eingabe anhand eines stetigen Ziels oder umgekehrt (eines der beiden Elemente ist kategorial, nicht jedoch beide) wird die *F*-Statistik verwendet.

Beides stetig. Bei der Rangbewertung einer stetigen Eingabe anhand eines stetigen Ziels wird die *t*-Statistik auf der Grundlage des Korrelationskoeffizienten verwendet.

Modellnuggets vom Typ "Merkmalauswahl"

Modellnuggets vom Typ "Merkmalauswahl" zeigen die Bedeutsamkeit der einzelnen Eingaben in Bezug auf ein ausgewähltes Ziel an (gemäß der Rangeinstufung durch den Merkmalauswahlknoten). Alle Felder, die vor der Rangeinstufung per Screening ausgeschlossen wurden, werden ebenfalls aufgeführt. Weitere Informationen finden Sie im Thema „Merkmalauswahlknoten" auf Seite 53.

Bei der Ausführung eines Streams, der ein Modellnugget vom Typ "Merkmalauswahl" enthält, fungiert das Modell als Filter, mit dem nur ausgewählte Eingaben (in der aktuellen Auswahl auf der Registerkarte "Modell" angezeigt) beibehalten werden. Sie können beispielsweise alle als bedeutsam eingestuften Felder auswählen (eine der Standardoptionen) oder manuell ein Subset der Felder auf der Registerkarte "Modell" auswählen. Das Zielfeld bleibt unabhängig von der Auswahl ebenfalls erhalten. Alle anderen Felder werden ausgeschlossen.

Die Filterung beruht ausschließlich auf dem Feldnamen; wenn Sie beispielsweise *Alter* und *Einkommen* auswählen, wird jedes Feld, das mit einem dieser Namen übereinstimmt, beibehalten. Das Modell aktualisiert die Rangwerte für die Felder nicht anhand neuer Daten; es filtert einfach nur die Felder anhand der ausgewählten Namen. Aus diesem Grund sollten Sie bei der Anwendung des Modells auf neue oder aktualisierte Daten Vorsicht walten lassen. Im Zweifelsfall wird eine erneute Generierung des Modells empfohlen.

Ergebnisse des Merkmalauswahlmodells

Auf der Registerkarte "Modell" für ein Modellnugget vom Typ "Merkmalauswahl" wird der Rang und die Wichtigkeit aller Eingaben im oberen Fensterbereich angezeigt. Zudem haben Sie die Möglichkeit, mithilfe der Kontrollkästchen in der Spalte auf der linken Seite Felder für die Filterung auszuwählen. Wenn Sie den Stream ausführen, werden nur die ausgewählten Felder übernommen. Die anderen Felder werden verworfen. Die Standardauswahl beruht auf den im Modellerstellungsknoten angegebenen Optionen, Sie können jedoch nach Bedarf weitere Felder auswählen bzw. deren Auswahl aufheben.

Im unteren Fensterbereich werden Eingaben aufgelistet, die gemäß des Prozentsatzes an fehlenden Werten oder aufgrund anderer im Modellierungsknoten angegebener Kriterien aus der Rangwertung ausgenommen wurden. Wie bei den in die Rangwertung aufgenommenen Feldern, können Sie mithilfe der Kontrollkästchen in der Spalte auf der linken Seite auswählen, ob diese Felder eingeschlossen oder verworfen werden sollen. Weitere Informationen finden Sie im Thema „Einstellungen für das Merkmalauswahlmodell" auf Seite 54.

- Um die Liste nach Rang, Feldname, Wichtigkeit oder einer anderen der angezeigten Spalten zu sortieren, klicken Sie auf die Spaltenüberschrift. Wenn Sie lieber die Symbolleiste verwenden, wählen Sie das gewünschte Element in der Liste "Sortieren nach" aus. Mit den nach unten bzw. oben zeigenden Pfeilen können Sie die Sortierrichtung ändern.
- Mithilfe der Symbolleiste können Sie alle Felder markieren oder die Markierungen aufheben und auf das Dialogfeld "Felder markieren" zugreifen, in dem Sie Felder nach Rang und Wichtigkeit auswählen können. Außerdem können Sie beim Klicken auf die Felder die Umschalt- oder Steuertaste gedrückt halten, um die Auswahl zu erweitern, und mithilfe der Leertaste eine Gruppe ausgewählter Felder aktivieren bzw. inaktivieren. Weitere Informationen finden Sie im Thema „Auswählen der Felder nach Wichtigkeit" auf Seite 57.
- Die Schwellenwerte für die Einordnung von Eingaben als "bedeutsam", "marginal" bzw. "unbedeutend" werden in der Legende unterhalb der Tabelle angezeigt. Diese Werte werden im Modellierungsknoten angegeben. Weitere Informationen finden Sie im Thema „Merkmalauswahloption" auf Seite 55.

Auswählen der Felder nach Wichtigkeit

Beim Scoren von Daten mithilfe eines Modellnuggets vom Typ "Merkmalauswahl" bleiben alle Felder, die (mithilfe der Kontrollkästchen in der Spalte auf der linken Seite) aus der Liste der in Ränge eingeteilten bzw. per Screening untersuchten Felder ausgewählt wurden, erhalten. Die anderen Felder werden verworfen. Um die Auswahl zu ändern, können Sie mithilfe der Symbolleiste auf das Dialogfeld "Felder markieren" zugreifen, in dem Sie Felder nach Rang oder Wichtigkeit auswählen können.

Alle Felder, die markiert sind als. Wählt alle als bedeutsam, marginal oder unbedeutend markierten Felder aus.

Obere Anzahl an Feldern. Ermöglicht die Auswahl der obersten *n* Felder nach Wichtigkeit.

Bedeutsamkeit größer als. Wählt alle Felder aus, deren Wichtigkeit den angegebenen Schwellenwert übersteigt.

Generieren eines Filters aus einem Merkmalauswahlmodell

Basierend auf den Ergebnissen eines Merkmalauswahlmodells können Sie mithilfe des Dialogfelds "Filter aus Merkmalauswahl generieren" mindestens einen Filterknoten generieren, der Subsets von Feldern basierend auf ihrer Wichtigkeit in Bezug auf das angegebene Ziel ein- oder ausschließt. Das Modellnugget kann zwar auch als Filter verwendet werden, hiermit jedoch erhalten Sie die Flexibilität, mit verschiedenen Subsets von Feldern zu experimentieren, ohne das Modell kopieren oder bearbeiten zu müssen. Das Zielfeld wird stets vom Filter beibehalten, unabhängig davon ob "Einschließen" oder "Ausschließen" ausgewählt wurde.

Ein-/Ausschließen. Sie können auswählen, welche Felder ein- bzw. ausgeschlossen werden sollen. Sie können beispielsweise die obersten 10 Felder einschließen oder alle als unbedeutend markierte Felder ausschließen.

Ausgewählte Felder. Schließt alle aktuell in der Tabelle ausgewählten Felder ein bzw. aus.

Alle Felder, die markiert sind als. Wählt alle als bedeutsam, marginal oder unbedeutend markierten Felder aus.

Obere Anzahl an Feldern. Ermöglicht die Auswahl der obersten *n* Felder nach Wichtigkeit.

Bedeutsamkeit größer als. Wählt alle Felder aus, deren Wichtigkeit den angegebenen Schwellenwert übersteigt.

Anomalieerkennungsknoten

Anomalieerkennungsmodelle ermitteln ungewöhnliche Fälle bzw. Ausreißer in den Daten. Im Gegensatz zu anderen Modellierungsmethoden, bei denen Regeln zu ungewöhnlichen Fällen gespeichert sind, speichern Anomalieerkennungsmodelle ausführliche Informationen darüber, wie das "normale" Verhalten aussieht. Auf diese Weise können Ausreißer selbst dann erkannt werden, wenn sie keinem bekannten Muster entsprechen. Dies ist insbesondere in Einsatzgebieten wie der Betrugserkennung von Nutzen, bei denen ständig neue Muster auftreten. Die Anomalieerkennung ist eine nicht überwachte Methode; dies bedeutet, dass kein Trainingsdataset mit bekannten Betrugsfällen als Grundlage erforderlich ist.

Während herkömmliche Methoden zur Erkennung von Ausreißern in der Regel nur ein bis zwei Variablen gleichzeitig betrachten, kann die Anomalieerkennung zahlreiche Felder untersuchen und somit Cluster oder Peergruppen bilden, in die ähnliche Datensätze fallen. Die einzelnen Datensätze können dann jeweils mit den anderen Datensätzen in der Peergruppe verglichen werden, um so mögliche Anomalien zu erkennen. Je weiter ein Fall vom normalen Zentrum entfernt ist, desto größer ist die Wahrscheinlich-

keit, dass dieser Fall ungewöhnlich ist. Der Algorithmus kann beispielsweise die Datensätze zu drei unterschiedlichen Clustern zusammenfassen und dann die Datensätze mit einem Flag versehen, die weit vom Zentrum des jeweiligen Clusters entfernt sind.

Jeder Datensatz wird einem Anomalieindex zugewiesen, der dem Verhältnis des Gruppenabweichungsindex zum Durchschnitt des Clusters darstellt, zu dem der Fall gehört. Je größer der Wert dieses Index ist, desto stärker ist die Abweichung des Falls vom Durchschnitt. Unter normalen Umständen würden Fälle mit einem Anomalieindex kleiner als 1 oder gegebenenfalls auch 1,5 nicht als Anomalien betrachtet, weil die Abweichung nahezu mit dem Durchschnitt übereinstimmt oder nur wenig höher liegt. Fälle mit einem Indexwert größer 2 sind dagegen vielversprechende Anomaliekandidaten, weil die Abweichung hierbei mindestens das Zweifache des Durchschnitts beträgt.

Die Anomalieerkennung ist eine Explorationsmethode, mit der ungewöhnliche Fälle oder Datensätze rasch erkannt werden, die als Kandidaten für die weitere Analyse infrage kommen. Diese Kandidaten gelten als *vermutete* Anomalien, die sich bei näherer Untersuchung als tatsächliche Anomalien herausstellen können (oder auch nicht). Unter Umständen stufen Sie einen Datensatz als völlig normal ein, den Sie jedoch beim Aufbauen eines Modells von den Daten abschirmen möchten. Umgekehrt gilt: Wenn der Algorithmus wiederholt falsche Anomalien zurückliefert, kann dies auf einen Fehler oder ein Artefakt bei der Datensammlung hinweisen.

Beachten Sie, dass bei der Anomalieerkennung ungewöhnliche Datensätze oder Fälle mithilfe einer Clusteranalyse ermittelt werden, die auf der im Modell ausgewählten Menge an Feldern beruht - ohne Berücksichtigung eines speziellen Zielfelds (abhängigen Felds) und unabhängig davon, ob diese Felder für das Muster relevant sind, das Sie vorherzusagen versuchen. Aus diesem Grund sollten Sie die Anomalieerkennung in Kombination mit der Merkmalauswahl oder einem anderen Verfahren für Screening und Rangordnung von Feldern verwenden. Beispielsweise können Sie mithilfe der Merkmalauswahl die wichtigsten Felder in Bezug auf ein bestimmtes Ziel ermitteln und anschließend mit der Anomalieerkennung die Datensätze finden, die in Bezug auf diese Felder besonders ungewöhnlich sind. (Eine alternative Vorgehensweise besteht darin, ein Entscheidungsbaummodell zu erstellen und anschließend alle falsch klassifizierten Datensätze als potenzielle Anomalien zu untersuchen. Diese Methode lässt sich jedoch nicht so leicht reproduzieren bzw. in größerem Maßstab automatisieren.)

Beispiel. Bei der Untersuchung landwirtschaftlicher Subventionen auf mögliche Fälle von Betrug kann die Anomalieerkennung verwendet werden, um Abweichungen von der Norm aufzudecken, indem diejenigen Datensätze gekennzeichnet werden, die Unregelmäßigkeiten aufweisen und weiter untersucht werden müssen. Sie sind in erster Linie an Subventionsanträgen interessiert, die für den Typ und die Größe des landwirtschaftlichen Betriebs offenbar zu viel (oder zu wenig) Geld beantragen.

Anforderungen. Ein oder mehrere Eingabefelder. Beachten Sie, dass nur bei Feldern, bei denen eine Rolle auf **Eingabe** gesetzt ist, Quellen- oder Typknoten als Eingabe verwendet werden können. Zielfelder (Rolle auf **Ziel** oder **Beides** gesetzt) werden ignoriert.

Stärken. Durch die Kennzeichnung von Fällen, die einem bekannten Regelset *nicht* entsprechen (anstatt diejenigen Fälle zu kennzeichnen, die den Regeln entsprechen) können Anomalieerkennungsmodelle ungewöhnliche Fälle ermitteln, selbst wenn diese keinem zuvor bekannten Muster folgen. Bei Verwendung in Kombination mit der Merkmalauswahl kann mithilfe der Anomalieerkennung ein Screening großer Datenmengen durchgeführt werden, um die relevantesten Datensätze relativ schnell zu ermitteln.

Anomalieerkennung - Modelloptionen

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Trennwert für Anomalie ermitteln auf der Grundlage von. Gibt die Methode für die Bestimmung des Trennwerts zur Kennzeichnung von Anomalien an. Die folgenden Optionen sind verfügbar:

- **Anomalieindex größer oder gleich einem Minimalwert.** Gibt den minimalen Trennwert für die Kennzeichnung von Anomalien an. Datensätze, die diesen Schwellenwert erreichen oder überschreiten, werden gekennzeichnet.
- **Prozentsatz der Datensätze mit den größten Anomalien in den Trainingsdaten.** Legt den Schwellenwert automatisch auf einem Niveau fest, bei dem der angegebene Prozentsatz an Datensätzen in den Trainingsdaten gekennzeichnet wird. Der resultierende Cutoff wird als Parameter in das Modell aufgenommen. Beachten Sie, dass mit dieser Option bestimmt wird, wie der Trennwert festgelegt wird, *nicht* jedoch der tatsächliche Prozentsatz der beim Scoring zu kennzeichnenden Datensätze. Die tatsächlichen Scoring-Ergebnisse können je nach den Daten abweichen.
- **Anzahl der Datensätze mit den größten Anomalien in den Trainingsdaten.** Legt den Schwellenwert automatisch auf einem Niveau fest, bei dem die angegebene Anzahl an Datensätzen in den Trainingsdaten gekennzeichnet wird. Der resultierende Schwellenwert wird als Parameter in das Modell aufgenommen. Beachten Sie, dass mit dieser Option bestimmt wird, wie der Trennwert festgelegt wird, *nicht* jedoch die konkrete Anzahl der beim Scoring zu kennzeichnenden Datensätze. Die tatsächlichen Scoring-Ergebnisse können je nach den Daten abweichen.

Hinweis: Unabhängig davon, wie der Trennwert festgelegt wird, hat er keine Auswirkungen auf den zugrunde liegenden Anomalieindexwert, der für die einzelnen Datensätze gemeldet wird. Er legt einfach den Schwellenwert fest, ab dem die Datensätze beim Schätzen oder Scoren des Modells als anomal gekennzeichnet werden sollen. Wenn Sie später eine größere oder kleinere Anzahl von Datensätzen untersuchen möchten, können Sie einen Auswahlknoten verwenden, um ein Subset der Datensätze auf der Grundlage des Anomalieindexwerts (`$O-AnomalyIndex > X`) zu identifizieren.

Anzahl der zu meldenden Anomaliefelder. Gibt an, wie viele Felder gemeldet werden sollen, um anzugeben, warum ein bestimmter Datensatz als Anomalie gekennzeichnet wird. Die Felder mit den größten Anomalien werden gemeldet. Diese Felder sind definiert als diejenigen, die die größte Abweichung von der Feldnorm für den Cluster aufweisen, dem der Datensatz zugeordnet ist.

Anomalieerkennung - Expertenoptionen

Um Optionen für fehlende Werte und andere Einstellungen anzugeben, setzen Sie den Modus auf der Registerkarte "Experten" auf **Experten**.

Anpassungskoeffizient. Wert, der zum Balancieren der relativen Gewichtung verwendet wird, die den stetigen Feldern (numerischer Bereich) und den kategorialen Feldern bei der Berechnung der Distanz zugewiesen wird. Größere Werte erhöhen den Einfluss der stetigen Felder. Dieser Wert darf kein Nullwert sein.

Anzahl der Peergruppen automatisch berechnen. Mit der Anomalieerkennung lässt sich eine schnelle Analyse einer großen Anzahl von möglichen Lösungen durchführen, um die optimale Anzahl an Peergruppen für die Trainingsdaten auszuwählen. Sie können den Bereich erweitern oder einengen, indem Sie die minimale bzw. maximale Anzahl an Peergruppen festlegen. Mit größeren Werten kann das System einen breiteren Bereich möglicher Lösungen durchsuchen. Allerdings erhöht sich dadurch die Verarbeitungszeit.

Anzahl der Peergruppen angeben. Wenn Sie wissen, wie viele Cluster in Ihr Modell aufgenommen werden sollen, wählen Sie diese Option und geben Sie die Anzahl der Peergruppen ein. Die Auswahl dieser Option führt im Allgemeinen zu einer besseren Leistung.

Rauschlevel und Rauschverhältnis. Diese Einstellungen bestimmen, wie Ausreißer bei der zweistufigen Clusterbildung behandelt werden. In der ersten Stufe wird ein Clustermerkmalsbaum (CF-Baum) verwendet, um die Daten aus einer sehr großen Anzahl einzelner Datensätze auf eine überschaubare Anzahl von Clustern zu verdichten. Der Baum wird anhand von Ähnlichkeitsmaßen erstellt. Wenn ein Knoten des Baums zu viele Datensätze enthalten würde, wird er in untergeordnete Knoten aufgespalten. In der zweiten Phase beginnt die hierarchische Clusterbildung an den Endknoten des CD-Baums. Die Rauschverar-

beitung ist beim ersten Datendurchlauf aktiviert und beim zweiten Datendurchlauf inaktiviert. Die Fälle im Rauschcluster aus dem ersten Datendurchlauf werden im zweiten Datendurchlauf den regulären Clustern zugewiesen.

- **Rauschlevel.** Geben Sie einen Wert zwischen 0 und 0,5 an. Diese Einstellung ist nur relevant, sofern der CF-Baum während der Wachstumsphase gefüllt wird, wenn er also keine weiteren Fälle in einem Blattknoten annehmen kann und kein Blattknoten aufgeteilt werden kann.

  Wenn der CF-Baum gefüllt wird und die Rausch-Ebene auf 0 gesetzt ist, wird der Schwellenwert erhöht und der CF-Baum wird mit allen Fällen neu erstellt. Nach der abschließenden Clusteranalyse werden die Werte, die keinem Cluster zugewiesen werden konnten, als Ausreißer beschriftet. Das Ausreißercluster erhält die Identifikationsnummer -1. Das Ausreißercluster wird nicht in die Anzahl der Cluster eingeschlossen, d. h., wenn Sie *n* Cluster und Rauschverarbeitung angeben, gibt der Algorithmus *n* Cluster und einen Rauschcluster aus. In der Praxis bedeutet dies, dass bei einer Erhöhung dieses Werts der Algorithmus mehr Spielraum hat, ungewöhnliche Datensätze in den Baum einzupassen. Er muss sie also nicht einem gesonderten Ausreißer-Cluster zuweisen.

  Wenn der CF-Baum gefüllt wird und die Rauschebene größer als 0 ist, wird der CF-Baum neu gebildet, nachdem alle Daten in "dünn besetzten" Blättern in einem eigenen Rausch-Blatt abgelegt wurden. Ein Blatt wird als "dünn besetzt" betrachtet, wenn das Verhältnis der Anzahl der Fälle im dünn besetzten Blatt zu der Anzahl der Fälle im größten Blatt kleiner ist als die Rauschebene. Nach der Neubildung des Baums werden die Ausreißer nach Möglichkeit im CF-Baum positioniert. Andernfalls werden die Ausreißer für die zweite Phase der Clusterbildung verworfen.
- **Rauschverhältnis.** Gibt an, welcher Anteil des der Komponente zugeordneten Arbeitsspeichers für die Rausch-Pufferung verwendet werden soll. Dieser Wert liegt im Bereich von 0,0 bis 0,5. Wenn das Hinzufügen eines gegebenen Falls zu einem Blatt des Baums zu einer Dichte unterhalb dieses Schwellenwerts führen würde, wird das Blatt nicht geteilt. Wenn die Dichte den Schwellenwert überschreitet, wird das Blatt geteilt und ein weiterer kleiner Cluster wird zum CF-Baum hinzugefügt. Durch die Erhöhung dieser Einstellung strebt der Algorithmus also möglicherweise schneller hin zu einem einfacheren Baum.

Fehlende Werte imputieren. Setzt bei stetigen Feldern für fehlende Werte den Feldmittelwert ein. Bei kategorialen Feldern werden fehlende Kategorien kombiniert und als gültige Kategorie behandelt. Wenn die Auswahl dieser Option aufgehoben wird, werden alle Datensätze mit fehlenden Werten aus der Analyse ausgeschlossen.

Modellnuggets vom Typ "Anomalieerkennung"

Modellnuggets vom Typ "Anomalieerkennung" enthalten alle Informationen, die vom Anomalieerkennungsmodell erfasst wurden, sowie Informationen zu den Trainingsdaten und dem Schätzvorgang.

Bei der Ausführung eines Streams, der ein Modellnugget vom Typ "Anomalieerkennung" enthält, wird eine Reihe neuer Felder zum Stream hinzugefügt. Diese werden durch die Auswahl auf der Registerkarte "Einstellungen" im Modellnugget festgelegt. Weitere Informationen finden Sie im Thema „Anomalieerkennungsmodell - Einstellungen” auf Seite 61. Neue Feldnamen basieren auf dem Modellnamen und tragen das Präfix *$O*, wie in der folgenden Tabellen zusammengefasst.

Tabelle 6. Generierung neuer Feldnamen.

Feldname	Beschreibung
$O-Anomaly	Flagfeld, das angibt, ob der Datensatz anomal ist oder nicht.
$O-AnomalyIndex	Der Anomalieindexwert für den Datensatz.
$O-PeerGroup	Gibt die Peergruppe an, der der Datensatz zugewiesen ist.
$O-Field-n	Name des Felds, das den *n*. Rang in der Reihenfolge der anomalsten Felder einnimmt (hinsichtlich der Abweichung von der Clusternorm).

Tabelle 6. Generierung neuer Feldnamen (Forts.).

$O-FieldImpact-n	Variabler Abweichungsindex für das Feld. Dieser Wert misst die Abweichung von der Feldnorm für den Cluster, dem der Datensatz zugewiesen ist.

Optional können Sie Scores für nicht anomale Datensätze unterdrücken, um die Lesbarkeit der Ergebnisse zu verbessern. Weitere Informationen finden Sie im Thema „Anomalieerkennungsmodell - Einstellungen".

Anomalieerkennungsmodelle - Details

Auf der Registerkarte "Modell" für ein generiertes Anomalieerkennungsmodell werden Informationen zu den Peergruppen im Modell angezeigt.

Beachten Sie, dass die gemeldeten Größen und Statistiken für die Peergruppe auf den Trainingsdaten beruhen und etwas vom tatsächlichen Scoring-Ergebnis abweichen können, selbst wenn dieselben Daten verwendet werden.

Anomalieerkennungsmodell - Übersicht

Auf der Registerkarte "Übersicht" für ein Modellnugget vom Typ "Anomalieerkennung" werden Informationen über die Felder, die Aufbaueinstellungen und den Schätzvorgang angezeigt. Außerdem werden die Anzahl der Peergruppen sowie die Anzahl des Trennwerts angezeigt, der verwendet wird, um Datensätze als anomal zu kennzeichnen.

Anomalieerkennungsmodell - Einstellungen

Über die Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modellnuggets angeben.

Anomale Datensätze kennzeichnen mit. Gibt an, wie anomale Datensätze in der Ausgabe behandelt werden sollen.

- **Flag und Index.** Erstellt ein Flagfeld, das für alle Datensätze, die den im Modell enthaltenen Trennwert überschreiten, auf *True* (Wahr) gesetzt ist. Der Anomalieindex wird außerdem für jeden Datensatz in einem separaten Feld angegeben. Weitere Informationen finden Sie im Thema „Anomalieerkennung - Modelloptionen" auf Seite 58.
- **Nur Flag.** Erstellt ein Flagfeld, jedoch ohne den Anomalieindex für die einzelnen Datensätze zu melden.
- **Nur Index.** Meldet den Anomalieindex, ohne ein Flagfeld zu erstellen.

Anzahl der zu meldenden Anomaliefelder. Gibt an, wie viele Felder gemeldet werden sollen, um anzugeben, warum ein bestimmter Datensatz als Anomalie gekennzeichnet wird. Die Felder mit den größten Anomalien werden gemeldet. Diese Felder sind definiert als diejenigen, die die größte Abweichung von der Feldnorm für den Cluster aufweisen, dem der Datensatz zugeordnet ist.

Datensätze verwerfen. Wählen Sie diese Option, um alle nicht anomalen Datensätze aus dem Stream zu verwerfen. Dadurch können Sie sich leichter auf potenzielle Anomalien in nachgeordneten Knoten konzentrieren. Alternativ können Sie festlegen, dass alle anomalen Datensätze verworfen werden sollen, um die nachfolgende Analyse auf diejenigen Datensätze zu begrenzen, die nicht auf der Grundlage des Modells als potenzielle Anomalien gekennzeichnet wurden.

Hinweis: Aufgrund kleiner Unterschiede beim Runden stimmt die tatsächliche Anzahl beim Scoring markierter Datensätze möglicherweise nicht mit der Anzahl der bei Trainieren des Modells markierten Datensätze überein, selbst wenn in beiden Fällen dieselben Daten verwendet wurden.

Kapitel 5. Knoten für die automatisierte Modellierung

Die automatisierten Modellierungsknoten schätzen und vergleichen eine Reihe von verschiedenen Modellierungsmethoden, sodass Sie eine Vielzahl von Ansätzen in einer einzigen Modellierungsausführung ausprobieren können. Sie können die zu verwendenden Modellierungsalgorithmen und die jeweils spezifischen Optionen auswählen, inklusive Kombinationen, die sich andernfalls gegenseitig ausschließen würden. Beispielsweise müssen Sie sich nicht zwischen den Methoden "Schnell", "Dynamisch" und "Reduzieren" für ein neuronales Netz entscheiden, sondern können alle drei Methoden ausprobieren. Der Knoten untersucht jede mögliche Kombination von Optionen, stuft jedes in Frage kommende Modell auf der Basis des angegebenen Werts und speichert die geeignetsten Kombinationen in Scoring oder weiterer Analyse.

Sie können je nach Anforderungen Ihrer Analyse zwischen drei Knoten für automatisierte Modellierung wählen:

Mit dem Knoten "Autom. Klassifikationsmerkmal" können Sie eine Reihe verschiedener Modelle für binäre Ergebnisse ("Ja" oder "Nein", "Abwanderung" oder "Keine Abwanderung" usw.) erstellen und vergleichen, um den besten Ansatz für die jeweilige Analyse auszuwählen. Es wird eine Reihe von Modellierungsalgorithmen unterstützt, sodass Sie die gewünschten Methoden, die spezifischen Optionen für die jeweilige Methode und die Kriterien zum Vergleich der Ergebnisse auswählen können. Der Knoten generiert eine Gruppe von Modellen, die auf den angegebenen Optionen beruhen, und erstellt anhand der von Ihnen angegebenen Kriterien eine Rangordnung der besten Kandidaten.

Der Knoten "Autonumerisch" schätzt und vergleicht mit einer Reihe verschiedener Methoden Modelle für die Ergebnisse stetiger numerischer Bereiche. Der Knoten arbeitet auf dieselbe Weise wie der Knoten "Autom. Klassifikationsmerkmal": Sie können die zu verwendenden Algorithmen auswählen und in einem Modellierungsdurchlauf mit mehreren Optionskombinationen experimentieren. Folgende Algorithmen werden unterstützt: Neuronale Netze, C&R-Baum, CHAID, lineare Regression, verallgemeinerte lineare Regression nd Support Vector Machines (SVM). Modelle können anhand von Korrelation, relativem Fehler bzw. Anzahl der verwendeten Variablen verglichen werden.

Mit dem Knoten "Autom. Cluster" können Sie Clustering-Modelle, die Gruppen und Datensätze mit ähnlichen Merkmalen identifizieren, schätzen und vergleichen. Die Funktionsweise des Knotens gleicht der von anderen Knoten für automatisierte Modellierung, und Sie können in einem einzigen Modellierungsdurchgang mit mehreren Optionskombinationen experimentieren. Modelle können mithilfe grundlegender Messwerte für Filterung und Rangfolge der Nützlichkeit von Clustermodellen verglichen werden, um ein Maß auf der Basis der Wichtigkeit von bestimmten Feldern zu liefern.

Die besten Modelle werden in einem einzelnen zusammengesetzten Modellnugget gespeichert, sodass Sie sie durchsuchen und vergleichen und zudem auswählen können, welche Modelle für das Scoring verwendet werden sollen.

- Nur für binäre, nominale und numerische Ziele können Sie mehrere Scoring-Modelle auswählen und die Scores in einem einzigen Modell-Ensemble kombinieren. Durch die Kombination der Vorhersagen aus mehreren Modellen lassen sich Begrenzungen, die einzelne Modelle aufweisen, vermeiden. Dadurch kann häufig eine höhere Gesamtgenauigkeit erreicht werden als mit einem der Modelle allein.
- Optional können Sie einen Drill-Down für die Ergebnisse durchführen und Modellierungsknoten oder Modellnuggets für jedes einzelne Modell generieren lassen, das Sie weiterverwenden oder eingehender untersuchen möchten.

Modelle und Ausführungszeit

Abhängig von dem Dataset und der Anzahl an Modellen kann die Ausführung von Knoten für automatisierte Modellierung Stunden oder noch länger dauern. Achten Sie bei der Auswahl von Optionen auf die Anzahl der erstellten Modelle. Nach Möglichkeit sollten Sie die Modellierungsdurchläufe für die Nacht oder das Wochenende planen, wenn die Systemressourcen weniger ausgelastet sind.

- Falls erforderlich kann die Anzahl der Datensätze im ursprünglichen Trainingsdurchlauf mithilfe eines Partitions- oder Stichprobenknotens reduziert werden. Nachdem Sie die Auswahl auf einige wenige infrage kommende Modelle eingeschränkt haben, kann das vollständige Dataset wiederhergestellt werden.
- Mit der Merkmalauswahl können Sie die Anzahl der Eingabefelder verringern. Weitere Informationen finden Sie im Thema „Merkmalauswahlknoten" auf Seite 53. Alternativ können Sie durch die anfänglichen Modellierungsausführungen Felder und Optionen ermitteln, deren weitere Untersuchung lohnenswert ist. Wenn beispielsweise die besten Modelle jeweils dieselben drei Felder zu verwenden scheinen, ist dies ein deutlicher Hinweis darauf, dass diese Felder beibehalten werden sollten.
- Optional können Sie die Zeit, die für die Schätzung eines Modells aufgewendet werden soll, begrenzen und die für Screening und Rangordnung der Modelle zu verwendenden Evaluierungsmaße angeben.

Knoten für die automatisierte Modellierung - Algorithmuseinstellungen

Sie können für die einzelnen Modelltypen entweder die Standardeinstellungen verwenden oder Optionen für den jeweiligen Modelltyp auswählen. Die einzelnen Optionen ähneln den Optionen, die in den gesonderten Modellierungsknoten verfügbar sind, mit dem Unterschied, dass nicht eine bestimmte Einstellung ausgewählt werden muss, sondern in den meisten Fällen beliebig viele Einstellungen verwendet werden können. So können Sie beispielsweise beim Vergleich von Netzmodellen mehrere verschiedene Trainingsmethoden auswählen und jede Methode mit und ohne Zufallsstartwert ausprobieren. Es werden alle möglichen Kombinationen der ausgewählten Optionen verwendet. Dadurch wird es sehr einfach, viele verschiedene Modelle in einem einzelnen Durchgang zu generieren. Seien Sie jedoch vorsichtig, da die Auswahl mehrerer Einstellungen die Anzahl der Modelle sehr schnell vervielfachen kann.

So wählen Sie Optionen für den jeweiligen Modelltyp aus:

1. Wählen Sie für den automatisierten Modellierungsknoten die Registerkarte **Experten** aus.
2. Klicken Sie auf die Spalte **Modellparameter** für den Modelltyp.
3. Wählen Sie im Dropdown-Menü die Option **Angeben**.
4. Wählen Sie im Dialogfeld **Algorithmuseinstellungen** die Optionen in der Spalte **Optionen** aus.

Hinweis: Weitere Optionen sind auf der Registerkarte "Experten" des Dialogfelds **Algorithmuseinstellungen** verfügbar.

Knoten für die automatisierte Modellierung - Stoppregeln

Für die Knoten "Automatisierte Modellierung" angegebene Stoppregeln beziehen sich auf die gesamte Knotenausführung, nicht auf das Stoppen einzelner, vom Knoten erstellter Modelle.

Gesamtausführungszeit beschränken. (Nur für neuronale Netze sowie K-Means-, Kohonen-, TwoStep-, SVM-, KNN-, Bayes-Netz- und C&R-Baum-Modelle) Stoppt die Ausführung nach einer angegebenen Anzahl an Stunden. Alle bis zu diesem Zeitpunkt generierten Modelle werden in das Modellnugget aufgenommen, es werden jedoch keine weiteren Modelle erstellt.

Anhalten, sobald gültige Modelle generiert werden. Stoppt die Ausführung, wenn ein Modell alle auf der Registerkarte "Verwerfen" (für den Knoten "Autom. Klassifikationsmerkmal" oder "Autom. Cluster") oder der Registerkarte "Modell" (für den Knoten "Autonumerisch") angegebenen Kriterien erfüllt. Weitere

Informationen finden Sie im Thema „Knoten "Autom. Klassifikationsmerkmal" - Optionen für Verwerfen” auf Seite 70. Weitere Informationen finden Sie im Thema „Knoten "Autom. Cluster" - Optionen für Verwerfen” auf Seite 77.

Knoten "Autom. Klassifikationsmerkmal"

Mit dem Knoten "Autom. Klassifikationsmerkmal" können Sie mit verschiedenen Methoden Modelle für nominale (Setziel) oder binäre Ziele (Ja/Nein) schätzen und vergleichen, wodurch Sie eine Vielzahl von Ansätzen in einer einzigen Modellausführung ausprobieren können. Sie können die gewünschten Algorithmen auswählen und mit mehreren Kombinationen von Optionen experimentieren. Beispielsweise müssen Sie sich nicht zwischen den Methoden "Schnell", "Dynamisch" und "Reduzieren" für ein neuronales Netz entscheiden, sondern können alle drei Methoden ausprobieren. Der Knoten prüft jede mögliche Optionskombination, stuft jedes in Frage kommende Modell basierend auf dem angegebenen Maß ein und speichert die besten Modelle für das Scoring oder die weitere Analyse. Weitere Informationen finden Sie in Kapitel 5, „Knoten für die automatisierte Modellierung”, auf Seite 63.

Beispiel. Ein Einzelhandelsunternehmen verfügt über historische Daten, die die Angebote verfolgen, die bestimmten Kunden in früheren Werbeaktionen unterbreitet wurden. Das Unternehmen möchte nun profitablere Ergebnisse erzielen, indem es für jeden Kunden das richtige Angebot ermittelt.

Anforderungen. Ein Zielfeld mit einem Messniveau des Typs *Nominal* oder *Flag* (mit der Rolle **Ziel**) und mindestens ein Eingabefeld (mit der Rolle **Eingabe**). Für ein Flagfeld gilt der für das Ziel definierte *Wahr*-Wert bei der Berechnung von Profiten, Lifts und verwandten Statistiken als Treffer. Eingabefelder können ein Messniveau von *Stetig* oder *Kategorial* aufweisen, mit der Einschränkung, dass einige Eingaben für bestimmte Modelltypen möglicherweise nicht geeignet sind. Ordinale Felder beispielsweise, die als Eingaben in Modellen vom Typ "C&R-Baum", "CHAID" und "QUEST" verwendet werden sollen, müssen einen numerischen Speichertyp (nicht "Zeichenfolge") aufweisen und werden andernfalls von diesen Modellen ignoriert. Ebenso können stetige Eingabefelder in einigen Fällen klassiert werden. Die Anforderungen sind dieselben wie bei Verwendung der einzelnen Modellierungsknoten; so funktioniert ein Bayes-Netzmodell immer auf dieselbe Weise, unabhängig davon, ob es über den Bayes-Netzknoten oder den Knoten "Autom. Klassifikationsmerkmal" generiert wurde.

Häufigkeits- und Gewichtungsfelder. Häufigkeit und Gewichtung dienen dazu, einigen Datensätzen eine größere Bedeutsamkeit zu verleihen als anderen, beispielsweise, weil der Benutzer weiß, dass ein Teil der übergeordneten Grundgesamtheit im erstellten Dataset unterrepräsentiert ist (Gewichtung) oder weil ein Datensatz für eine Reihe identischer Fälle steht (Häufigkeit). Häufigkeitsfelder können, sofern angegeben, von Modellen vom Typ "C&R-Baum", "CHAID", "QUEST", "Entscheidungsliste" und "Bayes-Netz" verwendet werden. Gewichtungsfelder können von Modellen vom Typ "C&R-Baum", "CHAID" und "C5.0" verwendet werden. Andere Modelltypen ignorieren diese Felder und erstellen die Modelle in jedem Fall. Häufigkeits- und Gewichtungsfelder werden nur für die Modellerstellung verwendet. Bei der Evaluation bzw. beim Scoren von Modellen werden sie nicht berücksichtigt. Weitere Informationen finden Sie im Thema „Verwenden von Häufigkeits- und Gewichtungsfeldern” auf Seite 33.

Unterstützte Modelltypen

Folgende Modelltypen werden unterstützt: "Neuronales Netz", "C&R-Baum", "QUEST", "CHAID", "C5.0", "Logistische Regression", "Entscheidungsliste", "Bayes-Netz", "Diskriminanz", "Nächster Nachbar" und "SVM". Weitere Informationen finden Sie im Thema „Knoten "Autom. Klassifikationsmerkmal" - Expertenoptionen” auf Seite 67.

Knoten "Autom. Klassifikationsmerkmal" - Modelloptionen

Über die Registerkarte "Modell" auf dem Knoten "Autom. Klassifikationsmerkmal" können Sie die Anzahl der zu erstellenden Modelle sowie die zum Vergleichen der Modelle verwendeten Kriterien angeben.

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Modelle in Ränge einteilen nach. Legt fest, welche Kriterien für Vergleich und Rangordnung von Modellen verwendet werden sollen. Zu den Optionen gehören die Gesamtgenauigkeit, die Fläche unter der ROC-Kurve, Profit, Lift und die Anzahl der Felder. Beachten Sie, dass alle diese Maße im Zusammenfassungsbericht angegeben werden, unabhängig davon, welches davon an dieser Stelle ausgewählt wird.

Hinweis: Bei einem nominalen Ziel (Setziel) ist die Rangfolge auf **Gesamtgenauigkeit** oder **Anzahl der Felder** eingeschränkt.

Bei der Berechnung von Profiten, Lifts und verwandten Statistiken gilt der für das Zielfeld definierte *Wahr*-Wert als Treffer.

- **Gesamtgenauigkeit** Der Prozentsatz der Datensätze, der korrekt vom Modell vorhergesagt wird, im Verhältnis zur Gesamtzahl der Datensätze.
- **Fläche unter der ROC-Kurve** Die ROC-Kurve bietet einen Index für die Leistungsfähigkeit eines Modells. Je höher die Kurve über der Bezugslinie liegt, desto genauer ist der Test.
- **Profit (Kumulativ)** Die Summe der Profite aus kumulativen Perzentilen (nach Konfidenz für die Vorhersage sortiert), wie auf der Grundlage der angegebenen Kosten, Einkünfte und Gewichtungskriterien berechnet. Normalerweise beginnt der Profit nahe 0 für das oberste Perzentil, nimmt stetig zu und sinkt dann wieder. Bei einem guten Modell zeigen die Profite eine klar ausgeprägte Spitze im Mittelteil, die zusammen mit dem Perzentil, in dem sie auftritt, angegeben wird. Bei einem Modell ohne Informationsgehalt verläuft die Profitkurve relativ gerade; die Linie kann ansteigen, abfallen oder auf demselben Niveau verbleiben, abhängig von der vorliegenden Kosten-Einnahmen-Struktur.
- **Lift (Kumulativ)** Der Quotient der Treffer in kumulativen Quantilen in Bezug auf die Gesamtstichprobe (wobei die Quantile nach der Konfidenz für die Vorhersage sortiert sind). Ein Lift-Wert von 3 für das oberste Perzentil zeigt beispielsweise eine Trefferquote an, die dreimal so hoch ist wie für die Stichprobe insgesamt. Bei einem guten Modell sollte der Lift deutlich über 1,0 für die obersten Quantile beginnen und dann stark in Richtung 1,0 für die unteren Quantile abfallen. Bei einem Modell ohne Informationsgehalt liegt der Lift-Wert ungefähr bei 1,0.
- **Anzahl der Felder** Teilt Modelle auf der Grundlage der verwendeten Eingabefelder in Ränge ein.

Modelle in Ränge einteilen mithilfe von. Wenn eine Partition verwendet wird, können Sie angeben, ob die Ränge auf dem Trainingsdataset oder auf dem Testdataset beruhen sollen. Bei großen Datasets lässt sich die Leistungsfähigkeit durch die Verwendung einer Partition für ein erstes Screening der Modelle u. U. erheblich verbessern.

Anzahl der zu verwendenden Modelle. Legt die maximale Anzahl der Modelle fest, die in dem vom Knoten erstellten Modellnugget aufgeführt werden sollen. Die Modelle mit dem höchsten Rang werden gemäß dem angegebenen Rangordnungskriterium aufgeführt. Beachten Sie, dass eine Erhöhung dieses Grenzwerts die Leistungsgeschwindigkeit verringern kann. Der höchste zulässige Wert ist 100.

Prädiktoreinfluss berechnen. Bei Modellen, die zu einem angemessenen Maß an Bedeutsamkeit führen, können Sie ein Diagramm anzeigen, in dem der relative Einfluss der einzelnen Prädiktoren bei der Modellschätzung angegeben wird. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und diejenigen Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass sich bei einigen Modellen der Zeitaufwand für die

Berechnung durch den Prädiktoreinfluss erhöhen kann. Außerdem wird diese Option nicht empfohlen, wenn Sie einfach einen allgemeinen Vergleich zwischen vielen verschiedenen Modellen wünschen. Diese Option ist von größerem Nutzen, wenn die Analyse bereits auf eine Handvoll Modelle eingeengt wurde, die detaillierter untersucht werden sollen. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Profitkriterien. *Anmerkung*. Nur für Flagziele. Der Profit entspricht dem Umsatz für jeden Datensatz abzüglich der Kosten für den betreffenden Datensatz. Die Profite für ein Quantil entsprechen einfach der Summe der Profite für alle Datensätze im Quantil. Profite gelten definitionsgemäß nur für Treffer, Kosten dagegen für alle Datensätze.

- **Kosten.** Geben Sie die Kosten für die einzelnen Datensätze an. Wählen Sie die Option **Fest** oder **Variabel** für den Umsatz. Bei festen Kosten geben Sie den Wert der Kosten ein. Bei variablen Kosten klicken Sie auf die Feldauswahlschaltfläche und bestimmen Sie ein Feld als Kostenfeld.
- **Umsatz.** Geben Sie den Umsatz für die einzelnen Datensätze ein, die als Treffer gelten. Wählen Sie die Option **Fest** oder **Variabel** für den Umsatz. Bei einem festen Umsatz geben Sie den Wert des Umsatzes ein. Bei einem variablen Umsatz klicken Sie auf die Feldauswahlschaltfläche und bestimmen Sie ein Feld als Umsatzfeld.
- **Gewichtung.** Wenn die Datensätze in den Daten für mehrere Einheiten stehen, können Sie die Ergebnisse mithilfe der Häufigkeitsgewichtungen anpassen. Geben Sie die Gewichtung für die einzelnen Datensätze im Feld **Fest** oder **Variabel** an. Bei einer festen Gewichtung geben Sie den Wert für die Gewichtung an (die Anzahl der Einheiten pro Datensatz). Bei variablen Gewichtungen klicken Sie auf die Schaltfläche für die Feldauswahl und bestimmen ein Feld als Gewichtungsfeld.

Liftkriterien. *Anmerkung*. Nur für Flagziele. Gibt das für Lift-Berechnungen zu verwendende Perzentil an. Diesen Wert können Sie auch beim Vergleichen der Ergebnisse ändern. Weitere Informationen finden Sie im Thema „Nugget für automatisierte Modellierung" auf Seite 77.

Knoten "Autom. Klassifikationsmerkmal" - Expertenoptionen

Über die Registerkarte "Experten" des Knotens "Autom. Klassifikationsmerkmal" können Sie eine Partition (falls verfügbar) anwenden, die zu verwendenden Algorithmen auswählen und Stoppregeln angeben.

Modelle auswählen. Standardmäßige werden alle Modelle zum Erstellen ausgewählt. Wenn Sie jedoch Analytic Server installiert haben, können Sie angeben, dass nur die Modelle ausgewählt werden, die mit Analytic Server ausgeführt werden können. Sie können diese Modelle so vordefinieren, dass sie aufgeteilte Modelle erstellen oder sehr große Datasets verarbeiten können.

Verwendete Modelle. Wählen Sie anhand der Kontrollkästchen in der Spalte auf der linken Seite die Modelltypen (Algorithmen) aus, die in den Vergleich aufgenommen werden sollen. Je mehr Typen Sie auswählen, desto mehr Modelle werden erstellt und desto länger dauert die Verarbeitung.

Modelltyp. Listet die verfügbaren Algorithmen auf (siehe unten).

Modellparameter. Sie können für die einzelnen Modelltypen entweder die Standardeinstellungen verwenden oder mithilfe von **Angeben** Optionen für den jeweiligen Modelltyp auswählen. Die einzelnen Optionen ähneln den in den separaten Modellierungskonten verfügbaren Optionen, mit dem Unterschied, dass mehrere Optionen bzw. Kombinationen ausgewählt werden können. Beispiel: Beim Vergleich von neuronalen Netzmodellen können Sie, anstatt eine der sechs Trainingsmethoden auszuwählen, alle sechs auswählen, um sechs Modelle in einem einzigen Durchgang zu trainieren.

Anzahl der Modelle. Listet die Anzahl der Modelle auf, die auf der Grundlage der aktuellen Einstellungen für die einzelnen Algorithmen erstellt wurden. Bei einer Kombination von Optionen kann die Anzahl der Modelle schnell recht groß werden. Daher wird dringend empfohlen, auf diesen Wert zu achten, insbesondere bei Verwendung großer Datasets.

Maximale Zeit für Erstellung eines einzelnen Modells beschränken. (Nur für K-Means-, Kohonen-, TwoStep-, SVM-, KNN-, Bayes Net- und Entscheidungslistenmodelle) Legt ein maximales Zeitlimit für jedes beliebige Modell fest. Wenn beispielsweise das Training für ein bestimmtes Modell aufgrund einer komplexen Interaktion unerwartet viel Zeit in Anspruch nimmt, wird durch diese Option vermieden, dass das Modell den gesamten Modellierungsdurchlauf aufhält.

Hinweis: Wenn das Ziel ein nominales Feld (Setfeld) ist, ist die Option "Entscheidungsliste" nicht verfügbar.

Unterstützte Algorithmen

Der Netzknoten verwendet ein vereinfachtes Modell der Art und Weise, wie ein menschliches Gehirn Informationen verarbeitet. Es funktioniert, indem eine große Anzahl miteinander verbundener einfacher Verarbeitungseinheiten simuliert wird, die abstrakten Versionen von Neuronen ähnlich sind. Neuronale Netze sind leistungsstarke Mehrzweckschätzer, für deren Training und Anwendung nur sehr geringe statistische oder mathematische Kenntnisse erforderlich sind.

Der C5.0-Knoten erstellt entweder einen Entscheidungsbaum oder ein Regelset. Das Modell teilt die Stichprobe auf der Basis des Felds auf, das auf der jeweiligen Ebene den maximalen Informationsgewinn liefert. Das Zielfeld muss kategorial sein. Es sind mehrere Aufteilungen in mehr als zwei Untergruppen zulässig.

Der Knoten für Klassifizierungs- und Regressionsbäume (C&R-Bäume) erstellt einen Entscheidungsbaum, mit dem Sie zukünftige Beobachtungen vorhersagen oder klassifizieren können. Bei dieser Methode wird eine rekursive Partitionierung verwendet, um die Trainingsdatensätze in Segmente aufzuteilen. Dabei wird bei jedem Schritt die Unreinheit verringert. Ein Knoten im Baum wird als "rein" betrachtet, wenn 100 % der Fälle im Knoten in eine bestimmte Kategorie des Zielfelds fallen. Ziel- und Eingabefelder können numerische Bereiche oder kategorial (nominal, ordinal oder Flags) sein. Alle Aufteilungen sind binär (nur zwei Untergruppen).

Der QUEST-Knoten bietet eine binäre Klassifizierungsmethode zum Erstellen von Entscheidungsbäumen, die dazu dient, die für Analysen von großen C&R-Bäumen erforderliche Verarbeitungszeit zu verkürzen. Gleichzeitig soll die in den Klassifizierungsbaummethoden festgestellte Tendenz verringert werden, die darin besteht, dass Eingaben bevorzugt werden, die mehr Aufteilungen erlauben. Eingabefelder können stetig (numerische Bereiche) sein, das Zielfeld muss aber kategorial sein. Alle Aufteilungen sind binär.

Der CHAID-Knoten erzeugt Entscheidungsbäume unter Verwendung von Chi-Quadrat-Statistiken zur Ermittlung optimaler Aufteilungen. Im Gegensatz zu den Knoten vom Typ "C&R-Baum" und "QUEST" kann CHAID nicht binäre Bäume generieren, d. h. Bäume mit Aufteilungen mit mehr als zwei Verzweigungen. Ziel- und Eingabefelder können in einem numerischen Bereich (stetig) oder kategorial sein. Exhaustive CHAID ist eine Änderung von CHAID, die noch gründlicher vorgeht, indem sie alle möglichen Aufteilungen untersucht, allerdings mehr Rechenzeit beansprucht.

Die logistische Regression ist ein statistisches Verfahren zur Klassifizierung von Datensätzen auf der Grundlage der Werte von Eingabefeldern. Sie ist analog zur linearen Regression, außer dass statt eines numerischen Bereichs ein kategoriales Zielfeld verwendet wird.

Der Knoten "Entscheidungsliste" kennzeichnet Untergruppen bzw. Segmente, die eine höhere oder geringere Wahrscheinlichkeit für ein bestimmtes binäres Ergebnis aufweisen als die Gesamtpopulation. Sie könnten beispielsweise nach Kunden suchen, deren Abwanderung unwahrscheinlich ist oder die mit großer Wahrscheinlichkeit positiv auf eine Kampagne reagieren. Sie können Ihr Fachwissen in das Modell integrieren, indem Sie eigene, benutzerdefinierte Segmente hinzufügen und eine Vorschau anzeigen, in der alternative Modelle nebeneinander angezeigt werden, um die Ergebnisse zu vergleichen. Entscheidungslistenmodelle bestehen aus einer Liste von Regeln, bei denen jede Regel eine Bedingung und ein Ergebnis aufweist. Regeln werden in der vorgegebenen Reihenfolge angewendet und die erste Regel, die zutrifft, bestimmt das Ergebnis.

Mithilfe des Bayes-Netzknotens können Sie ein Wahrscheinlichkeitsmodell erstellen, idem Sie beobachtete und aufgezeichnete Hinweise mit Weltwissen kombinieren, um die Wahrscheinlichkeit ihres Vorkommens zu ermitteln. Der Knoten ist speziell für Netze vom Typ "Tree Augmented Naïve Bayes" (TAN) und "Markov-Decke" gedacht, die in erster Linie zur Klassifizierung verwendet werden.

Bei der Diskriminanzanalyse werden strengere Annahmen als bei der logistischen Regression verwendet, sie kann jedoch eine wertvolle Alternative oder Ergänzung zu einer logistischen Regressionsanalyse sein, wenn diese Annahmen erfüllt sind.

Der Knoten "*k*-Nächste Nachbarn" (KNN) verknüpft einen neuen Fall mit der Kategorie oder dem Wert der *k* Objekte, die ihm im Prädiktorraum am nächsten liegen, wobei *k* eine Ganzzahl ist. Ähnliche Fälle liegen nah beieinander und Fälle mit geringer Ähnlichkeit sind weit voneinander entfernt.

Der Knoten "Support Vector Machine" (SVM) ermöglicht die Klassifizierung von Daten in eine von zwei Gruppen ohne Überanpassung. SVM eignet sich gut für umfangreiche Datasets, beispielsweise solche mit einer großen Anzahl an Eingabefeldern.

Fehlklassifizierungskosten

In einigen Zusammenhängen sind bestimmte Fehler kostspieliger als andere. Es kann beispielsweise kostspieliger sein, einen Antragsteller für einen Kredit mit einem hohen Risiko als niedriges Risiko zu klassifizieren (eine Art von Fehler) als einen Antragsteller mit einem niedrigen Risiko als hohes Risiko (eine andere Art von Fehler) zu klassifizieren. Anhand von Fehlklassifizierungskosten können Sie die relative Bedeutung verschiedener Arten von Vorhersagefehlern angeben.

Fehlklassifizierungskosten sind im Grunde Gewichtungen, die auf bestimmte Ergebnisse angewendet werden. Diese Gewichtungen werden in das Modell mit einbezogen und können die Vorhersage (als Schutz gegen kostspielige Fehler) ändern.

Mit Ausnahme von C5.0-Modellen werden Fehlklassifizierungskosten beim Scoren von Modellen nicht angewendet und bei der Rangbildung bzw. beim Vergleich von Modellen mithilfe eines Knotens vom Typ "Autom. Klassifikationsmerkmal", eines Evaluierungsdiagramms oder eines Analyseknotens nicht berücksichtigt. Ein Modell, das Kosten beinhaltet, führt nicht unbedingt zu weniger Fehlern als ein Modell, das keine Kosten beinhaltet, und es weist auch nicht unbedingt eine höhere Gesamtgenauigkeit auf, aber es ist möglicherweise in praktischer Hinsicht leistungsfähiger, da es eine integrierte Verzerrung zugunsten von *weniger teuren* Fehlern aufweist.

Die Kostenmatrix zeigt die Kosten für jede mögliche Kombination aus prognostizierter Kategorie und tatsächlicher Kategorie. Standardmäßig werden alle Fehlklassifizierungskosten auf 1,0 gesetzt. Um benutzerdefinierte Kostenwerte einzugeben, wählen Sie **Fehlklassifizierungskosten verwenden** und geben Ihre benutzerdefinierten Werte in die Kostenmatrix ein.

Um die Fehlklassifizierungskosten zu ändern, wählen Sie die Zelle aus, die der gewünschten Kombination aus vorhergesagten und tatsächlichen Werten entspricht, löschen den Inhalt der Zelle und geben die gewünschten Kosten für die Zelle ein. Die Kosten sind nicht automatisch symmetrisch. Wenn Sie z. B. die Kosten einer Fehlklassifizierung von *A* als *B* auf 2,0 setzen, weisen die Kosten für eine Fehlklassifizierung von *B* als *A* weiterhin den Standardwert 1,0 auf, es sei denn, Sie ändern diesen Wert ausdrücklich.

Knoten "Autom. Klassifikationsmerkmal" - Optionen für Verwerfen

Über die Registerkarte "Verwerfen" des Knotens "Autom. Klassifikationsmerkmal" können Sie automatisch Modelle verwerfen, die bestimmte Kriterien nicht erfüllen. Diese Modelle werden nicht im Zusammenfassungsbericht aufgeführt.

Sie können eine Untergrenze für "Gesamtgenauigkeit" sowie eine Obergrenze für die Anzahl der im Modell verwendeten Variablen festlegen. Für Flagziele können Sie zusätzlich eine Untergrenze für "Lift", "Profit" und "Fläche unter der Kurve" angeben. Hierbei werden Lift und Profit gemäß den Festlegungen im Modellierungsknoten bestimmt. Weitere Informationen finden Sie im Thema „Knoten "Autom. Klassifikationsmerkmal" - Modelloptionen” auf Seite 65.

Optional können Sie den Knoten so konfigurieren, dass die Ausführung gestoppt wird, sobald erstmals ein Modell generiert wurde, das alle angegebenen Kriterien erfüllt. Weitere Informationen finden Sie im Thema „Knoten für die automatisierte Modellierung - Stoppregeln” auf Seite 64.

Knoten "Autom. Klassifikationsmerkmal" - Einstellungsoptionen

Über die Registerkarte "Einstellungen" des Knotens "Autom. Klassifikationsmerkmal" können Sie die auf dem Nugget verfügbaren Optionen für die Scorezeit vorkonfigurieren.

Ensemble-Methode. Für Ziele können Sie aus den folgenden Ensemble-Methoden wählen:

- Voting
- Nach Konfidenz gewichtetes Voting
- Nach Raw Propensity gewichtetes Voting (nur bei Flagzielen).
- Höchste Konfidenz hat Vorrang
- Durchschnittliche Raw Propensity (nur bei Flagzielen)

Wenn Voting gebunden ist, Wert auswählen mithilfe von. Bei Voting-Methoden können Sie auswählen, wie Gleichstände aufgelöst werden sollen:

- **Zufallsauswahl.** Einer der gebundenen Werte (Werte mit Gleichstand) wird nach dem Zufallsprinzip ausgewählt.
- **Höchste Konfidenz.** Der gebundene Wert, der mit der höchsten Konfidenz vorhergesagt wurde, gewinnt. Beachten Sie, dass es sich hierbei nicht unbedingt um die höchste Konfidenz aller vorhergesagten Werte handelt.
- **Raw Propensity.** (Nur bei Flagzielen.) Der gebundene Wert, der mit der höchsten absoluten Neigung vorhergesagt wurde. Dabei berechnet sich die absolute Raw Propensity wie folgt:

```
abs(0,5 - Propensity) * 2
```

Knoten "Autonumerisch"

Der Knoten "Autonumerisch" schätzt und vergleicht Modelle für Ergebnisse stetiger numerischer Bereiche mithilfe einer Reihe verschiedener Methoden, wodurch Sie eine Vielzahl von Ansätzen in einer einzelnen Modellierungsausführung ausprobieren können. Sie können die gewünschten Algorithmen auswählen und mit mehreren Optionskombinationen experimentieren. Beispielsweise könnten Sie Immobilienwerte mithilfe von Modellen vom Typ "Neuronales Netz", "Lineare Regression", "C&R-Baum" und "CHAID" vorhersagen, um zu ermitteln, welches Modell die beste Leistung erbringt, und Sie könnten verschiedene Kombinationen der Regressionsmethoden "Schrittweise", "Vorwärts" und "Rückwärts" ausprobieren. Der Knoten untersucht jede mögliche Kombination von Optionen, stuft jedes in Frage kommende Modell auf der Basis des angegebenen Werts und speichert die geeignetsten Kombinationen in Scoring oder weiterer Analyse. Weitere Informationen finden Sie in Kapitel 5, „Knoten für die automatisierte Modellierung", auf Seite 63.

Beispiel. Eine Gemeinde möchte die Immobiliensteuern mit größerer Genauigkeit schätzen und Werte für bestimmte Immobilien nach Bedarf anpassen, ohne jedes einzelne Anwesen besichtigen zu müssen. Mithilfe des Knotens "Autonumerisch" kann der Analyst eine Reihe von Modellen generieren und vergleichen, die Immobilienwerte basierend auf dem Gebäudetyp, der Nachbarschaft, der Größe und anderen bekannten Faktoren vorhersagen.

Anforderungen. Ein einzelnes Zielfeld (mit der Rolle **Ziel**) und mindestens ein Eingabefeld (mit der Rolle **Eingabe**). Beim Ziel muss es sich um ein stetiges Feld (numerischer Bereich) handeln, beispielsweise *Alter* oder *Einkommen*. Eingabefelder können stetig oder kategorial sein, mit der Einschränkung, dass einige Eingaben für bestimmte Modelltypen nicht geeignet sind. So können beispielsweise Modelle vom Typ "C &R-Baum" kategoriale Zeichenfolgenfelder als Eingaben verwenden, während lineare Regressionsmodelle diese Felder nicht verwenden können und sie ignorieren, wenn sie angegeben sind. Die Anforderungen sind dieselben wie bei Verwendung der einzelnen Modellierungsknoten. So funktioniert beispielsweise ein CHAID-Modell immer auf dieselbe Weise, unabhängig davon, ob es aus einem CHAID-Knoten oder aus einem Knoten vom Typ "Autonumerisch" generiert wurde.

Häufigkeits- und Gewichtungsfelder. Häufigkeit und Gewichtung dienen dazu, einigen Datensätzen eine größere Bedeutsamkeit zu verleihen als anderen, beispielsweise, weil der Benutzer weiß, dass ein Teil der übergeordneten Grundgesamtheit im erstellten Dataset unterrepräsentiert ist (Gewichtung) oder weil ein Datensatz für eine Reihe identischer Fälle steht (Häufigkeit). Häufigkeitsfelder können, sofern angegeben, von Algorithmen vom Typ "C&R-Baum" und "CHAID" verwendet werden. Gewichtungsfelder können von Algorithmen vom Typ "C&R-Baum", "CHAID" und "GenLin" verwendet werden. Andere Modelltypen ignorieren diese Felder und erstellen die Modelle in jedem Fall. Häufigkeits- und Gewichtungsfelder werden nur für die Modellerstellung verwendet. Bei der Evaluation bzw. beim Scoren von Modellen werden sie nicht berücksichtigt. Weitere Informationen finden Sie im Thema „Verwenden von Häufigkeits- und Gewichtungsfeldern" auf Seite 33.

Unterstützte Modelltypen

Folgende Modelltypen werden unterstützt: "Netz", "C&R-Baum", "CHAID", "Regression", "GenLin", "Nächster Nachbar" und SVM. Weitere Informationen finden Sie im Thema „Knoten "Autonumerisch" - Expertenoptionen" auf Seite 73.

Knoten "Autonumerisch" - Modelloptionen

Auf der Registerkarte "Modell" des Knotens "Autonumerisch" können Sie die Anzahl der zu speichernden Modelle sowie die zum Vergleichen der Modelle verwendeten Kriterien angeben.

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Modelle in Ränge einteilen nach. Legt fest, welche Kriterien zum Vergleichen von Modellen verwendet werden sollen.

- **Korrelation.** Die Pearson-Korrelation zwischen dem beobachteten Wert für die einzelnen Datensätze und dem vom Modell vorhergesagten Wert. Die Korrelation ist ein Maß für den linearen Zusammenhang zwischen zwei Variablen. Dabei deuten Werte nahe bei 1 auf eine stärkere Beziehung hin. (Korrelationswerte liegen im Bereich zwischen -1 für eine perfekte negative Beziehung und +1 für eine perfekte positive Beziehung. Der Wert 0 bedeutet, dass keine lineare Beziehung besteht. Ein Modell mit einer negativen Korrelation weist den niedrigsten Rang auf.)
- **Anzahl der Felder.** Die Anzahl der Felder, die als Prädiktoren im Modell verwendet werden. Durch die Auswahl von Modellen mit weniger Feldern lässt sich in einigen Fällen die Datenvorbereitung rationalisieren und die Leistung verbessern.
- **Relativer Fehler.** Der relative Fehler ist der Quotient aus der Varianz der beobachteten Werte von den vom Modell vorhergesagten Werten und der Varianz der beobachteten Werte vom Mittelwert. Es wird also verglichen, wie gut die Leistungsfähigkeit des Modells in Bezug auf ein **Null**-Modell (leeres Modell) oder ein Modell nur mit dem **konstanten Term** ist, das einfach den Mittelwert des Zielfelds als Vorhersage ergibt. Bei einem guten Modell sollte dieser Wert kleiner als 1 sein, was darauf hinweist, dass das Modell genauer als das Nullmodell ist. Ein Modell mit einem relativen Fehler von mehr als 1 ist weniger genau als das Nullmodell und daher nicht brauchbar. Bei linearen Regressionsmodellen ist der relative Fehler gleich dem Quadrat der Korrelation und fügt ebenfalls keine neuen Informationen hinzu. Bei nicht linearen Modellen steht der relative Fehler in keinem Zusammenhang zur Korrelation und bietet ein zusätzliches Maß für die Bewertung der Leistungsfähigkeit des Modells.

Modelle in Ränge einteilen mithilfe von. Wenn eine Partition verwendet wird, können Sie angeben, ob die Ränge auf der Trainingspartition oder auf der Testpartition beruhen sollen. Bei großen Datasets lässt sich die Leistungsfähigkeit durch die Verwendung einer Partition für ein erstes Screening der Modelle u. U. erheblich verbessern.

Anzahl der zu verwendenden Modelle. Legt die maximale Anzahl der Modelle fest, die in dem vom Knoten erstellten Modellnugget aufgeführt werden sollen. Die Modelle mit dem höchsten Rang werden gemäß dem angegebenen Rangordnungskriterium aufgeführt. Durch Erhöhen dieses Grenzwerts können Sie Ergebnisse für einen größeren Anteil an Modellen vergleichen. Allerdings kann dadurch auch die Verarbeitungsgeschwindigkeit sinken. Der höchste zulässige Wert ist 100.

Prädiktoreinfluss berechnen. Bei Modellen, die zu einem angemessenen Maß an Wichtigkeit führen, können Sie ein Diagramm anzeigen, in dem die relative Wichtigkeit der einzelnen Prädiktoren bei der Modellschätzung angegeben wird. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und diejenigen Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass sich bei einigen Modellen der Zeitaufwand für die Berechnung durch den Prädiktoreinfluss erhöhen kann. Außerdem wird diese Option nicht empfohlen, wenn Sie einfach einen allgemeinen Vergleich zwischen vielen verschiedenen Modellen wünschen. Diese Option ist von größerem Nutzen, wenn die Analyse bereits auf eine Handvoll Modelle eingeengt wurde, die detaillierter untersucht werden sollen. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Modelle nicht behalten, wenn. Dient zur Angabe von Schwellenwerten für Korrelation, relativen Fehler und Anzahl der verwendeten Felder. Modelle, die eines dieser Kriterien nicht erfüllen, werden verworfen und nicht im Zusammenfassungsbericht aufgeführt.

- **Korrelation kleiner als.** Die minimale Korrelation (als absoluter Wert), die gegeben sein muss, damit ein Modell in den Zusammenfassungsbericht aufgenommen wird.
- **Anzahl der verwendeten Felder ist größer als.** Die maximale Anzahl an Feldern, die von einem aufzunehmenden Modell verwendet werden können.
- **Relativer Fehler ist größer als.** Der maximale relative Fehler, der für ein aufzunehmendes Modell zulässig ist.

Optional können Sie den Knoten so konfigurieren, dass die Ausführung gestoppt wird, sobald erstmals ein Modell generiert wurde, das alle angegebenen Kriterien erfüllt. Weitere Informationen finden Sie im Thema „Knoten für die automatisierte Modellierung - Stoppregeln" auf Seite 64.

Knoten "Autonumerisch" - Expertenoptionen

Auf der Registerkarte "Experten" des Knotens "Autonumerisch" können Sie die zu verwendenden Algorithmen und Optionen auswählen und Stoppregeln angeben.

Modelle auswählen. Standardmäßige werden alle Modelle zum Erstellen ausgewählt. Wenn Sie jedoch Analytic Server installiert haben, können Sie angeben, dass nur die Modelle ausgewählt werden, die mit Analytic Server ausgeführt werden können. Sie können diese Modelle so vordefinieren, dass sie aufgeteilte Modelle erstellen oder sehr große Datasets verarbeiten können.

Verwendete Modelle. Wählen Sie anhand der Kontrollkästchen in der Spalte auf der linken Seite die Modelltypen (Algorithmen) aus, die in den Vergleich aufgenommen werden sollen. Je mehr Typen Sie auswählen, desto mehr Modelle werden erstellt und desto länger dauert die Verarbeitung.

Modelltyp. Listet die verfügbaren Algorithmen auf (siehe unten).

Modellparameter. Sie können für die einzelnen Modelltypen entweder die Standardeinstellungen verwenden oder mithilfe von **Angeben** Optionen für den jeweiligen Modelltyp auswählen. Die einzelnen Optionen ähneln den in den separaten Modellierungskonten verfügbaren Optionen, mit dem Unterschied, dass mehrere Optionen bzw. Kombinationen ausgewählt werden können. Beispiel: Beim Vergleich von neuronalen Netzmodellen können Sie, anstatt eine der sechs Trainingsmethoden auszuwählen, alle sechs auswählen, um sechs Modelle in einem einzigen Durchgang zu trainieren.

Anzahl der Modelle. Listet die Anzahl der Modelle auf, die auf der Grundlage der aktuellen Einstellungen für die einzelnen Algorithmen erstellt wurden. Bei einer Kombination von Optionen kann die Anzahl der Modelle schnell recht groß werden. Daher wird dringend empfohlen, auf diesen Wert zu achten, insbesondere bei Verwendung großer Datasets.

Maximale Zeit für Erstellung eines einzelnen Modells beschränken. (Nur für K-Means-, Kohonen-, TwoStep-, SVM-, KNN-, Bayes Net- und Entscheidungslistenmodelle) Legt ein maximales Zeitlimit für jedes beliebige Modell fest. Wenn beispielsweise das Training für ein bestimmtes Modell aufgrund einer komplexen Interaktion unerwartet viel Zeit in Anspruch nimmt, wird durch diese Option vermieden, dass das Modell den gesamten Modellierungsdurchlauf aufhält.

Unterstützte Algorithmen

Der Netzknoten verwendet ein vereinfachtes Modell der Art und Weise, wie ein menschliches Gehirn Informationen verarbeitet. Es funktioniert, indem eine große Anzahl miteinander verbundener einfacher Verarbeitungseinheiten simuliert wird, die abstrakten Versionen von Neuronen ähnlich sind. Neuronale Netze sind leistungsstarke Mehrzweckschätzer, für deren Training und Anwendung nur sehr geringe statistische oder mathematische Kenntnisse erforderlich sind.

Der Knoten für Klassifizierungs- und Regressionsbäume (C&R-Bäume) erstellt einen Entscheidungsbaum, mit dem Sie zukünftige Beobachtungen vorhersagen oder klassifizieren können. Bei dieser Methode wird eine rekursive Partitionierung verwendet, um die Trainingsdatensätze in Segmente aufzuteilen. Dabei wird bei jedem Schritt die Unreinheit verringert. Ein Knoten im Baum wird als "rein" betrachtet, wenn 100 % der Fälle im Knoten in eine bestimmte Kategorie des Zielfelds fallen. Ziel- und Eingabefelder können numerische Bereiche oder kategorial (nominal, ordinal oder Flags) sein. Alle Aufteilungen sind binär (nur zwei Untergruppen).

Der CHAID-Knoten erzeugt Entscheidungsbäume unter Verwendung von Chi-Quadrat-Statistiken zur Ermittlung optimaler Aufteilungen. Im Gegensatz zu den Knoten vom Typ "C&R-Baum" und "QUEST" kann CHAID nicht binäre Bäume generieren, d. h. Bäume mit Aufteilungen mit mehr als zwei Verzweigungen. Ziel- und Eingabefelder können in einem numerischen Bereich (stetig) oder kategorial sein. Exhaustive CHAID ist eine Änderung von CHAID, die noch gründlicher vorgeht, indem sie alle möglichen Aufteilungen untersucht, allerdings mehr Rechenzeit beansprucht.

Die lineare Regression ist ein statistisches Verfahren zur Zusammenfassung von Daten und die Erstellung von Vorhersagen durch Anpassung einer geraden Linie oder Fläche, mit der die Diskrepanzen zwischen den vorhergesagten und den tatsächlichen Ausgabewerten minimiert werden.

Das verallgemeinerte lineare Modell erweitert das allgemeine lineare Modell so, dass die abhängige Variable über eine angegebene Verknüpfungsfunktion in linearem Zusammenhang zu den Faktoren und Kovariaten steht. Außerdem ist es mit diesem Modell möglich, dass die abhängige Variable eine von der Normalverteilung abweichende Verteilung aufweist. Es deckt die Funktionen einer großen Bandbreite an Statistikmodellen ab, darunter lineare Regression, logistische Regression, loglineare Modelle für Häufigkeitsdaten und Überlebensmodelle mit Intervallzensierung.

Der Knoten "*k*-Nächste Nachbarn" (KNN) verknüpft einen neuen Fall mit der Kategorie oder dem Wert der *k* Objekte, die ihm im Prädiktorraum am nächsten liegen, wobei *k* eine Ganzzahl ist. Ähnliche Fälle liegen nah beieinander und Fälle mit geringer Ähnlichkeit sind weit voneinander entfernt.

Der Knoten "Support Vector Machine" (SVM) ermöglicht die Klassifizierung von Daten in eine von zwei Gruppen ohne Überanpassung. SVM eignet sich gut für umfangreiche Datasets, beispielsweise solche mit einer großen Anzahl an Eingabefeldern.

Bei linearen Regressionsmodellen wird ein stetiges Ziel auf der Basis linearer Beziehungen zwischen dem Ziel und einem oder mehreren Prädiktoren vorhergesagt.

Knoten "Autonumerisch" - Einstellungsoptionen

Auf der Registerkarte "Einstellungen" des Knotens "Autonumerisch" können Sie die auf dem Nugget verfügbaren Optionen für die Scorezeit vorkonfigurieren.

Standardfehler berechnen. Für ein stetiges Ziel (numerischer Bereich) wird standardmäßig eine Standardfehlerberechnung durchgeführt, um den Unterschied zwischen den gemessenen oder geschätzten Werten und den wahren Werten zu berechnen sowie um zu zeigen, wie hoch die Übereinstimmung dieser Schätzungen war.

Knoten "Autom. Cluster"

Mit dem Knoten "Autom. Cluster" können Sie Clustering-Modelle, die Gruppen und Datensätze mit ähnlichen Merkmalen identifizieren, schätzen und vergleichen. Die Funktionsweise des Knotens gleicht der von anderen Knoten für die automatisierte Modellierung: Sie können in einem einzigen Modellierungsdurchlauf mit mehreren Kombinationen von Optionen experimentieren. Modelle können mithilfe grundlegender Messwerte für Filterung und Rangfolge der Nützlichkeit von Clustermodellen verglichen werden, um ein Maß auf der Basis der Wichtigkeit von bestimmten Feldern zu liefern.

Clustering-Modelle werden häufig verwendet, um Gruppen zu identifizieren, die als Eingabe für nachfolgende Analysen dienen können. Beispielsweise können Sie Kundengruppen auf der Basis von demografischen Merkmalen wie Einkommen oder von den Dienstleistungen, die sie in der Vergangenheit erworben haben, als Ziel nehmen. Dies ist ohne vorherige Kenntnis über die Gruppen und deren Merkmale möglich - Sie wissen u. U. gar nicht, wie viele Gruppen gesucht sind oder welche Funktionen für ihre Definition verwendet werden sollen. Clustering-Modelle werden häufig als nicht überwachte Lernmodelle bezeichnet, da sie kein Zielfeld verwenden und keine bestimmte Vorhersage liefern, die sich als "wahr" oder "falsch" bewerten lässt. Der Wert eines Clustering-Modells wird durch die Möglichkeit bestimmt, interessante Gruppierungen in den Daten zu erfassen und sinnvolle Beschreibungen dieser Gruppierungen zu liefern. Weitere Informationen finden Sie in Kapitel 11, „Clustermodelle”, auf Seite 215.

Anforderungen. Eines oder mehrere Felder, die relevante Merkmale definieren. Clustermodelle verwenden Zielfelder nicht auf die gleiche Weise wie andere Modelle, da die keine spezifischen Vorhersagen treffen, die sich als wahr oder falsch bewerten lassen. Stattdessen werden sie verwendet, um Gruppen von Fällen zu identifizieren, die möglicherweise zusammenhängen. Beispielsweise können Sie anhand eines Clustermodells nicht vorhersagen, ob ein bestimmter Kunde positiv oder negativ auf ein Angebot reagiert. Sie können jedoch ein Clustermodell verwenden, um Kunden Gruppen basierend auf ihrer Tendenz zu einer bestimmten Reaktion zuzuweisen. Gewichtungs- und Häufigkeitsfelder werden nicht verwendet.

Evaluierungsfelder. Wenn kein Ziel verwendet wird, können Sie optional eines oder mehrere Evaluierungsfelder für den Vergleich von Modellen angeben. Der Nutzen eines Clustermodells kann dadurch bewertet werden, dass gemessen wird, wie gut (oder schlecht) die Cluster diese Felder differenzieren.

Unterstützte Modelltypen

Unterstützte Modelltypen sind "TwoStep", "K-Means" und "Kohonen".

Knoten "Autom. Cluster" - Modelloptionen

Über die Registerkarte "Modell" des Knotens "Autom. Cluster" können Sie die Anzahl der zu speichernden Modelle sowie die zum Vergleichen der Modelle verwendeten Kriterien angeben.

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Modelle in Ränge einteilen nach. Legt fest, welche Kriterien für Vergleich und Rangordnung von Modellen verwendet werden sollen.

- **Silhouette.** Ein Index zur Messung von Clusterzusammenhalt und -abgrenzung. Weitere Informationen finden Sie unter *Rangeinteilung mit Silhouetten*.
- **Anzahl der Cluster.** Die Anzahl der Cluster im Modell.
- **Größe des kleinsten Clusters.** Die kleinste Clustergröße.

- **Größe des größten Clusters.** Die größte Clustergröße.
- **Kleinster/größter Cluster.** Das Größenverhältnis zwischen dem kleinsten und dem größten Cluster.
- **Wichtigkeit.** Die Bedeutung des Felds **Evaluierung** auf der Registerkarte **Felder**. Beachten Sie, dass dies nur berechnet werden kann, wenn das Feld **Evaluierung** angegeben wurde.

Modelle in Ränge einteilen mithilfe von. Wenn eine Partition verwendet wird, können Sie angeben, ob die Ränge auf dem Trainingsdataset oder auf dem Testdataset beruhen sollen. Bei großen Datasets lässt sich die Leistungsfähigkeit durch die Verwendung einer Partition für ein erstes Screening der Modelle u. U. erheblich verbessern.

Anzahl an zu behaltenden Modellen. Legt die maximale Anzahl der Modelle fest, die in dem vom Knoten erstellten Nugget aufgeführt werden sollen. Die Modelle mit dem höchsten Rang werden gemäß dem angegebenen Rangordnungskriterium aufgeführt. Beachten Sie, dass eine Erhöhung dieses Grenzwerts die Leistungsgeschwindigkeit verringern kann. Der höchste zulässige Wert ist 100.

Rangeinteilung mit Silhouetten

Die standardmäßige Rangeinteilung mit Silhouetten verwendet einen Standardwert von 0, da ein Wert kleiner 0 (also ein negativer Wert) angibt, dass der durchschnittliche Abstand zwischen einem Fall und Punkten in seinem zugeordneten Cluster größer ist als der minimale durchschnittliche Abstand zu Punkten in einem anderen Cluster. Daher können Modelle mit einer negativen Silhouette einfach verworfen werden.

Die Rangeinteilung ist eigentlich ein modifizierter Silhouetten-Koeffizient, der die Konzepte von Clusterzusammenhalt (Favorisierung von Modellen mit eng zusammengehörenden Clustern) und Clusterabgrenzung (Favorisierung von Modellen mit stark separierten Clustern) kombiniert. Der durchschnittliche Silhouetten-Koeffizient ist einfach der Durchschnitt für alle Fälle der folgenden Berechnung für jeden Einzelfall:

$(B - A) / \max(A, B)$

A ist die Entfernung zwischen Fall und Zentroid des Clusters, zu dem der Fall gehört und *B* ist der minimale Abstand zwischen Fall und Zentroid jedes anderen Clusters.

Der Silhouettenkoeffizient (und sein Durchschnittswert) liegen zwischen -1 (stellvertretend für ein sehr schlechtes Modell) und 1 (stellvertretend für ein ausgezeichnetes Modell). Der Durchschnitt kann auf der Ebene aller Fälle (führt zu einer Gesamtsilhouette) oder auf Clusterebene (führt zu einer Clustersilhouette) durchgeführt werden. Entfernungen können mithilfe euklidischer Entfernungen berechnet werden.

Knoten "Autom. Cluster" - Expertenoptionen

Über die Registerkarte "Experten" des Knotens "Autom. Cluster" können Sie eine Partition (falls verfügbar) anwenden, die zu verwendenden Algorithmen auswählen und Stoppregeln angeben.

Verwendete Modelle. Wählen Sie anhand der Kontrollkästchen in der Spalte auf der linken Seite die Modelltypen (Algorithmen) aus, die in den Vergleich aufgenommen werden sollen. Je mehr Typen Sie auswählen, desto mehr Modelle werden erstellt und desto länger dauert die Verarbeitung.

Modelltyp. Listet die verfügbaren Algorithmen auf (siehe unten).

Modellparameter. Sie können für die einzelnen Modelltypen entweder die Standardeinstellungen verwenden oder mithilfe von **Angeben** Optionen für den jeweiligen Modelltyp auswählen. Die einzelnen Optionen ähneln den in den separaten Modellierungskonten verfügbaren Optionen, mit dem Unterschied, dass mehrere Optionen bzw. Kombinationen ausgewählt werden können. Beispiel: Beim Vergleich von neuronalen Netzmodellen können Sie, anstatt eine der sechs Trainingsmethoden auszuwählen, alle sechs auswählen, um sechs Modelle in einem einzigen Durchgang zu trainieren.

Anzahl der Modelle. Listet die Anzahl der Modelle auf, die auf der Grundlage der aktuellen Einstellungen für die einzelnen Algorithmen erstellt wurden. Bei einer Kombination von Optionen kann die Anzahl der Modelle schnell recht groß werden. Daher wird dringend empfohlen, auf diesen Wert zu achten, insbesondere bei Verwendung großer Datasets.

Maximale Zeit für Erstellung eines einzelnen Modells beschränken. (Nur für K-Means-, Kohonen-, TwoStep-, SVM-, KNN-, Bayes Net- und Entscheidungslistenmodelle) Legt ein maximales Zeitlimit für jedes beliebige Modell fest. Wenn beispielsweise das Training für ein bestimmtes Modell aufgrund einer komplexen Interaktion unerwartet viel Zeit in Anspruch nimmt, wird durch diese Option vermieden, dass das Modell den gesamten Modellierungsdurchlauf aufhält.

Unterstützte Algorithmen

Der K-Means-Knoten teilt das Dataset in unterschiedliche Gruppen (oder Cluster) auf. Bei dieser Methode wird eine festgelegte Anzahl von Clustern definiert, den Clustern werden iterativ Datensätze zugewiesen und die Clusterzentren werden angepasst, bis eine weitere Verfeinerung keine wesentliche Verbesserung des Modells mehr darstellen würde. Statt zu versuchen, ein Ergebnis vorherzusagen, versucht *K*-Means mithilfe eines als "nicht überwachtes Lernen" bezeichneten Prozesses Muster im Set der Eingabefelder zu entdecken.

Der Kohonen-Knoten erstellt eine Art von neuronalem Netz, das verwendet werden kann, um ein Clustering des Datasets in einzelne Gruppen vorzunehmen. Wenn das Netz voll trainiert ist, sollten ähnliche Datensätze auf der Ausgabekarte eng nebeneinander stehen, während Datensätze, die sich unterscheiden, weit voneinander entfernt sein sollten. Die Zahl der von jeder Einheit im Modellnugget erfassten Beobachtungen gibt Aufschluss über die starken Einheiten. Dadurch wird ein Eindruck von der ungefähren Zahl der Cluster vermittelt.

Der TwoStep-Knoten verwendet eine aus zwei Schritten bestehende Clusterbildungsmethode. Im ersten Schritt wird ein einzelner Durchlauf durch die Daten vorgenommen, bei dem die Eingangsrohdaten zu einem verwaltbaren Set von Subclustern komprimiert werden. Im zweiten Schritt werden die Subcluster mithilfe einer hierarchischen Methode zur Clusterbildung nach und nach in immer größere Cluster zusammengeführt. TwoStep hat den Vorteil, dass die optimale Anzahl an Clustern für die Trainingsdaten automatisch geschätzt wird. Mit dem Verfahren können gemischte Feldtypen und große Datasets effizient verarbeitet werden.

Knoten "Autom. Cluster" - Optionen für Verwerfen

Über die Registerkarte "Verwerfen" des Knotens "Autom. Cluster" können Sie automatisch Modelle verwerfen, die bestimmte Kriterien nicht erfüllen. Diese Modelle werden nicht auf dem Modellnugget aufgeführt.

Sie können den Silhouetten-Mindestwert, Clusterzahlen, Clustergrößen und Wichtigkeit des Evaluierungsfelds für das Modell angeben. Silhouette sowie Anzahl und Größe von Clustern richten sich nach den Angaben im Modellierungsknoten. Weitere Informationen finden Sie im Thema „Knoten "Autom. Cluster" - Modelloptionen” auf Seite 75.

Optional können Sie den Knoten so konfigurieren, dass die Ausführung gestoppt wird, sobald erstmals ein Modell generiert wurde, das alle angegebenen Kriterien erfüllt. Weitere Informationen finden Sie im Thema „Knoten für die automatisierte Modellierung - Stoppregeln” auf Seite 64.

Nugget für automatisierte Modellierung

Beim Ausführen eines Knotens für automatisierte Modellierung schätzt der Knoten in Frage kommende Modelle für alle möglichen Einstellungskombinationen, stuft jedes in Frage kommende Modell auf der Basis des angegebenen Werts und speichert die geeignetsten Modelle in einem zusammengesetzten Nugget für automatisierte Modellierung. Dieses Modellnugget enthält ein Set aus einem oder mehreren vom

Knoten generierten Modellen, die individuell durchgesehen oder zur Verwendung beim Scoring ausgewählt werden können. Für jedes Modell werden Modelltyp und Erstellungszeit zusammen mit der Anzahl anderer Maße passend zum Modelltyp aufgeführt. Sie können die Tabelle nach einer beliebigen Spalte sortieren, um rasch die interessantesten Modelle ermitteln zu können.

- Zum Durchsehen von einem der einzelnen Modellnuggets doppelklicken Sie auf das Nuggetsymbol. Sie können dann einen Modellknoten für dieses Modell im Streamerstellungsbereich generieren oder eine Kopie des Modellnuggets in der Modellpalette.
- Piktogramme von Diagrammen erlauben eine rasche visuelle Einschätzung für jeden Modelltyp, wie unten zusammengefasst. Durch Doppelklicken auf einem Piktogramm können Sie ein Diagramm in voller Größe generieren. Das Diagramm in voller Größe zeigt bis zu 1.000 Punkten und beruht auf einer Stichprobe, wenn das Dataset mehr Punkte enthält. (Bei Streudiagrammen wird das Diagramm jedes Mal, wenn es angezeigt wird, neu generiert, sodass alle Änderungen in den Daten oberhalb des Knotens, wie beispielsweise die Aktualisierung einer Zufallsstichprobe oder Partition, wenn **Startwert für Zufallsgenerator festlegen** nicht ausgewählt ist, jedes Mal berücksichtigt werden, wenn das Streudiagramm erneut gezeichnet wird.)
- Mithilfe der Symbolleiste können Sie bestimmte Spalten auf der Registerkarte "Modell" ein- bzw. ausblenden oder die Spalte ändern, die für die Sortierung der Tabelle verwendet wird. (Außerdem können Sie die Sortierung durch Klicken auf die Spaltenüberschriften ändern.)
- Mit der Schaltfläche "Löschen" können Sie nicht verwendete Modelle permanent entfernen.
- Um die Anordnung der Spalten zu ändern, klicken Sie auf eine Spaltenüberschrift und ziehen Sie die Spalte an die gewünschte Position.
- Wenn eine Partition gerade verwendet wird, können Sie die Ergebnisse für die Trainings- bzw. Testpartition anzeigen lassen.

Die jeweils verwendeten Spalten hängen vom Typ der zu vergleichenden Modelle ab, wie unten angegeben.

Binäre Ziele

- Bei binären Modellen zeigt das Piktogramm die Verteilung der tatsächlichen Werte, überlagert mit den vorhergesagten Werten, und bietet so einen schnellen Überblick darüber, wie viele Datensätze in den einzelnen Kategorien korrekt vorhergesagt wurden.
- Rangordnungskriterien entsprechen den Optionen im Knoten "Autom. Klassifikationsmerkmal". Weitere Informationen finden Sie im Thema „Knoten "Autom. Klassifikationsmerkmal" - Modelloptionen” auf Seite 65.
- Für den maximalen Profit wird außerdem das Perzentil angegeben, in dem dieses Maximum auftritt.
- Beim kumulativen Lift können Sie das ausgewählte Perzentil mithilfe der Symbolleiste ändern.

Nominale Ziele

- Bei nominalen Modellen (Setmodellen) zeigt das Piktogramm die Verteilung der tatsächlichen Werte, überlagert mit den vorhergesagten Werten, und bietet so einen schnellen Überblick darüber, wie viele Datensätze in den einzelnen Kategorien korrekt vorhergesagt wurden.
- Rangordnungskriterien entsprechen den Optionen im Knoten "Autom. Klassifikationsmerkmal". Weitere Informationen finden Sie im Thema „Knoten "Autom. Klassifikationsmerkmal" - Modelloptionen” auf Seite 65.

Stetige Ziele

- Bei stetigen Modelle (numerischer Bereich) bietet das Diagramm für jedes Modell eine grafische Darstellung der vorhergesagten gegenüber den beobachteten Werten und ermöglicht einen schnellen Überblick über die Korrelation zwischen diesen Werten. Bei einem guten Modell sollten die Punkte tendenziell entlang der Diagonalen gruppiert und nicht zufällig im ganzen Diagramm verstreut sein.
- Rangordnungskriterien entsprechen den Optionen im Knoten "Autonumerisch". Weitere Informationen finden Sie im Thema „Knoten "Autonumerisch" - Modelloptionen” auf Seite 71.

Clusterziele.

- Bei Clustermodellen bietet das Diagramm für jedes Modell eine grafische Darstellung von Zahlen gegenüber Clustern und ermöglicht einen schnellen Überblick über die Clusterverteilung.
- Rangordnungskriterien entsprechen den Optionen im Knoten "Autom. Cluster". Weitere Informationen finden Sie im Thema „Knoten "Autom. Cluster" - Modelloptionen” auf Seite 75.

Auswählen von Modellen zum Scoring

In der Spalte **Verwenden?** können Sie Modelle für das Scoring auswählen.

- Nur für binäre, nominale und numerische Ziele können Sie mehrere Scoring-Modelle auswählen und die Scores in einem einzigen Modellnugget kombinieren. Durch die Kombination der Vorhersagen aus mehreren Modellen lassen sich Begrenzungen, die einzelne Modelle aufweisen, vermeiden. Dadurch kann häufig eine höhere Gesamtgenauigkeit erreicht werden als mit einem der Modelle allein.
- Für Clustermodelle kann nur jeweils ein Scoring-Modell ausgewählt werden. Standardmäßig wird das Modell mit dem höchsten Rang zuerst ausgewählt.

Generieren von Knoten und Modellen

Sie können eine Kopie des zusammengesetzten Nuggets für automatisierte Modellierung oder des Knotens für automatisierte Modellierung, aus dem das Nugget erstellt wurde, generieren. Dies kann beispielsweise dann hilfreich sein, wenn Sie nicht über den Original-Stream verfügen, aus dem das Nugget für automatisierte Modellierung erstellt wurde. Für alle im Nugget für automatisierte Modellierung aufgeführten Modelle kann ein Modellnugget oder ein Modellierungsknoten generiert werden.

Nugget für automatisierte Modellierung

Wählen Sie im Modell "Generieren" die Option **Modelle in Palette**, um der Modellpalette das Nugget für automatisierte Modellierung hinzuzufügen. Das generierte Modell kann gespeichert oder in der vorliegenden Form verwendet werden, ohne dass der Stream erneut ausgeführt werden muss.

Alternativ können Sie im Menü "Generieren" die Option **Modellierungsknoten generieren** auswählen, um dem Streamerstellungsbereich den Modellierungsknoten hinzuzufügen. Anhand dieses Knotens können Sie die ausgewählten Modelle erneut schätzen, ohne den gesamten Modellierungsdurchlauf wiederholen zu müssen.

Einzelnes Modellierungsnugget

1. Doppelklicken Sie im Menü **Modell** auf das benötigte Nugget. In einem neuen Dialogfeld wird eine Kopie dieses Nuggets erstellt.
2. Wählen Sie im Menü "Generieren" des neuen Dialogfelds die Option **Modelle in Palette**, um der Modellpalette das einzelne Nugget hinzuzufügen.
3. Alternativ können Sie im Menü "Generieren" des neuen Dialogfelds die Option **Modellierungsknoten generieren** auswählen, um dem Streamerstellungsbereich den einzelnen Modellierungsknoten hinzuzufügen.

Generieren von Evaluierungsdiagrammen

Bei binären Modellen können Sie Evaluierungsdiagramme generieren, die eine visuelle Möglichkeit zur Bewertung und zum Vergleich der einzelnen Modelle bieten. Evaluierungsdiagramme sind für Modelle, die mithilfe der Knoten vom Typ "Autonumerisch" oder "Autom. Cluster" generiert wurden, nicht verfügbar.

1. Wählen Sie im Nugget für automatisierte Modellierung des Knotens "Autom. Klassifikationsmerkmal" in der Spalte *Verwenden?* die Modelle aus, die Sie auswerten möchten.
2. Wählen Sie im Menü "Generieren" die Option **Evaluierungsdiagramm(e)** aus. Das Dialogfeld "Evaluierungsdiagramm" wird angezeigt.

3. Wählen Sie den gewünschten Diagrammtyp und die anderen gewünschten Optionen.

Evaluierungsdiagramme

Über die Registerkarte "Modell" des Nuggets für automatisierte Modellierung können Sie einen Drill-Down durchführen, um einzelne Diagramme für jedes der angezeigten Modelle anzuzeigen. Für die Nuggets "Automatisches Klassifikationsmerkmal" und "Autonumerisch" werden auf der Registerkarte "Diagramm" sowohl ein Diagramm als auch den Prädiktoreinfluss angezeigt, die die Ergebnisse aller Modelle zusammen widerspiegeln. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors” auf Seite 43.

Für "Automatisches Klassifikationsmerkmal" wird ein Verteilungsdiagramm angezeigt, während für "Autonumerisch" ein Multiplot (auch Streudiagramm genannt) angezeigt wird.

Kapitel 6. Entscheidungsbäume

Entscheidungsbaummodelle

Mithilfe von Entscheidungsbaummodellen können Sie Klassifizierungssysteme entwickeln, die zukünftige Beobachtungen basierend auf einer Reihe von Entscheidungsregeln vorhersagen oder klassifizieren. Wenn die Daten in Klassen aufgeteilt sind, die Sie interessieren (z. B. Darlehen mit hohem Risiko im Gegensatz zu Darlehen mit niedrigem Risiko, Abonnenten gegenüber Personen ohne Abonnement, Wähler im Gegensatz zu Nichtwählern oder Bakterienarten), können Sie mit diesen Daten Regeln erstellen, die Sie zur Klassifizierung alter oder neuer Fälle mit maximaler Genauigkeit verwenden können. So können Sie z. B. einen Baum erstellen, der das Kreditrisiko oder die Kaufabsicht basierend auf Alter und anderen Faktoren klassifiziert.

Dieser Ansatz, manchmal bekannt als **Regelinduktion**, hat mehrere Vorteile. Zunächst wird die Argumentationskette hinter dem Modell deutlich, wenn Sie durch die Struktur blättern. Dies steht im Gegensatz zu anderen "Blackbox"-Modellierungstechniken, bei denen die interne Logik nicht so leicht zu durchschauen ist.

Zudem berücksichtigt der Prozess in seiner Regel automatisch nur die Attribute, die im Entscheidungsfindungsprozess wirklich von Bedeutung sind. Attribute, die nicht zur Genauigkeit des Baums beitragen, werden ignoriert. Dies kann zu sehr hilfreichen Informationen über die Daten führen und kann dazu verwendet werden, die Daten auf die relevanten Felder zu reduzieren, bevor ein anderes Verfahren zum Maschinenlernen trainiert wird, z. B. ein neuronales Netz.

Entscheidungsbaummodellnuggets können in eine Sammlung von Wenn-dann-Regeln (ein **Regelset**) umgewandelt werden, die die Informationen in vielen Fällen in einer verständlicheren Form darstellen. Die Entscheidungsbaumdarstellung ist nützlich, wenn Sie sehen möchten, wie die Attribute in den Daten die Gesamtheit in Subsets **teilen** oder **aufteilen**, die für das Problem relevant sind. Die Regelsetdarstellung ist dann nützlich, wenn Sie sehen möchten, in welchem Zusammenhang bestimmte Elementgruppen mit einer bestimmten Schlussfolgerung stehen. Die folgende Regel präsentiert uns z. B. ein **Profil** für eine Gruppe von Fahrzeugen, die einen Kauf wert sind:

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

Baumerstellungsalgorithmen

Vier Algorithmen sind für die Durchführung der Klassifizierungs- und Segmentierungsanalyse verfügbar. Diese Algorithmen führen im Grunde alle dieselben Operationen durch. Sie prüfen alle Felder Ihres Datasets, um das Feld zu finden, das die beste Klassifizierung oder Vorhersage liefert, indem sie die Daten in Untergruppen aufteilen. Der Vorgang wird rekursiv angewendet, wobei die Untergruppen in immer kleinere Einheiten aufgeteilt werden, bis der Baum erstellt ist (wie von bestimmten Stoppkriterien definiert). Die bei der Baumerstellung verwendeten Ziel- und Eingabefelder können je nach verwendetem Algorithmus stetig (numerischer Bereich) oder kategorial sein. Wenn ein stetiges Ziel verwendet wird, wird ein Regressionsbaum generiert; wenn ein kategoriales Ziel verwendet wird, wird ein Klassifizierungsbaum generiert.

Der Knoten für Klassifizierungs- und Regressionsbäume (C&R-Bäume) erstellt einen Entscheidungsbaum, mit dem Sie zukünftige Beobachtungen vorhersagen oder klassifizieren können. Bei dieser Methode wird eine rekursive Partitionierung verwendet, um die Trainingsdatensätze in Segmente aufzuteilen. Dabei wird bei jedem Schritt die Unreinheit verringert. Ein Knoten im Baum wird als "rein" betrachtet, wenn 100 % der Fälle im Knoten in eine bestimmte Kategorie des Zielfelds fallen. Ziel- und Eingabefelder können numerische Bereiche oder kategorial (nominal, ordinal oder Flags) sein. Alle Aufteilungen sind binär (nur zwei Untergruppen).

Der CHAID-Knoten erzeugt Entscheidungsbäume unter Verwendung von Chi-Quadrat-Statistiken zur Ermittlung optimaler Aufteilungen. Im Gegensatz zu den Knoten vom Typ "C&R-Baum" und "QUEST" kann CHAID nicht binäre Bäume generieren, d. h. Bäume mit Aufteilungen mit mehr als zwei Verzweigungen. Ziel- und Eingabefelder können in einem numerischen Bereich (stetig) oder kategorial sein. Exhaustive CHAID ist eine Änderung von CHAID, die noch gründlicher vorgeht, indem sie alle möglichen Aufteilungen untersucht, allerdings mehr Rechenzeit beansprucht.

Der QUEST-Knoten bietet eine binäre Klassifizierungsmethode zum Erstellen von Entscheidungsbäumen, die dazu dient, die für Analysen von großen C&R-Bäumen erforderliche Verarbeitungszeit zu verkürzen. Gleichzeitig soll die in den Klassifizierungsbaummethoden festgestellte Tendenz verringert werden, die darin besteht, dass Eingaben bevorzugt werden, die mehr Aufteilungen erlauben. Eingabefelder können stetig (numerische Bereiche) sein, das Zielfeld muss aber kategorial sein. Alle Aufteilungen sind binär.

Der C5.0-Knoten erstellt entweder einen Entscheidungsbaum oder ein Regelset. Das Modell teilt die Stichprobe auf der Basis des Felds auf, das auf der jeweiligen Ebene den maximalen Informationsgewinn liefert. Das Zielfeld muss kategorial sein. Es sind mehrere Aufteilungen in mehr als zwei Untergruppen zulässig.

Allgemeine Verwendung der baumbasierten Analyse

Im Folgenden werden einige allgemeine Anwendungsbereiche der baumbasierten Analyse erläutert:

Segmentierung. Ermitteln Sie Personen, die wahrscheinlich Mitglieder einer bestimmten Klasse sind.

Schichtung. Weisen Sie Fälle zu einer von mehreren Kategorien zu, z. B. Gruppen mit hohem, mittlerem oder niedrigem Risiko.

Vorhersage. Erstellen Sie Regeln und verwenden Sie sie zum Vorhersagen zukünftiger Ereignisse. Vorhersage kann auch den Versuch bezeichnen, Vorhersageattribute Werten einer stetigen Variable zuzuordnen.

Datenreduktion und Variablenscreening. Wählen Sie ein geeignetes Subset von Prädiktoren aus einer Vielzahl von Variablen aus und erstellen Sie damit ein formal parametrisches Modell.

Interaktionsidentifizierung. Ermitteln Sie Beziehungen, die nur für bestimmte Untergruppen gelten, und geben Sie diese in einem formal parametrischen Modell an.

Kategoriezusammenführung und Einteilung von stetigen Variablen. Codieren Sie Gruppenprädiktorkategorien und stetige Variablen mit minimalem Informationsverlust um.

Interactive Tree Builder

Sie können ein Baummodell entweder automatisch erstellen, indem Sie den Algorithmus auf jeder Ebene den besten Split auswählen lassen, oder Sie können den interaktiven Tree Builder nutzen und den Baum vor dem Speichern des Modellnuggets auf der Grundlage Ihres Fachwissens verfeinern oder vereinfachen.

1. Erstellen Sie einen Stream und fügen Sie einen der drei Entscheidungsbaumknoten "C&R-Baum", "CHAID" oder "QUEST" hinzu.

   Hinweis: Für C5.0-Bäume wird keine interaktive Baumerstellung unterstützt.
2. Öffnen Sie den Knoten, wählen Sie auf der Registerkarte "Felder" die Ziel- und Prädiktorfelder und geben Sie bei Bedarf zusätzliche Modelloptionen an. Spezifische Anleitungen finden Sie in der Dokumentation zu den einzelnen Baumerstellungsknoten.
3. Wählen Sie auf dem Panel "Ziel" der Registerkarte "Erstellungsoptionen" den Befehl **interaktive Sitzung starten**.
4. Klicken Sie auf **Ausführen**, um den Tree Builder zu starten.

Der aktuelle Baum wird ab dem Stammknoten angezeigt. Sie können den Baum Ebene für Ebene bearbeiten und reduzieren und, bevor Sie eines oder mehrere Modelle erstellen, auf Gewinne, Risiken und zugehörige Informationen zugreifen.

Kommentare

- Für C&R-Baum-, CHAID- und QUEST-Knoten müssen alle im Modell verwendeten ordinalen Felder numerisch (und nicht als Zeichenfolge) gespeichert sein. Im Bedarfsfall können Sie die Felder mit dem Umcodierungsknoten konvertieren.
- Optional können Sie mit einem Partitionsfeld die Daten in Trainings- und Teststichproben trennen.
- Sie können ein Modell mit Tree Builder oder, wie andere IBM SPSS Modeler-Modelle, direkt aus dem Modellierungsknoten generieren. Weitere Informationen finden Sie im Thema „Direktes Erstellen eines Baummodells” auf Seite 94.

Erweitern und Reduzieren des Baums

Auf der Registerkarte "Viewer" im Tree Builder können Sie den aktuellen Baum ab dem Stammknoten anzeigen.

1. Um den Baum zu erweitern, wählen Sie in den Menüs folgende Optionsfolge:

   Baum > **Baum erweitern**

   Das System erstellt den Baum, indem jeder Zweig so lange rekursiv aufgeteilt wird, bis ein oder mehrere Stoppkriterien erfüllt sind. Auf der Grundlage der verwendeten Modellbildungsmethode wird bei jeder Aufteilung automatisch der beste Prädiktor ausgewählt.
2. Sie können auch nur eine Ebene hinzufügen, indem Sie **Baum um eine Ebene erweitern** auswählen.
3. Um eine Verzweigung unter einem bestimmten Knoten hinzuzufügen, wählen Sie den Knoten aus und klicken Sie auf **Verzweigung erweitern**.
4. Um den für die Aufteilung verwendeten Prädiktor festzulegen, wählen Sie den gewünschten Knoten aus und klicken Sie auf **Verzweigung mit benutzerdefinierter Aufteilung erweitern**. Weitere Informationen finden Sie im Thema „Definieren benutzerdefinierter Aufteilungen” auf Seite 84.
5. Um eine Verzweigung zu reduzieren, wählen Sie einen Knoten aus und klicken Sie auf **Verzweigung entfernen**. Der ausgewählte Knoten wird gelöscht.
6. Um die untere Ebene des Baums zu entfernen, wählen Sie **Eine Ebene entfernen**.
7. Ausschließlich für C&R-und QUEST-Bäume können Sie **Struktur erweitern und reduzieren** wählen, um die Reduktion auf der Grundlage eines Kostenkomplexitätsalgorithmus durchzuführen, der die Risikoschätzung auf der Grundlage der Anzahl der Terminalknoten anpasst, was in der Regel zu einer einfacheren Struktur führt. Weitere Informationen finden Sie im Thema „C&R-Baumknoten” auf Seite 96.

Lesen von Aufteilungsregeln auf der Registerkarte "Viewer"

Bei der Anzeige von Aufteilungsregeln auf der Registerkarte "Viewer" bedeuten eckige Klammern, dass der angegebene Wert im Bereich enthalten ist, während runde Klammern anzeigen, dass der Wert aus dem Bereich ausgeschlossen ist. Der Audruck (23,37] bedeutet somit von 23 (ausgeschlossen) bis einschließlich 37, also von etwas über 23 bis 37. Auf der Registerkarte "Modell" wird dieselbe Bedingung wie folgt angezeigt:

```
Age > 23 and Age <= 37
```

Unterbrechen der Baumerweiterung. Um die Baumerweiterung zu unterbrechen (wenn er beispielsweise länger als erwartet dauert), klicken Sie in der Symbolleiste auf die Schaltfläche "Ausführung anhalten".

Abbildung 28. Schaltfläche "Ausführung anhalten"

Diese Schaltfläche ist nur während der Strukturerweiterung aktiviert. Sie stoppt den aktuellen Aufbauvorgang am gerade erreichten Punkt, wobei alle bereits hinzugefügten Knoten erhalten bleiben, ohne dass Änderungen gespeichert werden oder das Fenster geschlossen wird. Der Tree Builder bleibt geöffnet, sodass Sie nach Bedarf ein Modell erstellen, Direktiven aktualisieren oder Ausgaben im entsprechenden Format exportieren können.

Definieren benutzerdefinierter Aufteilungen

Im Dialogfeld "Aufteilung definieren" können Sie den Prädiktor auswählen und Bedingungen für die einzelnen Aufteilungen angeben.

1. Wählen Sie in Tree Builder auf der Registerkarte "Viewer" einen Knoten aus und treffen Sie in den Menüs folgende Auswahl:

   Baum > **Verzweigung mit benutzerdefinierter Aufteilung erweitern**
2. Wählen Sie in der Dropdown-Liste den gewünschten Prädiktor aus oder klicken Sie auf die Schaltfläche **Prädiktoren**, damit zu jedem Prädiktor Einzelheiten angezeigt werden. Weitere Informationen finden Sie im Thema „Anzeigen der Prädiktordetails” auf Seite 85.
3. Für jede Aufteilung können Sie die Standardbedingungen übernehmen oder über **Angepasst** die für die Aufteilung gewünschten Bedingungen festlegen.
   - Für stetige Prädiktoren (numerischer Bereich), können Sie die Felder **Bereichswerte bearbeiten** verwenden, um den Wertebereich anzugeben, der in jeden neuen Knoten fällt.
   - Für kategoriale Prädiktoren können Sie die Felder **Set-Werte bearbeiten** oder **Ordinale Werte bearbeiten** verwenden, um die spezifischen Werte (oder Wertebereiche, wenn es sich um einen ordinalen Prädiktor handelt) anzugeben, die jedem neuen Knoten zugeordnet sind.
4. Wählen Sie **Erweitern**, damit die Verzweigung mit dem ausgewählten Prädiktor noch einmal erweitert wird.

Unabhängig von Stoppregeln kann der Baum in der Regel ohne Prädiktor aufgeteilt werden. Ausgenommen sind lediglich die Fälle, wo es sich um einen reinen Knoten handelt (d. h. 100 % der Fälle fallen in dieselbe Zielklasse, wodurch nichts mehr aufgeteilt werden kann) oder wenn der ausgewählte Prädiktor konstant ist (also nichts gegen ihn aufgeteilt werden kann).

Fehlende Werte in. Nur für CHAID-Knoten gibt es bei der Festlegung einer benutzerdefinierten Aufteilung die Option, fehlende Werte eines bestimmten Prädiktors einem bestimmten untergeordneten Knoten zuzuordnen. (Bei "C&R-Baum" und "QUEST" werden fehlende Werte mit im Algorithmus definierten Ersatzfeldern bearbeitet. Weitere Informationen finden Sie im Thema „Aufteilungsdetails und Ersatztrenner” auf Seite 85.)

Anzeigen der Prädiktordetails

Das Dialogfeld "Prädiktor auswählen" zeigt Statistiken der für den aktuellen Split verfügbaren Prädiktoren (oder "Konkurrenten", wie sie zuweilen auch genannt werden).

- Bei CHAID und Exhaustive CHAID wird für jeden kategorialen Prädiktor die Chi-Quadrat-Statistik aufgeführt. Wenn es sich bei einem Prädiktor um einen numerischen Bereich handelt, wird die *F*-Statistik angezeigt. Die Chi-Quadrat-Statistik ist ein Maß dafür, wie unabhängig das Zielfeld vom Aufteilungsfeld ist. Eine hoch ausfallende Chi-Quadrat-Statistik weist in der Regel auf eine geringere Wahrscheinlichkeit hin, d. h., die Chance, dass zwei Felder unabhängig sind, ist geringer, was wiederum bedeutet, dass die Aufteilung "gut" ist. Freiheitsgrade werden aufgenommen, weil diese die Tatsache berücksichtigen, dass es bei einer dreifachen Aufteilung einfacher ist, eine hoch ausfallende Statistik und eine geringe Wahrscheinlichkeit zu erhalten, als dies bei einer zweifachen Aufteilung der Fall ist.
- Bei "C&R-Baum" und "QUEST" wird für jeden Prädiktor die Verbesserung angezeigt. Je größer die Verbesserung ist, desto stärker reduziert der Einsatz des Prädiktors die zwischen den über- und den untergeordneten Knoten entstehende Unreinheit. (Ein reiner Knoten liegt dann vor, wenn alle Fälle in eine einzige Zielkategorie fallen. Je geringer die Unreinheit des gesamten Baums ist, desto besser passt das Modell zu den Daten.) Anders ausgedrückt, weist ein hoher Verbesserungswert in der Regel auf eine brauchbare Aufteilung für diesen Baumtyp hin. Das verwendete Unreinheitsmaß wird im Baumerstellungsknoten festgelegt.

Aufteilungsdetails und Ersatztrenner

Sie können jeden Knoten auf der Registerkarte "Viewer" auswählen und über die Schaltfläche für die Aufteilungsinformationen (rechts in der Symbolleiste) die Details über die Aufteilung des Knotens anzeigen. Zusammen mit der verwendeten Aufteilungsregel werden die relevanten Statistiken angezeigt. Für kategoriale C&R-Bäume werden die Verbesserung und die Assoziation angezeigt. Die Assoziation ist ein Maß für die Entsprechung zwischen einem Ersatztrenner und dem primären Aufteilungsfeld, wobei der "beste" Ersatztrenner in der Regel derjenige ist, der das Aufteilungsfeld am stärksten imitiert. Bei "C&R-Baum" und "QUEST" werden auch alle anstelle des primären Prädiktors verwendeten Ersatztrenner aufgeführt.

Um die Aufteilung des ausgewählten Knotens zu bearbeiten, klicken Sie auf der linken Seite des Ersatztrennerbereichs auf das Symbol, damit das Dialogfeld "Aufteilung definieren" geöffnet wird. (Sie können das Verfahren abkürzen, indem Sie in der Liste einen Ersatztrenner auswählen und diesen durch Anklicken des Symbols zum primären Aufteilungsfeld machen.)

Ersatzfelder. Wo zutreffend, werden für den ausgewählten Knoten alle Ersatzfelder für das primäre Aufteilungsfeld angezeigt. Ersatzfelder sind alternative Felder, die verwendet werden, wenn der primäre Prädiktorwert für einen bestimmten Datensatz fehlt. Die maximale Anzahl an Ersatzfeldern, die für eine bestimmte Aufteilung erlaubt ist, ist im Baumerstellungsknoten angegeben, die tatsächliche Anzahl richtet sich jedoch nach den Trainingsdaten. Im Allgemeinen gilt: Je mehr fehlende Daten, desto mehr Ersatzfelder werden wahrscheinlich verwendet. Für andere Entscheidungsbaummodelle bleibt diese Registerkarte leer.

Hinweis: Damit Ersatzfelder im Modell berücksichtigt werden, müssen sie während der Trainingsphase ermittelt werden. Wenn die Trainingsstichprobe keine fehlenden Werte enthält, werden keine Ersatzfelder ermittelt. Alle Datensätze mit fehlenden Werten, die während des Testens oder der Bewertung gefunden werden, fallen automatisch in den untergeordneten Knoten mit der größten Anzahl an Datensätzen. Wenn während des Testens oder Bewertens fehlende Werte erwartet werden, können Sie sicher sein, dass auch in den Trainingsstichproben Werte fehlen. Für CHAID-Bäume sind keine Ersatzfelder verfügbar.

Obwohl bei CHAID-Knoten keine Ersatztrenner verwendet werden, haben Sie in einer benutzerdefinierten Aufteilung die Option, diese einem bestimmten untergeordneten Knoten zuzuordnen. Weitere Informationen finden Sie im Thema „Definieren benutzerdefinierter Aufteilungen” auf Seite 84.

Anpassen der Baumansicht

Auf der Registerkarte "Viewer" von Tree Builder wird der aktuelle Baum angezeigt. Standardmäßig sind alle Verzweigungen des Baums erweitert. Sie können Verzweigungen allerdings ausblenden und erweitern oder weitere Einstellungen anpassen.

- Klicken Sie auf das Minuszeichen (**-**) in der rechten unteren Ecke eines übergeordneten Knotens, um alle zugehörigen untergeordneten Knoten auszublenden. Klicken Sie auf das Pluszeichen (**+**) in der rechten unteren Ecke eines übergeordneten Knotens, um dessen untergeordnete Knoten anzuzeigen.
- Die Ausrichtung des Baums (von oben nach unten, von links nach rechts, von rechts nach links) können Sie im Menü "Ansicht" oder über die Symbolleiste ändern.
- Zum Anzeigen oder Ausblenden von Feld- und Wertbeschriftungen klicken Sie auf die Schaltfläche "Feld- und Wertbeschriftungen anzeigen" in der Hauptsymbolleiste.
- Die Schaltflächen mit den Vergrößerungsgläsern vergrößern oder verkleinern die Anzeige. Wenn Sie rechts in der Symbolleiste auf die Baumübersichtsschaltfläche klicken, wird ein Diagramm des gesamten Baums angezeigt.
- Wenn ein Partitionsfeld verwendet wird, können Sie die Baumansicht zwischen der Trainings- und der Testpartition umschalten (**Ansicht > Partition**). Wenn die Teststichprobe angezeigt wird, kann der Baum zwar angezeigt, aber nicht bearbeitet werden. (Die aktuelle Partition wird in der unteren linken Fensterecke in der Statusleiste angegeben.)
- Zum Anzeigen von Details über die aktuelle Aufteilung klicken Sie auf die Schaltfläche "Aufteilungsinformationen" (die Schaltfläche "i" ganz rechts in der Symbolleiste). Weitere Informationen finden Sie im Thema „Aufteilungsdetails und Ersatztrenner" auf Seite 85.
- Für jeden Knoten Statistiken, Diagramme oder beides anzeigen (siehe unten).

Anzeigen von Statistiken und Diagrammen

Knotenstatistiken. Bei einem kategorialen Zielfeld zeigt die Tabelle in jedem Knoten die Anzahl und den Prozentsatz der in jeder Kategorie enthaltenen Datensätze sowie den Prozentsatz der gesamten vom Knoten dargestellten Stichprobe. Bei einem stetigen Zielfeld (numerischer Bereich) zeigt die Tabelle die durchschnittliche und die Standardabweichung, die Anzahl der Datensätze und den vorhergesagten Wert des Zielfelds.

Knotendiagramme. Bei einem kategorialen Zielfeld wird das Diagramm als Balkendiagramm der in jeder Kategorie des Zielfelds enthaltenen Prozentsätze ausgegeben. Vor jeder Tabellenzeile befindet sich ein Farbmuster, das die Farbe angibt, die für die Zielfeldkategorie in den Diagrammen des Knotens verwendet wird. Bei einem stetigen Zielfeld (numerischer Bereich) wird das Diagramm als Histogramm des Zielfelds der im Knoten enthaltenen Datensätze angezeigt.

Gewinne

Die Registerkarte "Gewinne" zeigt Statistiken aller im Baum vorhandenen Endknoten. Gewinne bieten ein Maß dafür, wie weit der Mittelwert oder der Anteil eines bestimmten Knotens vom Gesamtmittelwert abweicht. In der Regel gilt, je größer der Unterschied, desto besser kann der Baum als Tool zur Entscheidungsfindung eingesetzt werden. Ein Index- oder "Liftwert" von 148 % für einen Knoten zeigt beispielsweise an, dass die Wahrscheinlichkeit, dass in diesem Knoten enthaltene Datensätze unter die Zielkategorie fallen, fast anderthalb Mal so hoch ist, wie dies für das gesamte Dataset der Fall ist.

Für C&R-Baum- and QUEST-Knoten, für die ein Set zur Verhinderung übermäßiger Anpassung angegeben ist, werden zwei Sets von Statistikdaten angezeigt:

- Baumaufbauset - Trainingsstichprobe ohne Set zur Verhinderung übermäßiger Anpassung
- Set zur Verhinderung übermäßiger Anpassung

Für andere interaktive C&R-Bäume und QUEST-Bäume und für alle interaktiven CHAID-Bäume werden nur die Baumaufbausetstatistiken angezeigt.

Die Registerkarte "Gewinne" bietet die folgenden Möglichkeiten:

- Anzeige knotenweiser, kumulativer oder quantiler Statistiken.
- Anzeige von Gewinnen oder Profiten.
- Umschalten der Anzeige zwischen Tabellen und Diagrammen.
- Auswählen der Zielkategorie (nur für kategoriale Ziele).
- Sortieren der Tabelle in auf- oder absteigender Reihenfolge, basierend auf dem Indexprozentsatz. Wenn Statistiken für mehrere Partitionen angezeigt werden, wird die Sortierung immer in der Trainings- und nicht in der Teststichprobe vorgenommen.

Eine in der Tabelle der Gewinne getroffene Auswahl wird in der Regel in der Baumansicht aktualisiert und umgekehrt. Wenn Sie beispielsweise in der Tabelle eine Zeile auswählen, dann wird im Baum der entsprechende Knoten ausgewählt.

Klassifizierungsgewinne

Bei Klassifizierungsbäumen (mit einer kategorialen Zielvariablen) zeigt der Indexprozentsatz des Gewinns, wie stark der Anteil einer bestimmten Zielkategorie jedes Knotens von dem Gesamtanteil abweicht.

Knotenweise Statistiken

In dieser Ansicht enthält die Tabelle in jeder Zeile einen Endknoten. Beispiel: Wenn Ihre Direktmarketingkampagne insgesamt eine Rücklaufquote von 10 % erzielt hat, dabei aber 20 % der unter Knoten X fallenden Datensätze positiv waren, dann liegt der Indexprozentsatz dieses Knotens bei 200 %. Dieser Wert drückt aus, dass die Wahrscheinlichkeit, dass die in dieser Gruppe enthaltenen Teilnehmer kaufen, doppelt so hoch ist wie in der Gesamtpopulation.

Für CRT-Baum- and QUEST-Knoten, für die ein Set zur Verhinderung übermäßiger Anpassung angegeben ist, werden zwei Sets von Statistikdaten angezeigt:

- Baumaufbauset - Trainingsstichprobe ohne Set zur Verhinderung übermäßiger Anpassung
- Set zur Verhinderung übermäßiger Anpassung

Für andere interaktive CRT-Bäume und QUEST-Bäume und für alle interaktiven CHAID-Bäume werden nur die Baumaufbausetstatistiken angezeigt.

Knoten. Die ID des aktuellen Knotens (die auf der Registerkarte "Viewer" ausgegeben wird).

Knoten: n. Die Gesamtzahl der Datensätze in diesem Knoten.

Knoten (%). Der Prozentsatz aller im Dataset enthaltenen Datensätze, die unter diesen Knoten fallen.

Gewinn: n. Die Anzahl der Datensätze mit der ausgewählten Zielkategorie, die unter diesen Knoten fallen. Dieser Wert gibt an, wie viele der insgesamt im Dataset enthaltenen Datensätze, die unter die Zielkategorie fallen, sich in diesem Knoten befinden.

Gewinn (%). Der Prozentsatz aller in der Zielkategorie enthaltenen Datensätze des gesamten Datasets, die unter diesen Knoten fallen.

Antwort (%). Der Prozentsatz der Datensätze des aktuellen Knotens, die unter die Zielkategorie fallen. Treffer werden in diesem Kontext manchmal auch als "Hits" bezeichnet.

Index (%). Der Trefferprozentsatz des aktuellen Knotens, ausgedrückt als Prozentsatz des Trefferprozentsatzes des gesamten Datasets. Beispiel: Ein Indexwert von 300 % zeigt an, dass die Wahrscheinlichkeit, dass in diesem Knoten enthaltene Datensätze unter die Zielkategorie fallen, dreimal so hoch ist, wie dies im gesamten Dataset der Fall ist.

Kumulative Statistiken

In der kumulativen Ansicht enthält die Tabelle einen Knoten pro Zeile, wobei die Statistiken aber kumulativ und auf- oder absteigend nach Indexprozentsatz sortiert sind. Wenn beispielsweise eine absteigende Sortierung vorliegt, wird in der ersten Zeile der Knoten mit dem höchsten Indexprozentsatz ausgegeben und in den darauffolgenden Zeilen erscheinen die kumulierten Werte der jeweiligen Zeile mit den darüberliegenden Zeilen.

Der kumulierte Indexprozentsatz sinkt mit jeder Zeile, da Knoten mit immer niedrigeren Trefferprozenten hinzugefügt werden. Der kumulative Index der letzten Zeile liegt immer bei 100 %, da an diesem Punkt das gesamte Dataset enthalten ist.

Quantile

In dieser Ansicht wird in jeder Zeile anstelle eines Knotens ein Quantil angezeigt. Bei den Quantilen handelt es sich entweder um Quartile, Quintile (Fünftel), Dezile (Zehntel), Vingtile (Zwanzigstel) oder Perzentile (Hundertstel). Sofern mehrere Knoten benötigt werden, um den Prozentsatz zu erreichen, können in einem Quantil mehrere Knoten aufgeführt sein (wenn z. B. Quartile angezeigt werden, die beiden höchsten Knoten jedoch weniger als 50 % aller Fälle enthalten). Die übrige Tabelle ist kumulativ und genau wie die kumulative Ansicht zu interpretieren.

Klassifizierung der Profite und ROI

Die Gewinnstatistiken für Klassifizierungsbäume können auch mit Profit und ROI (Return on Investment) ausgegeben werden. Mit dem Dialogfeld "Profite definieren" können Sie die Einnahmen und die Ausgaben für die einzelnen Kategorien angeben.

1. Öffnen Sie die Registerkarte "Gewinne" (mit dem Symbol $/$) und klicken Sie in der Symbolleiste auf die Schaltfläche "Profit", um das Dialogfeld zu öffnen.
2. Geben Sie für jede Kategorie des Zielfelds Einnahmen- und Ausgabenwerte ein.

Wenn es Sie beispielsweise 0,48 $ kostet, ein Angebot an jeden Kunden zu senden, und die Einnahme bei einer positiven Antwort für ein dreimonatiges Abonnement 9,95 $ beträgt, kostet Sie jede *Nein*-Antwort 0,48 $ und jede *Ja*-Antwort bringt Ihnen 9,47 $ (9,95-0,48).

In der Tabelle "Gewinne" wird der **Profit** als die Summe der Einnahmen abzüglich der Ausgaben für alle im Endknoten enthaltenen Datensätze berechnet. **ROI** ist der Gesamtprofit geteilt durch die Gesamtausgaben in einem Knoten.

Kommentare

- Profitwerte wirken sich nur auf die in der Tabelle "Gewinne" angezeigten durchschnittlichen Profit- und ROI-Werte aus und bieten eine für Ihr Endergebnis brauchbarere statistische Anzeige. Die grundlegende Baummodellstruktur bleibt unverändert. Profite dürfen nicht mit Fehlklassifizierungskosten verwechselt werden, die im Baumerstellungsknoten angegeben sind und zum Schutz gegen teuere Fehler als Faktor in das Modell eingehen.
- Profitangaben sind zwischen interaktiven Baumerstellungssitzungen nicht persistent.

Regressionsgewinne

Für Regressionsbäume kann eine knotenweise, kumulativ knotenweise und eine quantile Ansicht gewählt werden. Durchschnittswerte werden in der Tabelle ausgegeben. Diagramme sind nur für Quantile verfügbar.

Gewinndiagramme

Alternativ zu Tabellen können auf der Registerkarte "Gewinne" Diagramme angezeigt werden.

1. Wählen Sie auf der Registerkarte "Gewinne" das Symbol für Quantile (das dritte von links auf der Symbolleiste). (Für knotenweise oder kumulative Statistiken sind keine Diagramme verfügbar.)
2. Klicken Sie auf das Diagrammsymbol.

3. Wählen Sie in der Dropdown-Liste die Einheit aus (Perzentile, Dezile etc.), die angezeigt werden soll.
4. Wählen Sie **Gewinne**, **Treffer** oder **Lift**, um die angezeigte Messung zu ändern.

Gewinndiagramm

Das Gewinndiagramm bildet die Werte der Tabellenspalte *Gewinn (%)* ab. Gewinne sind als der Anteil der in jedem Inkrement enthaltenen Treffer im Verhältnis zur Gesamtzahl der im Baum enthaltenen Treffer definiert. Dabei kommt folgende Gleichung zum Einsatz:

(Treffer im Inkrement / Gesamtzahl Treffer) x 100 %

Das Diagramm illustriert, wie weit Sie das Netz auswerfen müssen, um einen bestimmten Prozentsatz aller im Baum enthaltenen Treffer zu erzielen. Die diagonale Linie bildet die für die gesamte Stichprobe erwarteten Treffer ab, wenn das Modell nicht verwendet wird. In diesem Fall ist die Trefferrate konstant, da die Wahrscheinlichkeit eines Treffers für alle Personen gleich ist. Um das Ergebnis zu verdoppeln, müssen Sie doppelt so viele Personen ansprechen. Die gekrümmte Linie zeigt an, wie weit Sie Ihre Treffer verbessern können, wenn Sie nur die einschließen, deren Prozentsatz hinsichtlich des Gewinns höher ausfällt. Wenn Sie beispielsweise die obersten 50 % einschließen, erhalten Sie über 70 % der positiven Treffer. Je steiler die Kurve, desto höher ist der Gewinn.

Liftdiagramm

Das Liftdiagramm bildet die Werte der Tabellenspalte *Index (%)* ab. Dieses Diagramm vergleicht den Prozentsatz der in jedem Inkrement enthaltenen Datensätze, bei denen es sich um Treffer handelt, mit dem Prozentsatz aller im Trainingsdataset enthaltenen Treffer. Dabei wird folgende Gleichung zugrunde gelegt:

(Treffer im Inkrement / Datensätze im Inkrement) / (Gesamtzahl Treffer / Gesamtzahl Datensätze)

Trefferdiagramm

Das Trefferdiagramm bildet die Werte der Tabellenspalte *Treffer (%)* ab. Ergebnis ist der Prozentsatz der im Inkrement enthaltenen Datensätze, bei denen es sich um Treffer handelt, wobei folgende Gleichung angewendet wird:

(Treffer im Inkrement / Datensätze im Inkrement) x 100 %

Gewinnbasierte Auswahl

Mit dem Dialogfeld "Gewinnbasierte Auswahl" können Sie automatisch Endknoten mit den höchsten (oder niedrigsten) Gewinnen basierend auf einer angegebenen Regel oder einem Schwellenwert auswählen. Auf der Grundlage dieser Auswahl können Sie dann einen Auswahlknoten generieren.

1. Wählen Sie auf der Registerkarte "Gewinne" die knotenweise oder die kumulative Anzeige sowie die Zielkategorie aus, die die Grundlage für die Auswahl bilden soll. (Eine Auswahl basiert auf der aktuellen Tabellenanzeige und ist für Quantile nicht verfügbar.)
2. Wählen Sie auf der Registerkarte "Gewinne" folgende Menüoptionen:

   Bearbeiten > **Endknoten auswählen** > **Gewinnbasierte Auswahl**

   Nur auswählen. Sie können zutreffende *oder* nicht zutreffende Knoten auswählen, um beispielsweise *alle bis auf* die 100 höchsten Datensätzen auszuwählen.

   Nach Gewinninformationen zuordnen. Ordnet Knoten auf der Grundlage der Gewinnstatistiken der aktuellen Zielkategorie zu. Dazu gehören:

   - Knoten, deren Gewinn, Antwort oder Lift (Index) mit einem angegebenen Schwellenwert übereinstimmt, beispielsweise einer Antwort größer oder gleich 50 %.
   - Die *n* höchsten Knoten, basierend auf dem Gewinn für die Zielkategorie.
   - Die höchsten Knoten, bis zu einer vorgegebenen Anzahl Knoten.
   - Die höchsten Knoten, bis zu einem vorgegebenen Prozentsatz der Trainings-Daten.

3. Klicken Sie auf **OK**, um die Auswahl auf der Registerkarte "Viewer" zu aktualisieren.
4. Um auf der Grundlage der aktuellen Auswahl auf der Registerkarte "Viewer" einen neuen Auswahlknoten zu erstellen, wählen Sie im Menü "Generieren" die Option **Auswahlknoten**. Weitere Informationen finden Sie im Thema „Generieren von Filter- und Auswahlknoten" auf Seite 93.

Hinweis: Da Sie zur Zeit Knoten und keine Datensätze oder Prozentsätze auswählen, kann nicht immer eine perfekte Übereinstimmung mit dem Auswahlkriterium erzielt werden. Das System wählt komplette Knoten *bis zu* der vorgegebenen Ebene aus. Wenn Sie beispielsweise die höchsten 12 Fälle auswählen, von denen 10 im ersten und 2 im zweiten Knoten liegen, dann wird nur der erste Knoten ausgewählt.

Risiken

Risiken geben an, wie groß die Gefahr einer Fehlklassifizierung auf jeder Ebene ist. Die Registerkarte "Risiken" enthält eine punktuelle Risikoschätzung und (für kategoriale Ausgaben) eine Fehlklassifizierungstabelle.

- Bei numerischen Vorhersagen bildet das Risiko eine Gesamtschätzung der in jedem Endknoten vorhandenen Varianz.
- Bei kategorialen Vorhersagen bildet das Risiko den Anteil der falsch klassifizierten Fälle, angepasst um alle A-priori- oder Fehlklassifizierungskosten.

Speichern der Baummodelle und Ergebnisse

Die Ergebnisse Ihrer interaktiven Baumerstellung können Sie auf mehrere Arten speichern oder exportieren:

- Generieren eines Modells auf der Grundlage des aktuellen Baums (**Generieren > Modell erzeugen**).
- Speichern der für die Erweiterung des aktuellen Baums verwendeten Richtlinien. Wenn Sie den Baumerstellungsknoten das nächste Mal ausführen, wird der aktuelle Baum automatisch wieder mit allen von Ihnen festgelegten benutzerdefinierten Aufteilungen aufgebaut.
- Exportieren der Modell-, Gewinn- und Risikoinformationen. Weitere Informationen finden Sie im Thema „Exportieren der Modell-, Gewinn- und Risikoinformationen" auf Seite 93.

Über den Tree Builder oder ein Baummodellnugget können Sie Folgendees ausführen:

- Generieren Sie einen Filter oder wählen Sie basierend auf dem aktuellen Baum einen Knoten aus. Weitere Informationen finden Sie im Thema „Generieren von Filter- und Auswahlknoten" auf Seite 93.
- Generieren Sie ein Regelsetnugget, das die Baumstruktur als Set von Regeln darstellt, das die Endverzweigungen des Baums definiert. Weitere Informationen finden Sie im Thema „Generieren eines Regelsets aus einem Entscheidungsbaum" auf Seite 93.
- Bei Baummodellnuggets können Sie außerdem das Modell im PMML-Format exportieren. Weitere Informationen finden Sie im Thema „Modellpalette" auf Seite 40. Wenn das Modell benutzerdefinierte Aufteilungen beinhaltet, werden diese Informationen nicht in der exportierten PMML gespeichert. (Die Aufteilung wird beibehalten, die Tatsache, dass sie benutzerdefiniert und nicht vom Algorithmus gewählt ist, jedoch nicht.)
- Generieren Sie ein Diagramm auf der Basis des ausgewählten Teils des aktuellen Baums. *Hinweis*: Dies funktioniert bei einem Nugget nur, wenn es an andere Knoten in einem Stream angehängt ist. Weitere Informationen finden Sie im Thema „Erzeugen von Diagrammen" auf Seite 114.

Hinweis: Der interaktive Baum als solcher kann nicht gespeichert werden. Damit Ihre Arbeit nicht verloren geht, müssen Sie ein Modell generieren und/oder die Direktiven aktualisieren, bevor Sie das Tree Builder-Fenster schließen.

Generieren eines Modells mit Tree Builder

Um ein Modell auf der Grundlage des aktuellen Baums zu generieren, wählen Sie in den Tree Builder-Menüs folgende Optionen:

Generieren > **Modell**

Im Dialogfeld "Neues Modell generieren" können Sie aus den folgenden Optionen auswählen:

Modellname. Sie können einen benutzerdefinierten Namen angeben oder den Namen auf der Grundlage des Namens des Modellierungsknotens automatisch generieren lassen.

Knoten erstellen auf. Sie können den Knoten zum **Erstellungsbereich**, zur Palette **Generierte Modelle** oder zu **Beiden** hinzufügen.

Direktiven einbeziehen. Zum Einschließen der Direktiven vom aktuellen Baum im generierten Modell, markieren Sie dieses Kontrollkästchen. Auf diese Weise können Sie den Baum bei Bedarf erneut generieren. Weitere Informationen finden Sie im Thema „Direktiven für Bäume".

Direktiven für Bäume

Bei Modellen vom Typ "C&R-Baum", "CHAID" und "QUEST" werden durch Direktiven die Bedingungen festgelegt, nach denen der Baum um jeweils eine Ebene erweitert wird. Direktiven werden jedes Mal angewendet, wenn der interaktive Tree Builder vom Knoten aus gestartet wird.

- Direktiven werden meist verwendet, um einen zuvor interaktiv erstellten Baum noch einmal zu generieren. Weitere Informationen finden Sie im Thema „Aktualisieren der Direktiven" auf Seite 92. Sie können Direktiven auch manuell bearbeiten. Dabei sollten Sie allerdings vorsichtig sein.
- Die Direktiven sind sehr spezifisch für die Struktur des Baums, den sie beschreiben. Daher kann jede Änderung der zugrunde liegenden Daten oder der Modellierungsoptionen dazu führen, dass ein bislang gültiger Satz an Direktiven Fehler verursacht. Wenn beispielsweise der CHAID-Algorithmus eine zweifache Aufteilung auf der Grundlage aktualisierter Daten in eine dreifache Aufteilung ändert, führen alle zuvor auf der Grundlage der zweifachen Aufteilung erstellten Richtlinien zu Fehlern.

Hinweis: Wenn Sie ein Modell direkt erstellen (ohne Tree Builder), werden sämtliche Baumdirektiven ignoriert.

Bearbeiten von Direktiven

1. Um gespeicherte Direktiven anzuzeigen oder zu bearbeiten, öffnen Sie den Baumerstellungsknoten und wählen Sie das Panel "Ziel" auf der Registerkarte "Erstellungsoptionen".
2. Wählen Sie **Interaktive Sitzung starten**, um die Steuerungen zu aktivieren, wählen Sie **Interaktiv erstellte Direktiven verwenden** und klicken Sie auf **Direktiven**.

Syntax der Direktiven

Direktiven legen die Bedingungen für die Erweiterung des Baums fest, beginnend mit dem Stammknoten. Um den Baum beispielsweise um eine Ebene zu erweitern:

```
Grow Node Index 0 Children 1 2
```

Da kein Prädiktor angegeben ist, wählt der Algorithmus die beste Aufteilung aus.

Beachten Sie, dass die erste Aufteilung immer im Stammknoten (`Index 0`) erfolgen muss und dass die Indexwerte für beide untergeordneten Elemente angegeben werden müssen (in diesem Fall `1` und `2`). Die Angabe `Grow Node Index 2 Children 3 4` ist erst dann gültig, wenn Sie den Stammknoten erweitert und damit den Knoten 2 erzeugt haben.

So erweitern Sie den Baum:

```
Grow Tree
```

So erweitern und reduzieren Sie den Baum (nur C&R-Baum):

```
Grow_And_Prune Tree
```

So legen Sie eine benutzerdefinierte Aufteilung für einen stetigen Prädiktor fest:

```
Grow Node Index 0 Children 1 2 Spliton
  ( "EDUCATE", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

So nehmen Sie eine Aufteilung eines nominalen Prädiktors mit zwei Werten vor:

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

So nehmen Sie eine Aufteilung eines nominalen Prädiktors mit mehreren Werten vor:

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ORGS", Group( "2.0","4.0" )
             Group( "0.0","1.0","3.0","6.0" ))
```

So nehmen Sie eine Aufteilung eines ordinalen Prädiktors vor:

```
Grow Node Index 4 Children 5 6 Spliton
  ( "CHILDS", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

Hinweis: Beim Anlegen von benutzerdefinierten Aufteilungen, muss bei Feldnamen und Werten (`EDUCATE`, `GENDER`, `CHILDS` usw.) die Groß-/Kleinschreibung beachtet werden.

Direktiven für CHAID-Bäume

Direktiven für CHAID-Bäume reagieren besonders empfindlich auf Änderungen der Daten oder des Modells, da sie, anders als "C&R-Baum" und "QUEST", nicht auf die Verwendung binärer Aufteilungen beschränkt sind. Die folgende Syntax sieht beispielsweise völlig korrekt aus, führt aber zu einem Fehler, wenn der Algorithmus den Stammknoten in mehr als zwei untergeordnete Elemente aufteilt:

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

Bei CHAID ist es möglich, dass der Knoten 0 auch 3 oder 4 untergeordnete Elemente besitzt, was dazu führt, dass die zweite Zeile einen Fehler verursacht.

Verwenden von Direktiven in Scripts

Direktiven können auch in Scripts eingebettet werden, indem sie in dreifache Anführungszeichen eingeschlossen werden.

Aktualisieren der Direktiven

Um Ihre Ergebnisse einer interaktiven Baumerstellung zu erhalten, können Sie die für das Generieren des aktuellen Baums aufgestellten Direktiven speichern. So können Sie den Baum in seinem aktuellen Status zur weiteren Bearbeitung neu erstellen, was bei einem gespeicherten Modellnugget nicht mehr möglich ist.

Um Richtlinien zu aktualisieren, wählen Sie in den Tree Builder-Menüs folgende Optionen:

Datei > **Direktiven aktualisieren**

Direktiven werden in dem Modellierungsknoten gespeichert, mit dem der Baum erstellt wurde ("C&R-Baum", "QUEST" oder "CHAID") und können dazu genutzt werden, den aktuellen Baum noch einmal zu generieren. Weitere Informationen finden Sie im Thema „Direktiven für Bäume” auf Seite 91.

Exportieren der Modell-, Gewinn- und Risikoinformationen

Über Tree Builder können Sie Modell-, Gewinn- und Risikostatistiken als Text-, HTML- oder Bildformate exportieren.

1. Wählen Sie im Tree Builder-Fenster die Registerkarte oder die Ansicht, die Sie exportieren wollen.
2. Wählen Sie die folgenden Befehle aus den Menüs aus:
   Datei > **Exportieren**
3. Wählen Sie **Text**, **HTML** oder **Diagramm** sowie die Elemente aus, die Sie aus dem Untermenü exportieren wollen.

Soweit zutreffend, wird der Export auf der Grundlage der aktuellen Auswahl durchgeführt.

Exportieren von Text- oder HTML-Formaten. Sie können Gewinn- oder Risikostatistiken für die Trainings- oder die Testpartition (sofern definiert) exportieren. Der Export erfolgt auf der Grundlage der aktuellen Auswahl auf der Registerkarte "Gewinne". Sie können beispielsweise "knotenweise", "kumulativ" oder "quantile Statistiken" auswählen.

Exportieren von Grafiken. Sie können den aktuellen Baum so exportieren, wie er auf der Registerkarte "Viewer" angezeigt wird. Oder Sie exportieren die Gewinndiagramme für die Trainings- oder Testpartition (sofern definiert). Zu den verfügbaren Formaten gehören *.JPEG*, *.PNG* und *.BMP*. Bei Gewinnen basiert der Export auf der aktuell auf der Registerkarte "Gewinne" (nur verfügbar, wenn ein Diagramm angezeigt wird) vorgenommenen Auswahl.

Generieren von Filter- und Auswahlknoten

Im Tree Builder-Fenster oder beim Durchsuchen eines Modellnuggets für ein Entscheidungsbaummodell wählen Sie in den Menüs folgende Optionen aus:

Generieren > **Filterknoten**

oder

> **Auswahlknoten**

Filterknoten. Generiert einen Knoten, der alle vom aktuellen Baum nicht verwendeten Felder herausfiltert. Mit diesem Verfahren schränken Sie das Dataset schnell ein, sodass es nur noch solche Felder enthält, die vom Algorithmus als wichtige Felder ausgewählt werden. Wenn oberhalb dieses Entscheidungsbaumknotens ein Typknoten liegt, leitet das Modellnugget des Filtermodells alle Felder mit der Rolle *Ziel* weiter.

Auswahlknoten. Generiert einen Knoten, der alle Datensätze auswählt, die in den aktuellen Knoten fallen. Diese Option setzt voraus, dass auf der Registerkarte "Viewer" eine oder mehrere Verzweigungen ausgewählt sind.

Das Modellnugget wird im Streamerstellungsbereich abgelegt.

Generieren eines Regelsets aus einem Entscheidungsbaum

Sie können ein Modellnugget vom Typ "Regelset" generieren, das die Baumstruktur als Menge von Regeln darstellt, mit denen die Endzweigungen des Baums definiert werden. Regelsets enthalten meist die wichtigsten Informationen eines gesamten Entscheidungsbaums, allerdings mit einem weniger komplexen Modell. Der wichtigste Unterschied besteht darin, dass es bei einem Regelset möglich ist, dass für einen bestimmten Datensatz mehr als eine oder aber überhaupt keine Regel gilt. Sie können beispielsweise alle Regeln nehmen, die als Ergebnis *Nein* vorhersagen, und dann alle, die *Ja* vorhersagen. Wenn mehrere Regeln gelten, dann wird jeder Regel ein gewichtetes "Votum" zugeordnet, das auf der dieser Regel zuge-

ordneten Konfidenz basiert, und die endgültige Vorhersage ergibt sich aus der Kombination der gewichteten Voten aller für den fraglichen Datensatz geltenden Regeln. Wenn keine Regel gilt, wird dem Datensatz eine Standardvorhersage zugeordnet.

Regelsets können nur aus Bäumen mit kategorialen Zielfeldern generiert werden (nicht aus Regressionsbäumen).

Im Tree Builder-Fenster oder beim Durchsuchen eines Modellnuggets für ein Entscheidungsbaummodell wählen Sie in den Menüs folgende Optionen aus:

Generieren > **Regelset**

Regelsatzname. Mit dieser Option können Sie den Namen des neuen Modellnuggets vom Typ "Regelset" angeben.

Knoten erstellen auf. Steuert den Standort des neuen Modellnuggets vom Typ "Regelset". Wählen Sie **Erstellungsbereich**, **Generierte Modelle** oder **Beides** aus.

Mindestzahl Instanzen. Legen Sie fest, wie viele Instanzen (Anzahl der Datensätze, für die die Regel gilt) im Modellnugget vom Typ "Regelset" mindestens beibehalten werden sollen. Regeln, deren Unterstützung unter dem angegebenen Wert liegt, werden nicht in das neue Regelset aufgenommen.

Minimale Konfidenz. Geben Sie die minimale Konfidenz für Regeln an, die im Modellnugget vom Typ "Regelset" erhalten bleiben sollen. Regeln, deren Konfidenz unter dem angegebenen Wert liegt, werden nicht in das neue Regelset aufgenommen.

Direktes Erstellen eines Baummodells

Alternativ zum interaktiven Tree Builder können Sie ein Entscheidungsbaummodell auch direkt aus dem Knoten erstellen, sobald der Stream ausgeführt wird. Dies ist konsistent mit anderen Modellerstellungsknoten. Für C5.0-Baummodelle, die nicht vom interaktiven Tree Builder unterstützt werden, kann ausschließlich diese Methode genutzt werden.

1. Erstellen Sie einen Stream und fügen Sie einen der Entscheidungsbaumknoten, C&R-Baum, CHAID, QUEST oder C5.0, hinzu.
2. Wählen Sie für "C&R-Baum", "QUEST" oder "CHAID" im Fenster "Ziel" der Registerkarte "Erstellungsoptionen" eines der Hauptziele aus. Wenn Sie "Einzelnen Baum aufbauen" wählen, stellen Sie sicher, dass der Modus auf **Modell erzeugen** gesetzt ist.

   Für C5.0 setzen Sie auf der Registerkarte "Modell" die Option **Ausgabetyp** auf **Entscheidungsbaum**.
3. Wählen Sie die Ziel- und Prädiktorfelder aus und legen Sie die zusätzlich benötigten Modelloptionen fest. Spezifische Anleitungen finden Sie in der Dokumentation zu den einzelnen Baumerstellungsknoten.
4. Führen Sie den Stream aus, damit das Modell generiert wird.

Kommentare

- Wenn Sie Bäume nach dieser Methode generieren, werden die Direktiven für Bäume ignoriert.
- Egal ob interaktiv oder direkt, beide Methoden zur Erstellung von Entscheidungsbäumen führen zu ähnlichen Modellen. Die Frage ist lediglich, wie viel Kontrolle Sie während des Ablaufs ausüben wollen.

Entscheidungsbaumknoten

Die Entscheidungsbaumknoten in IBM SPSS Modeler gewähren Zugriff auf die früher eingeführten Baumerstellungsalgorithmen:

- C&R-Baum
- QUEST
- CHAID
- C5.0

Weitere Informationen finden Sie im Thema „Entscheidungsbaummodelle" auf Seite 81.

Die Algorithmen gleichen sich dahingehend, dass alle einen Entscheidungsbaum aufbauen können, indem sie rekursiv die Daten in immer kleinere Untergruppen aufteilen. Es gibt jedoch einige entscheidende Unterschiede.

Eingabefelder. Die folgenden Typen (Messniveaus) sind für die Eingabefelder (Prädiktoren) möglich: stetig, kategorial, Flag, nominal oder ordinal.

Zielfelder. Es kann nur ein Zielfeld angegeben werden. Bei C&R-Baum und CHAID sind folgende Typen für das Ziel möglich: stetig, kategorial, Flag, nominal oder ordinal. Bei QUEST kann es kategorial, ein Flag oder nominal sein. Bei C5.0 kann das Ziel ein Flag, nominal oder ordinal sein.

Aufteilungstyp. C&R-Baum und QUEST unterstützen nur binäre Aufteilungen (d. h., jeder Knoten des Baums kann nur in zwei Verzweigungen aufgeteilt werden). Dagegen unterstützen CHAID und C5.0 die Aufteilung in mehr als zwei Verzweigungen gleichzeitig.

Für die Aufteilung verwendete Methode. Die Algorithmen unterscheiden sich in den Kriterien, die für die Aufteilungsentscheidung verwendet werden. Wenn ein C&R-Baum eine kategoriale Ausgabe vorhersagt, wird ein Streuungsmaß verwendet (standardmäßig der Gini-Koeffizient, Sie können dies jedoch ändern). Für stetige Ziele wird die Methode der kleinsten quadratischen Abweichung verwendet. CHAID verwendet den Chi-Quadrat-Test; bei QUEST wird ein Chi-Quadrat-Test für kategoriale Prädiktoren verwendet und die Varianzanalyse bei stetigen Eingaben. Bei C5.0 wird ein Informationstheorie-Maß verwendet, das Informationsgewinnverhältnis.

Behandlung fehlender Werte. Alle Algorithmen lassen fehlende Werte in den Prädiktorfeldern zu, wenden jedoch unterschiedliche Methoden für deren Behandlung an. C&R-Baum und QUEST verwenden bei Bedarf Ersatzvorhersagefelder, um einen Datensatz mit fehlenden Werten beim Training weiter durch den Baum zu leiten. CHAID erstellt für die fehlenden Werte eine separate Kategorie und lässt sie zur Verwendung bei der Baumerstellung zu. Bei C5.0 wird eine Fraktionierungsmethode eingesetzt, die einen Bruchbereich eines Datensatzes entlang jeder Verzweigung des Baums von einem Knoten weiterreicht, bei dem die Aufteilung auf einem Feld mit fehlendem Wert basiert.

Beschneiden. Bei C&R-Baum, QUEST und C5.0 gibt es die Option, den Baum vollständig aufzubauen und anschließend durch Entfernen von Aufteilungen der untersten Ebene zu beschneiden, die keine signifikante Auswirkung auf die Genauigkeit des Baums haben. Bei allen Entscheidungsbaumalgorithmen ist es jedoch möglich, die Mindestgröße der Untergruppen zu steuern, was dazu beiträgt, Verzweigungen mit wenigen Datensätzen zu vermeiden.

Interaktive Baumerstellung. C&R-Baum, QUEST und CHAID bieten eine Option zum Starten einer interaktiven Sitzung. Auf diese Weise können Sie Ihren Baum jeweils auf einer Ebene erstellen, die Aufteilungen bearbeiten und den Baum beschneiden, bevor Sie das Modell erstellen. C5.0 bietet keine Option zur interaktiven Bearbeitung.

A-priori-Wahrscheinlichkeiten. C&R-Baum und QUEST unterstützen die Spezifikation von A-priori-Wahrscheinlichkeiten für Kategorien bei der Vorhersage eines kategorialen Zielfelds. A-priori-Wahrschein-

lichkeiten sind Schätzungen der gesamten relativen Häufigkeit für jede Zielkategorie in der Gesamtheit, aus der die Trainingsdaten gezogen werden. Mit anderen Worten: Es handelt sich um die Wahrscheinlichkeitsschätzungen, die Sie für jeden möglichen Zielwert vornehmen würden, bevor Sie etwas über die Prädiktorwerte wissen. CHAID und C5.0 unterstützen die Spezifizierung von A-priori-Wahrscheinlichkeiten nicht.

Regelsets. Für Modelle mit kategorialen Zielfeldern bieten die Entscheidungsbaumknoten die Option, das Modell in Form eines Regelsets zu erstellen, was bei der Interpretation teilweise einfacher ist als ein komplexer Entscheidungsbaum. Bei C&R-Baum, QUEST und CHAID können Sie ein Regelset in einer interaktiven Sitzung erstellen. Bei C5.0 können Sie diese Option auf dem Modellierungsknoten angeben. Außerdem ist es bei allen Entscheidungsbaummodellen möglich, ein Regelset im Modellnugget zu erstellen. Weitere Informationen finden Sie im Thema „Generieren eines Regelsets aus einem Entscheidungsbaum" auf Seite 93.

C&R-Baumknoten

Der Klassifizierungs- und Regressionsbaumknoten (C&R - Classification & Regression) ist eine baumbasierte Klassifizierungs- und Vorhersagemethode. Ähnlich wie C5.0 verwendet diese Methode eine rekursive Partitionierung, um die Trainingsdatensätze in Segmente mit ähnlichen Ausgabefeldwerten aufzuteilen. Der Knoten vom Typ "C&R-Baum" beginnt mit der Untersuchung der Eingabefelder, um die beste Aufteilung zu finden, die anhand der Reduktion in einem aus der Aufteilung resultierenden Unreinheitsindex gemessen wird. Die Aufteilung definiert zwei Untergruppen, die anschließend beide in zwei weitere Untergruppen aufgeteilt werden. Dies wird so lange fortgesetzt, bis die Stoppkriterien erreicht sind. Alle Aufteilungen erfolgen binär (nur zwei Untergruppen).

Reduzierung

Bei C&R-Bäumen haben Sie die Option, den Baum zuerst zu erweitern und dann auf der Grundlage eines Kostenkomplexitätsalgorithmus, der die Risikoschätzung basierend auf der Anzahl der Endknoten anpasst, zu reduzieren. Diese Methode, die eine große Erweiterung des Baumes ermöglicht, bevor dieser nach komplexeren Kriterien reduziert wird, kann zu kleineren Bäumen mit besseren Kreuzvalidierungseigenschaften führen. Wenn die Anzahl der Endknoten vergrößert wird, verringert dies in der Regel das Risiko für die aktuellen (Trainings-)Daten. Das tatsächliche Risiko kann aber größer sein, wenn das Modell auf unbekannte Daten verallgemeinert wird. Angenommen, es liegt der Extremfall vor, dass Sie für jeden im Trainingsdataset vorhandenen Datensatz einen separaten Endknoten besitzen. Das geschätzte Risiko liegt bei 0 %, da jeder Datensatz in einen eigenen Knoten fällt, das Fehlklassifizierungsrisiko für unbekannte (Test-)Daten liegt allerdings mit großer Sicherheit über 0 %. Die Kostenkomplexitätsmessung versucht, dies zu kompensieren.

Beispiel. Ein Kabelfernsehunternehmen hat eine Marketingstudie in Auftrag gegeben, um zu ermitteln, welche Kunden ein Abonnement für einen interaktiven Nachrichtenservice über Kabel erwerben würden. Mithilfe der Daten aus der Studie können Sie einen Stream erstellen, in dem das Zielfeld die Absicht angibt, das Abonnement zu erwerben, und in dem als Prädiktorfelder Alter, Geschlecht, Bildung, Einkommenskategorie, wöchentlicher Fernsehkonsum und Anzahl der Kinder verwendet werden. Wenn Sie einen Knoten vom Typ "C&R-Baum" auf den Stream anwenden, können Sie die Antworten vorhersagen und klassifizieren, um die höchste Rücklaufquote für Ihre Kampagne zu erzielen.

Anforderungen. Um ein C&R-Baummodell zu trainieren, benötigen Sie mindestens ein *Eingabe*-Feld und genau ein *Ziel*-Feld. Ziel- und Eingabefelder können stetig (in einem numerischen Bereich) oder kategorial sein. Felder, die auf *Beides* oder *Keine* gesetzt sind, werden ignoriert. Die Typen der im Modell verwendeten Felder müssen vollständig als Instanz generiert sein und alle im Modell verwendeten Ordinalfelder (sortiertes Set) müssen numerisch (und nicht als Zeichenfolge) gespeichert sein. Im Bedarfsfall können Sie die Felder mit dem Umcodierungsknoten konvertieren.

Stärken. C&R-Baummodelle sind bei Problemen mit fehlenden Daten und einer großen Feldanzahl sehr stabil. Sie benötigen für die Schätzung in der Regel keine langen Trainingsphasen. Darüber hinaus sind

C&R-Baummodelle tendenziell leichter zu verstehen als einige andere Modelltypen. Die aus dem Modell abgeleiteten Regeln lassen sich sehr direkt interpretieren. Im Gegensatz zu C5.0 können C&R-Bäume stetige genauso wie kategoriale Ausgabefelder verarbeiten.

CHAID-Knoten

CHAID (Chi-squared Automatic Interaction Detection) ist eine Klassifizierungsmethode für die Erstellung von Entscheidungsbäumen mit Chi-Quadrat-Statistiken zur Identifizierung der optimalen Splits.

CHAID untersucht zuerst die zwischen allen Eingabefeldern und dem Ergebnis vorhandenen Kreuztabellen und testet die Signifikanz mit einem Chi-Quadrat-Unabhängigkeitstest. Wenn mehr als eine dieser Beziehungen statistisch signifikant ist, wählt CHAID das signifikanteste Eingabefeld aus (kleinster *P*-Wert). Wenn eine Eingabe mehr als zwei Kategorien besitzt, werden diese verglichen und solche Kategorien gegeneinander reduziert, deren Ergebnis keinen Unterschied aufweist. Dies erfolgt, indem sukzessive alle Kategorienpaare mit dem am wenigsten signifikanten Unterschied verbunden werden. Diese Kategoriezusammenführung wird gestoppt, wenn die Abweichung aller verbleibenden Kategorien das angegebene Testniveau erreicht hat. Bei nominalen Eingabefeldern können alle Kategorien zusammengeführt werden. Bei einem ordinalen Set können nur zusammenhängende Kategorien zusammengeführt werden.

Exhaustive CHAID ist eine Änderung von CHAID, die noch gründlicher vorgeht, indem sie alle für jeden Prädiktor möglichen Aufteilungen untersucht, allerdings mehr Rechenzeit beansprucht.

Anforderungen. Ziel- und Eingabefelder können stetig oder kategorial sein. Knoten können auf jeder Ebene in zwei oder mehr Untergruppen aufgeteilt werden. Alle im Modell verwendeten ordinalen Felder müssen numerisch (nicht als Zeichenfolge) gespeichert sein. Im Bedarfsfall können Sie die Felder mit dem Umcodierungsknoten konvertieren.

Stärken. Im Gegensatz zu den Knoten vom Typ "C&R-Baum" und "QUEST" kann CHAID nicht binäre Bäume generieren, d. h. Bäume mit Aufteilungen mit mehr als zwei Verzweigungen. CHAID erstellt daher tendenziell breitere Bäume als die binären Aufbaumethoden. CHAID funktioniert mit allen Eingaben und akzeptiert sowohl Fallgewichtungs- als auch Häufigkeitsvariablen.

QUEST-Knoten

QUEST (Quick, Unbiased, Efficient Statistical Tree - Schneller, unverzerrter, effizienter Statistikbaum) ist eine binäre Klassifizierungsmethode zum Erstellen eines Entscheidungsbaums. Diese Methode wurde primär in der Absicht entwickelt, die Verarbeitungszeit zu verkürzen, die für Analysen von großen C&R-Bäumen mit vielen Variablen oder mit vielen Fällen benötigt wird. Ein zweites Ziel von QUEST war die Senkung der in den Klassifizierungsbaummodellen festgestellten Tendenz, Eingaben zu bevorzugen, die mehr Aufteilungen erlauben. Dabei handelt es sich um stetige Eingabefelder (numerischer Bereich) oder um solche mit vielen Kategorien.

- QUEST verwendet eine Folge von auf signifikanten Tests basierenden Regeln, um die im Knoten vorhandenen Eingabefelder zu bewerten. Zu Auswahlzwecken muss gegebenenfalls für jede in einem Knoten vorhandene Eingabe nur ein einziger Test durchgeführt werden. Im Gegensatz zu "C&R-Baum" werden nicht alle Aufteilungen untersucht. Und im Gegensatz zu "C&R-Baum" und "CHAID" werden beim Bewerten eines Eingabefelds für die Auswahl die Kategoriekombinationen nicht getestet. Dies beschleunigt die Analyse.
- Aufteilungen werden festgelegt, indem eine quadratische Diskriminanzanalyse durchgeführt wird, die die ausgewählte Eingabe für Gruppen verwendet, die durch die Zielkategorien gebildet werden. Diese Methode führt gegenüber einer erschöpfenden Suche (C&R-Baum) wiederum zu einer Steigerung der Geschwindigkeit bei der Bestimmung der optimalen Aufteilung.

Anforderungen. Eingabefelder können stetig (numerische Bereiche) sein, das Zielfeld muss aber kategorial sein. Alle Aufteilungen sind binär. Gewichtungsfelder können nicht eingesetzt werden. Alle im Modell verwendeten ordinalen Felder (sortiertes Set) müssen numerisch (nicht als Zeichenfolge) gespeichert sein. Im Bedarfsfall können Sie die Felder mit dem Umcodierungsknoten konvertieren.

Stärken. Genau wie "CHAID", aber im Gegensatz zu "C&R-Baum", verwendet "QUEST" statistische Tests, um zu entscheiden, ob ein Eingabefeld verwendet wird. Das Verfahren trennt auch die Eingabeauswahl von der Aufteilung und verwendet jeweils unterschiedliche Kriterien. Dies stellt einen Unterschied zu CHAID dar, wo das statistische Testergebnis, das die Variablenauswahl bestimmt, auch die Aufteilung erzeugt. "C&R-Baum" verfährt ähnlich, indem die Messung der Unreinheitsänderung sowohl die Auswahl des Eingabefelds als auch die Aufteilung bestimmt.

Entscheidungsbaumknoten - Feldoptionen

Auf der Registerkarte "Felder" geben Sie an, ob Sie die Feldrolleneinstellungen verwenden möchten, die bereits in den vorgeordneten Knoten definiert wurden, oder ob Sie die Feldzuweisungen manuell vornehmen möchten.

Vordefinierte Rollen verwenden Bei dieser Option werden die Rolleneinstellungen (Ziele, Prädiktoren usw.) aus einem Typknoten oberhalb verwendet (oder die Registerkarte "Typen" eines weiter oben im Stream gelegenen Quellenknotens).

Benutzerdefinierte Feldzuweisungen verwenden. Wählen Sie diese Option, wenn Sie die Ziele, Prädiktoren und andere Rollen in diesem Bildschirm manuell zuweisen möchten.

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts im Bildschirm Objekte aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

Klicken Sie auf die Schaltfläche **Alle**, damit alle Felder in der Liste ausgewählt werden, oder klicken Sie auf die Schaltfläche für ein individuelles Messniveau, um alle Felder mit diesem Messniveau auszuwählen.

Ziel. Wählen Sie ein Feld als Ziel für die Vorhersage aus.

Prädiktoren (Eingaben). Wählen Sie ein oder mehrere Felder als Eingaben für die Vorhersage aus.

Analysegewichtung. (Nur CHAID und C&R-Baum) Geben Sie das Feld hier an, um es als Fallgewichtung zu verwenden. Fallgewichtungen werden verwendet, um Differenzen in der Varianz zwischen den Ebenen des Ausgabefelds zu berücksichtigen. Weitere Informationen finden Sie im Thema „Verwenden von Häufigkeits- und Gewichtungsfeldern” auf Seite 33.

Entscheidungsbaumknoten - Erstellungsoptionen

Auf der Registerkarte "Erstellungsoptionen" legen Sie sämtliche Optionen zum Erstellen des Modells fest. Selbstverständlich können Sie auch einfach auf die Schaltfläche **Ausführen** klicken, um ein Modell mit allen Standardoptionen zu erstellen, doch in der Regel werden Sie Einstellungen zur Anpassung an Ihre eigenen Zwecke vornehmen wollen.

Hier können Sie wählen, ob Sie ein neues Modell erstellen oder ein vorhandenes aktualisieren möchten. Zudem legen Sie das Hauptziel des Knotens fest: Erstellung eines Standardmodells, Erstellung eines Modells mit erweiterter Genauigkeit oder Stabilität oder Erstellung eines Modells zur Verwendung für sehr umfangreiche Datasets.

Was möchten Sie tun?

Neues Modell erstellen. (Standard) Erstellt jedes Mal ein vollständig neues Modell, wenn Sie einen Stream mit diesem Modellierungsknoten ausführen.

Training des bestehenden Modells fortsetzen. In der Standardeinstellung wird bei jeder Ausführung eines Modellierungsknotens ein völlig neuees Modell erstellt. Bei Auswahl dieser Option wird das Training

mit dem letzten, vom Knoten erfolgreich aufgebauten Modell fortgesetzt. Dies ermöglicht die Aktualisierung eines bestehenden Modells, ohne dass ein Zugriff auf die ursprünglichen Daten erforderlich ist, und kann zu einer wesentlich schnelleren Leistung führen, da *ausschließlich* die neuen bzw. aktualisierten Datensätze in den Stream eingespeist werden. Details zum vorherigen Modell werden zusammen mit dem Modellierungsknoten gespeichert, wodurch diese Option auch dann verwendet werden kann, wenn das vorherige Modellnugget nicht mehr im Stream oder in der Modellpalette verfügbar ist.

Hinweis: Diese Option ist nur aktiviert, wenn Sie **Modell für sehr umfangreiche Datasets erstellen** als Ziel auswählen.

Wie lautet Ihr Hauptziel?

- **Einen einzelnen Baum erstellen.** Erstellt ein Standardmodell mit einzelnem Entscheidungsbaum. Standardmodelle können grundsätzlich einfacher interpretiert und schneller gescort werden, als Modelle, die unter Verwendung der anderen Zieloptionen erstellt werden.

  Modalwert. Legt fest, welche Methode für die Modellbildung verwendet wird. **Modell erzeugen** erstellt ein Modell automatisch, sobald der Stream ausgeführt wird. Die Option **Interaktive Sitzung starten** öffnet den Tree Builder, mit dem Sie Ihr Modell Ebene für Ebene erstellen, Aufteilungen bearbeiten und nach Wunsch reduzieren können, bevor Sie das Modellnugget erstellen.

  Direktiven verwenden. Mit dieser Option legen Sie Richtlinien fest, die angewendet werden, wenn aus dem Knoten ein interaktiver Baum generiert wird. Sie können beispielsweise die Aufteilungen der ersten und zweiten Ebene bestimmen, die dann beim Tree Builder-Start automatisch angewendet werden. Richtlinien einer interaktiven Baumerstellung können Sie auch speichern, um den Baum zu einem späteren Zeitpunkt noch einmal zu erstellen. Weitere Informationen finden Sie im Thema „Aktualisieren der Direktiven" auf Seite 92.
- **Modellgenauigkeit verbessern (Boosting).** Mit dieser Option wählen Sie eine spezielle Methode aus, die als **Boosting** bekannt ist, um die Modellgenauigkeitsquote zu erhöhen. Das Boosting funktioniert so, dass mehrere Modelle in einer Folge erstellt werden. Das erste Modell wird auf die übliche Weise erstellt. Anschließend wird ein zweites Modell erstellt, bei dem besonders die Datensätze berücksichtigt werden, bei denen es im ersten Modell zu Fehlklassifizierungen kam. Das dritte Modell wird in Bezug auf die im zweiten Modell enthaltenen Fehler erstellt usw. Zum Schluss werden die Fälle klassifiziert, indem der gesamte Modellsatz auf ihnen angewendet wird, wobei ein gewichtetes Voting-Verfahren genutzt wird, um die einzelnen Vorhersagen zu einer Gesamtvorhersage zu kombinieren. Das Boosting kann die Genauigkeit eines Entscheidungsbaummodells signifikant verbessern, macht aber auch ein längeres Training notwendig.
- **Modellstabilität verbessern (Bagging).** Mit dieser Option wählen Sie ein spezielles Verfahren, das als **Bagging** (Bootstrap-Aggregation) bekannt ist, um die Stabilität des Modells zu verbessern und eine Überanpassung zu vermeiden. Mit dieser Option werden mehrere Modelle erstellt und kombiniert, um zuverlässigere Vorhersagen zu erhalten. Mithilfe dieser Option gebildete Modelle benötigen mehr Zeit bei der Erstellung und beim Scoren als Standardmodelle.
- **Modell für sehr umfangreiche Datasets erstellen.** Wählen Sie diese Option aus, wenn Sie mit Datasets arbeiten, die zu groß für die Erstellung eines Modells mithilfe der anderen Zieloptionen sind. Diese Option unterteilt die Daten in kleinere Datenblöcke und erstellt auf jedem Block ein Modell. Die genauesten Modelle werden anschließend automatisch ausgewählt und zu einem einzigen Modellnugget zusammengefasst. Durch Auswahl der Option **Training des bestehenden Modells fortsetzen** auf diesem Bildschirm können Sie eine inkrementelle Modellaktualisierung durchführen. *Hinweis*: Diese Option ist für sehr umfangreiche Datasets gedacht und benötigt eine Verbindung zu IBM SPSS Modeler Server.

Entscheidungsbaumknoten - Grundeinstellungen

Hier nehmen Sie die Grundeinstellungen für die Erstellung des Entscheidungsbaums vor.

Aufbau-Algorithmen für Bäume. (Nur CHAID) Wählen Sie den **CHAID**-Algorithmus aus, den Sie verwenden möchten. **Exhaustive CHAID** ist eine Variante von CHAID, die noch gründlicher vorgeht, indem sie alle für jeden Prädiktor möglichen Aufteilungen untersucht, allerdings mehr Rechenzeit beansprucht.

Maximale Baumtiefe. Legen Sie die maximale Anzahl der Ebenen unter dem Stammknoten fest (wie oft die Stichprobe rekursiv aufgeteilt wird). Der Standardwert ist 5. Wählen Sie **Benutzerdefiniert** aus und geben Sie einen Wert ein, um eine andere Anzahl an Niveaus anzugeben.

Reduzieren (nur C&R-Baum und QUEST)

Baum beschneiden, um eine Überanpassung zu vermeiden. Die Reduzierung besteht im Entfernen von Aufteilungen der unteren Ebene, die keine signifikante Auswirkung auf die Genauigkeit des Baums besitzen. Durch Reduzierung kann der Baum vereinfacht werden, wodurch er leichter zu interpretieren ist. In einigen Fällen kann er auch besser verallgemeinert werden. Lassen Sie diese Option inaktiviert, wenn Sie den vollständigen Baum ohne Reduktion erhalten wollen.

- **Maximale Risikodifferenz (in Standardfehlern).** Mit dieser Option können Sie eine etwas liberalere Reduzierungsregel angeben. Die Standardfehlerregel ermöglicht dem Algorithmus, den einfachsten Baum auszuwählen, dessen Risikoschätzung nahe (oder möglichst über) der des untergeordneten Baums mit dem kleinsten Risiko liegt. Der Wert gibt die Größe der zulässigen Differenz zwischen der Risikoschätzung für den reduzierten Baum und der des Baums mit dem kleinsten Risiko an. Wenn Sie beispielsweise 2 angeben, kann ein Baum ausgewählt werden, dessen Risikoschätzung (2 × Standardfehler) größer als die des vollständigen Baums ist.

Maximale Anzahl der Ersatztrenner. Ersatztrenner sind eine Methode für die Behandlung fehlender Werte. Für jede Aufteilung ermittelt der Algorithmus die Eingabefelder, die dem ausgewählten Aufteilungsfeld am ähnlichsten sind. Diese Felder sind die **Ersatztrenner** für diese Aufteilung. Wenn ein Datensatz klassifiziert werden muss, aber in einem Aufteilungsfeld ein Wert fehlt, kann für die Aufteilung der entsprechende Wert eines Ersatztrennerfelds verwendet werden. Eine höhere Einstellung sorgt für mehr Flexibilität bei der Behandlung fehlender Werte, verursacht allerdings auch eine stärke Arbeitsspeicherverwendung und längere Trainingszeiten.

Entscheidungsbaumknoten - Stoppregeln

Diese Optionen steuern, wie der Baum konstruiert wird. Grenzregeln legen fest, wann mit dem Aufteilen bestimmter Verzweigungen des Baums aufgehört wird. Damit legen Sie die minimale Verzweigungsgröße fest. Dies verhindert Aufteilungen, die zu sehr kleinen Untergruppen führen. Die Option **Mindestanzahl der Datensätze in übergeordneter Verzweigung** verhindert eine Aufteilung, wenn die Anzahl der im aufzuteilenden Knoten (die **übergeordnete Verzweigung**) enthaltenen Datensätze geringer ist, als der hier angegebene Wert. Die Option **Mindestanzahl der Datensätze in untergeordneter Verzweigung** verhindert eine Aufteilung, wenn die Anzahl der in einer durch die Aufteilung erzeugten Verzweigung (die **untergeordnete Verzweigung**) enthaltenen Datensätze geringer ist, als der hier angegebene Wert.

- **Prozentsatz verwenden.** Hiermit können Sie Größen als Prozentsätze der gesamten Trainingsdaten angeben.
- **Absolutwert verwenden.** Hiermit können Sie Größen als absolute Anzahl von Datensätzen angeben.

Entscheidungsbaumknoten - Ensembles

Diese Einstellungen legen das Verhalten der Ensemblebildung fest, die erfolgt, wenn auf der Registerkarte "Ziele" die Option "Boosting", "Bagging" oder "Sehr große Datasets" ausgewählt ist. Optionen, die für das ausgewählte Ziel nicht gelten, werden ignoriert.

Bagging und sehr umfangreiche Datasets. Beim Scoren eines Ensembles wird diese Regel angewendet, um die vorhergesagten Werte aus den Basismodellen für die Berechnung des Scorewerts für das Ensemble zu kombinieren.

- **Standardkombinationsregel für kategoriale Ziele.** Ensemblevorhersagewerte für kategoriale Ziele können mithilfe von "Voting", "höchster Wahrscheinlichkeit" oder "höchste mittlere Wahrscheinlichkeit" kombiniert werden. Mit **Voting** wird die Kategorie gewählt, die in allen Basismodellen am häufigsten die höchste Wahrscheinlichkeit erreicht. Mit **Höchste Wahrscheinlichkeit** wird die Kategorie gewählt, die in allen Basismodellen den höchsten Einzelwert bei der höchsten Wahrscheinlichkeit erzielt. Mit

Höchste mittlere Wahrscheinlichkeit wird die Kategorie mit dem höchsten Wert ausgewählt, wenn der Mittelwert der Kategoriewahrscheinlichkeiten aus der Menge aller Basismodelle berechnet wird.

- **Standardkombinationsregel für stetige Ziele.** Ensemblevorhersagewerte für stetige Ziele können unter Verwendung des Mittelwerts oder Medians der Vorhersagewerte aus den Basismodellen kombiniert werden.

Hinweis: Wenn als Ziel die Verbesserung der Modellgenauigkeit ausgewählt wurde, wird die Auswahl zum Kombinieren der Regeln ignoriert. Beim Boosting wird für das Scoring der kategorialen Ziele stets eine gewichtete Mehrheit verwendet und für das Scoring stetiger Ziele ein gewichteter Median.

Boosting und Bagging. Geben Sie die Anzahl der zu erstellenden Basismodelle an, wenn als Ziel die Verbesserung der Modellgenauigkeit oder -stabilität angegeben ist. Im Falle des Bagging ist das die Anzahl der Bootstrap-Stichproben. Muss eine positive ganze Zahl sein.

C&R-Baum- und QUEST-Knoten - Kosten & A-priori

Fehlklassifizierungskosten

In einigen Zusammenhängen sind bestimmte Fehler kostspieliger als andere. Es kann beispielsweise kostspieliger sein, einen Antragsteller für einen Kredit mit einem hohen Risiko als niedriges Risiko zu klassifizieren (eine Art von Fehler) als einen Antragsteller mit einem niedrigen Risiko als hohes Risiko (eine andere Art von Fehler) zu klassifizieren. Anhand von Fehlklassifizierungskosten können Sie die relative Bedeutung verschiedener Arten von Vorhersagefehlern angeben.

Fehlklassifizierungskosten sind im Grunde Gewichtungen, die auf bestimmte Ergebnisse angewendet werden. Diese Gewichtungen werden in das Modell mit einbezogen und können die Vorhersage (als Schutz gegen kostspielige Fehler) ändern.

Mit Ausnahme von C5.0-Modellen werden Fehlklassifizierungskosten beim Scoren von Modellen nicht angewendet und bei der Rangbildung bzw. beim Vergleich von Modellen mithilfe eines Knotens vom Typ "Autom. Klassifikationsmerkmal", eines Evaluierungsdiagramms oder eines Analyseknotens nicht berücksichtigt. Ein Modell, das Kosten beinhaltet, führt nicht unbedingt zu weniger Fehlern als ein Modell, das keine Kosten beinhaltet, und es weist auch nicht unbedingt eine höhere Gesamtgenauigkeit auf, aber es ist möglicherweise in praktischer Hinsicht leistungsfähiger, da es eine integrierte Verzerrung zugunsten von *weniger teuren* Fehlern aufweist.

Die Kostenmatrix zeigt die Kosten für jede mögliche Kombination aus prognostizierter Kategorie und tatsächlicher Kategorie. Standardmäßig werden alle Fehlklassifizierungskosten auf 1,0 gesetzt. Um benutzerdefinierte Kostenwerte einzugeben, wählen Sie **Fehlklassifizierungskosten verwenden** und geben Ihre benutzerdefinierten Werte in die Kostenmatrix ein.

Um die Fehlklassifizierungskosten zu ändern, wählen Sie die Zelle aus, die der gewünschten Kombination aus vorhergesagten und tatsächlichen Werten entspricht, löschen den Inhalt der Zelle und geben die gewünschten Kosten für die Zelle ein. Die Kosten sind nicht automatisch symmetrisch. Wenn Sie z. B. die Kosten einer Fehlklassifizierung von *A* als *B* auf 2,0 setzen, weisen die Kosten für eine Fehlklassifizierung von *B* als *A* weiterhin den Standardwert 1,0 auf, es sei denn, Sie ändern diesen Wert ausdrücklich.

A-priori-Wahrscheinlichkeiten

Mit diesen Optionen können Sie die A-priori-Wahrscheinlichkeiten für Kategorien bei einer Vorhersage eines kategorialen Zielfelds angeben. **A-priori-Wahrscheinlichkeiten** sind Schätzungen der gesamten relativen Häufigkeit für jede Zielkategorie in der Gesamtheit, aus der die Trainingsdaten gezogen werden. Mit anderen Worten: Es handelt sich um die Wahrscheinlichkeitsschätzungen, die Sie für jeden möglichen Zielwert vornehmen würden, *bevor* Sie etwas über die Prädiktorwerte wissen. Es gibt drei Methoden, A-priori-Werte festzulegen:

- **Auf Trainingsdaten basierend.** Dies ist die Standardeinstellung. A-priori-Wahrscheinlichkeiten basieren auf den relativen Häufigkeiten der Kategorien in den Trainingsdaten.
- **Für alle Klassen gleich.** A-priori-Wahrscheinlichkeiten für alle Kategorien werden als $1/k$ definiert, wobei k die Zahl der Zielkategorien darstellt.
- **Benutzerdefiniert.** Sie können eigene A-priori-Wahrscheinlichkeiten angeben. Die Startwerte für A-priori-Wahrscheinlichkeiten werden für alle Klassen gleich gesetzt. Sie können die Wahrscheinlichkeiten für einzelne Kategorien auf benutzerdefinierte Werte einstellen. Um die Wahrscheinlichkeit einer bestimmten Kategorie anzupassen, wählen Sie die Wahrscheinlichkeitszelle in der Tabelle aus, die der gewünschten Kategorie entspricht, löschen den Inhalt der Zelle und geben den gewünschten Wert ein.

Die A-priori-Wahrscheinlichkeiten für alle Kategorien sollten sich auf 1,0 summieren (die **Wahrscheinlichkeitsbeschränkung**). Wenn sie keine Summe von 1,0 bilden, wird eine Warnnachricht ausgegeben und es besteht die Möglichkeit, die Werte automatisch normalisieren zu lassen. Diese automatische Anpassung behält die Anteile über die Kategorien hinweg bei, während die Wahrscheinlichkeitsbeschränkung erzwungen wird. Sie können diese Anpassung jederzeit durchführen, indem Sie auf die Schaltfläche **Normalisieren** klicken. Um die Tabelle auf gleiche Werte für alle Kategorien zurückzusetzen, klicken Sie auf die Schaltfläche **Gleichsetzen**.

A-priori-Wahrscheinlichkeiten anhand der Fehlklassifizierungskosten korrigieren. Mit dieser Option können Sie die A-priori-Wahrscheinlichkeiten basierend auf den Fehlklassifizierungskosten (auf der Registerkarte "Kosten" angegeben) anpassen. Dadurch können Sie Kosteninformationen für Bäume, die mit dem Twoing-Unreinheitsmaß arbeiten, direkt in den Vorgang zur Baumerweiterung aufnehmen. (Wenn diese Option nicht ausgewählt ist, werden Kosteninformationen nur beim Klassifizieren der Datensätze und bei der Berechnung der Risikoschätzungen für Bäume auf der Grundlage des Twoing-Maßes verwendet.)

CHAID-Knoten - Kosten

In einigen Zusammenhängen sind bestimmte Fehler kostspieliger als andere. Es kann beispielsweise kostspieliger sein, einen Antragsteller für einen Kredit mit einem hohen Risiko als niedriges Risiko zu klassifizieren (eine Art von Fehler) als einen Antragsteller mit einem niedrigen Risiko als hohes Risiko (eine andere Art von Fehler) zu klassifizieren. Anhand von Fehlklassifizierungskosten können Sie die relative Bedeutung verschiedener Arten von Vorhersagefehlern angeben.

Fehlklassifizierungskosten sind im Grunde Gewichtungen, die auf bestimmte Ergebnisse angewendet werden. Diese Gewichtungen werden in das Modell mit einbezogen und können die Vorhersage (als Schutz gegen kostspielige Fehler) ändern.

Mit Ausnahme von C5.0-Modellen werden Fehlklassifizierungskosten beim Scoren von Modellen nicht angewendet und bei der Rangbildung bzw. beim Vergleich von Modellen mithilfe eines Knotens vom Typ "Autom. Klassifikationsmerkmal", eines Evaluierungsdiagramms oder eines Analyseknotens nicht berücksichtigt. Ein Modell, das Kosten beinhaltet, führt nicht unbedingt zu weniger Fehlern als ein Modell, das keine Kosten beinhaltet, und es weist auch nicht unbedingt eine höhere Gesamtgenauigkeit auf, aber es ist möglicherweise in praktischer Hinsicht leistungsfähiger, da es eine integrierte Verzerrung zugunsten von *weniger teuren* Fehlern aufweist.

Die Kostenmatrix zeigt die Kosten für jede mögliche Kombination aus prognostizierter Kategorie und tatsächlicher Kategorie. Standardmäßig werden alle Fehlklassifizierungskosten auf 1,0 gesetzt. Um benutzerdefinierte Kostenwerte einzugeben, wählen Sie **Fehlklassifizierungskosten verwenden** und geben Ihre benutzerdefinierten Werte in die Kostenmatrix ein.

Um die Fehlklassifizierungskosten zu ändern, wählen Sie die Zelle aus, die der gewünschten Kombination aus vorhergesagten und tatsächlichen Werten entspricht, löschen den Inhalt der Zelle und geben die gewünschten Kosten für die Zelle ein. Die Kosten sind nicht automatisch symmetrisch. Wenn Sie z. B. die Kosten einer Fehlklassifizierung von *A* als *B* auf 2,0 setzen, weisen die Kosten für eine Fehlklassifizierung von *B* als *A* weiterhin den Standardwert 1,0 auf, es sei denn, Sie ändern diesen Wert ausdrücklich.

Knoten "C&R" - Erweitert

Mit den erweiterten Optionen können Sie die Feinabstimmung der Baumerstellung vornehmen.

Minimale Änderung in Unreinheit. Legt die minimale Änderung in der Unreinheit fest, damit im Baum eine neue Aufteilung erstellt wird. **Unreinheit** bezieht sich auf das Ausmaß, in dem durch den Baum definierte Untergruppen in jeder Gruppe eine große Reihe von Ausgabefeldwerten besitzen. Bei kategorialen Zielen wird ein Knoten als "rein" betrachtet, wenn 100 % der im Knoten vorhandenen Fälle in eine bestimmte Kategorie des Zielfelds fallen. Das Ziel der Baumerstellung besteht darin, Untergruppen mit ähnlichen Ausgabewerten zu erstellen, also die Unreinheit in jedem Knoten zu minimieren. Wenn die beste für eine Verzweigung mögliche Aufteilung die Unreinheit um weniger als den vorgegebenen Betrag reduziert, wird die Aufteilung nicht durchgeführt.

Unreinheitsmaß für kategoriale Ziele. Geben Sie für kategoriale Zielfelder die für die Messung der im Baum vorhandenen Unreinheit verwendete Methode an. (Für stetige Ziele wird diese Option ignoriert und als Unreinheitsmaß immer die **kleinste quadratische Abweichung** verwendet.)

- **Gini** ist ein allgemeines Unreinheitsmaß, das auf Wahrscheinlichkeiten der Zugehörigkeit zu einer Kategorie einer Verzweigung basiert.
- **Twoing** ist ein Unreinheitsmaß, das die binäre Aufteilung betont und eher zu einer Aufteilung in annähernd gleichgroße Verzweigungen führt.
- **Ordinal** fügt eine weitere Einschränkung hinzu, indem nur zusammenhängende Zielklassen zu Gruppen zusammengefasst werden können, was nur bei Ordinalzielen möglich ist. Wenn diese Option für ein nominales Ziel ausgewählt ist, wird standardmäßig das Standard-Twoing-Maß verwendet.

Set zur Verhinderung übermäßiger Anpassung. Bei dem Algorithmus werden Datensätze intern in ein Modellerstellungsset und ein Set zur Verhinderung übermäßiger Anpassung aufgeteilt. Letzeres ist ein unabhängiges Set an Datensätzen, das dazu verwendet wird, Fehler während des Trainings zu erfassen. So kann verhindert werden, dass die Methode zufällige Variationen in den Daten modelliert. Geben Sie einen Prozentsatz an Datensätzen an. Der Standardwert ist 30.

Ergebnisse replizieren. Durch Einstellen eines Startwerts für den Zufallsgenerator können Analysen reproduziert werden. Geben Sie eine ganze Zahl ein oder klicken Sie auf **Generieren**. Dadurch wird eine pseudozufällige Ganzzahl zwischen 1 und 2147483647 (einschließlich) erzeugt.

QUEST-Knoten - Erweitert

Mit den erweiterten Optionen können Sie die Feinabstimmung der Baumerstellung vornehmen.

Signifikanzniveau für Aufteilung. Legt das Signifikanzniveau (Alpha) für das Aufteilen von Knoten fest. Der Wert muss zwischen 0 und 1 liegen. Niedrigere Werte führen in der Regel zu Bäumen mit weniger Knoten.

Set zur Verhinderung übermäßiger Anpassung. Bei dem Algorithmus werden Datensätze intern in ein Modellerstellungsset und ein Set zur Verhinderung übermäßiger Anpassung aufgeteilt. Letzeres ist ein unabhängiges Set an Datensätzen, das dazu verwendet wird, Fehler während des Trainings zu erfassen. So kann verhindert werden, dass die Methode zufällige Variationen in den Daten modelliert. Geben Sie einen Prozentsatz an Datensätzen an. Der Standardwert ist 30.

Ergebnisse replizieren. Durch Einstellen eines Startwerts für den Zufallsgenerator können Analysen reproduziert werden. Geben Sie eine ganze Zahl ein oder klicken Sie auf **Generieren**. Dadurch wird eine pseudozufällige Ganzzahl zwischen 1 und 2147483647 (einschließlich) erzeugt.

CHAID-Knoten- Erweitert

Mit den erweiterten Optionen können Sie die Feinabstimmung der Baumerstellung vornehmen.

Signifikanzniveau für Aufteilung. Legt das Signifikanzniveau (Alpha) für das Aufteilen von Knoten fest. Der Wert muss zwischen 0 und 1 liegen. Niedrigere Werte führen in der Regel zu Bäumen mit weniger Knoten.

Signifikanzniveau für Zusammenführung. Legt das Signifikanzniveau (Alpha) für das Zusammenführen von Kategorien fest. Der Wert muss größer als 0 und kleiner oder gleich 1 sein. Um zu vermeiden, dass Kategorien zusammengeführt werden, geben Sie den Wert 1 an. Bei stetigen Zielen bedeutet dies, dass die Anzahl der Kategorien für die Variable im Endbaum mit der angegebenen Anzahl von Intervallen übereinstimmt. Diese Option ist für Exhaustive CHAID nicht verfügbar.

Signifikanzwerte mit der Bonferroni-Methode anpassen. Passt beim Testen der verschiedenen Kategoriekombinationen eines Prädiktors die Signifikanzwerte an. Die Werte werden auf der Grundlage der Anzahl der Tests angepasst, die sich direkt auf die Anzahl der Kategorien und das Messniveau eines Prädiktors bezieht. Dies ist in der Regel wünschenswert, da eine bessere Kontrolle der falsch positiven Fehlerrate stattfindet. Das Inaktivieren dieser Option erhöht die Leistung Ihrer Analyse beim Auffinden tatsächlicher Differenzen, führt aber zu einer höheren falsch positiven Rate. Insbesondere für kleine Stichproben kann das Inaktivieren dieser Option ratsam sein.

Erneute Aufteilung zusammengeführter Kategorien in einem Knoten erlauben. Der CHAID-Algorithmus versucht Kategorien zusammenzuführen, um den einfachsten, das Modell beschreibenden Baum zu erzeugen. Wenn diese Option ausgewählt ist, können zusammengeführte Kategorien erneut aufgeteilt werden, wenn dies zu einer besseren Lösung führt.

Chi-Quadrat für kategoriale Ziele. Für kategoriale Ziele können Sie die für die Berechnung der Chi-Quadrat-Statistik verwendete Methode angeben.

- **Pearson.** Diese Methode liefert schnellere Berechnungen, sollte bei kleineren Stichproben jedoch nur nach sorgfältiger Erwägung verwendet werden.
- **Likelihood-Quotient.** Diese Methode ist robuster als Pearson, benötigt aber mehr Rechenzeit. Diese Methode eignet sich ideal für kleine Stichproben. Bei stetigen Zielen wird immer diese Methode verwendet.

Minimale Änderung in erwarteten Zellhäufigkeiten. Beim Schätzen der Zellhäufigkeiten (des nominalen Modells und des ordinalen Modells der Zeileneffekte) kommt eine iterative Prozedur (Epsilon) zum Einsatz, um ein Konvergieren gegen die optimale Schätzung zu erreichen, die im Chi-Quadrat-Test für eine bestimmte Aufteilung verwendet wird. Epsilon bestimmt, wie groß die Änderung sein muss, damit Iterationen fortgesetzt werden. Wenn die aus der letzten Iteration resultierende Änderung unter dem festgelegten Wert liegt, wird die Iteration beendet. Wenn Sie Probleme damit haben, dass der Algorithmus nicht konvergiert, können Sie diesen Wert erhöhen oder die maximale Anzahl der Iterationen so lange reduzieren, bis die Konvergenz stattfindet.

Maximale Anzahl der Iterationen für Konvergenz. Legt die maximale Anzahl der Iterationen fest, nach der aufgehört wird, egal ob eine Konvergenz stattgefunden hat oder nicht.

Ergebnisse replizieren. Durch Einstellen eines Startwerts für den Zufallsgenerator können Analysen reproduziert werden. Geben Sie eine ganze Zahl ein oder klicken Sie auf **Generieren**. Dadurch wird eine pseudozufällige Ganzzahl zwischen 1 und 2147483647 (einschließlich) erzeugt.

Modelloptionen für Entscheidungsbaumknoten

Auf der Registerkarte "Modelloptionen" können Sie bestimmen, ob Sie einen Namen für das Modell angeben oder automatisch erstellen lassen möchten. Außerdem können Sie auswählen, ob Sie Informationen zum Prädiktoreinfluss sowie Scores für Raw Propensity und Adjusted Propensity für Flagziele erhalten möchten.

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Modellevaluation

Prädiktoreinfluss berechnen. Bei Modellen, die zu einem angemessenen Maß an Bedeutsamkeit führen, können Sie ein Diagramm anzeigen, in dem der relative Einfluss der einzelnen Prädiktoren bei der Modellschätzung angegeben wird. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und diejenigen Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass die Berechnung des Prädiktoreinflusses bei einigen Modellen längere Zeit in Anspruch nehmen kann, insbesondere bei der Arbeit mit großen Datasets, und daher bei einigen Modellen standardmäßig inaktiviert ist. Der Prädiktoreinfluss ist für Entscheidungslistenmodelle nicht verfügbar. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Propensity-Scores

Propensity-Scores können im Modellierungsknoten oder auf der Registerkarte "Einstellungen" im Modellnugget aktiviert werden. Diese Funktionalität ist nur verfügbar, wenn das ausgewählte Ziel ein Flagfeld ist. Weitere Informationen finden Sie im Thema „Propensity-Scores" auf Seite 36.

Raw-Propensity-Scores berechnen. Raw-Propensity-Scores werden ausschließlich auf der Grundlage der Trainingsdaten aus dem Modell abgeleitet. Wenn das Modell den Wert *wahr* (wird antworten) vorhersagt, ist die Neigung mit P identisch. Dabei ist P die Wahrscheinlichkeit der Vorhersage. Wenn das Modell den Wert "falsch" vorhersagt, wird die Neigung als (1 - P) berechnet.

- Wenn Sie bei der Modellerstellung diese Option auswählen, werden standardmäßig Propensity-Scores im Modellnugget aktiviert. Sie können jedoch immer festlegen, dass Raw-Propensity-Scores im Modellnugget aktiviert werden sollen, unabhängig davon, ob Sie sie im Modellierungsknoten auswählen.
- Beim Scoren des Modells werden Raw-Propensity-Scores in einem Feld hinzugefügt, bei dem die Buchstaben *RP* an das Standardpräfix angehängt sind. Wenn sich die Vorhersagen beispielsweise im Feld *$R-Abwanderung* befinden, lautet der Name des Felds für den Propensity-Score *$RRP-Abwanderung*.

Adjusted-Propensity-Scores berechnen. Raw Propensitys basieren ausschließlich auf vom Modell angegebenen Schätzungen. Beim Modell kann jedoch eine Überanpassung vorliegen, was zu übermäßig optimistischen Schätzungen für die Neigung führt. Adjusted Propensitys versuchen, dies zu kompensieren, indem untersucht wird, wie leistungsfähig das Modell bei den Test- bzw. Validierungspartitionen ist, und die Neigungen entsprechend angepasst werden, um eine bessere Schätzung zu erzielen.

- Diese Einstellung ist nur möglich, wenn ein gültiges Partitionsfeld im Stream vorhanden ist.
- Anders als rohe Konfidenzscores müssen Adjusted-Propensity-Scores bei der Erstellung des Modells berechnet werden; andernfalls sind sie beim Scoren des Modellnuggets nicht verfügbar.
- Beim Scoren des Modells werden Adjusted-Propensity-Scores in einem Feld hinzugefügt, bei dem die Buchstaben *AP* an das Standardpräfix angehängt sind. Wenn sich die Vorhersagen beispielsweise im Feld *$R-Abwanderung* befinden, lautet der Name des Felds für den Propensity-Score *$RAP-Abwanderung*. Adjusted-Propensity-Scores sind bei logistischen Regressionsmodellen nicht verfügbar.
- Bei der Berechnung der Adjusted-Propensity-Scores darf die für die Berechnung verwendete Test- bzw. Validierungspartition nicht ausbalanciert worden sein. Um dies zu vermeiden, müssen Sie darauf achten, dass in etwaigen weiter oben im Stream befindlichen Balancierungsknoten die Option **Balancierung nur für Trainingsdaten durchführen** ausgewählt wurde. Zusätzlich gilt: Wenn eine komplexe Stichprobe gezogen wurde, werden dadurch die Adjusted-Propensity-Scores ungültig.
- Adjusted-Propensity-Scores sind bei verstärkten Baum- und Regelsetmodellen nicht verfügbar. Weitere Informationen finden Sie im Thema „Verbesserte C5.0-Modelle" auf Seite 113.

Basierend auf. Um Adjusted-Propensity-Scores berechnen zu können, muss im Stream ein Partitionsfeld vorhanden sein. Sie können angeben, ob die Test- bzw. Validierungspartition für diese Berechnung verwendet werden soll. Um bestmögliche Ergebnisse zu erzielen, sollte die Test- bzw. Validierungspartition mindestens so viele Datensätze enthalten wie die Partition, die zum Trainieren des ursprünglichen Modells verwendet wurde.

C5.0-Knoten

Hinweis: Diese Funktion ist in SPSS Modeler Professional und SPSS Modeler Premium verfügbar.

Dieser Knoten verwendet den C5.0-Algorithmus, um entweder einen **Entscheidungsbaum** oder ein **Regelset** zu erstellen. Ein C5.0-Modell teilt die Stichprobe auf der Basis des Felds auf, das den maximalen **Informationsgewinn** liefert. Jede durch die erste Aufteilung definierte Teilstichprobe wird anschließend wieder aufgeteilt, üblicherweise auf der Grundlage eines anderen Felds. Der Prozess wird so lange fortgesetzt, bis die Unterstichproben nicht weiter aufgeteilt werden können. Zum Schluss werden die Aufteilungen der untersten Ebene noch einmal untersucht, wobei solche entfernt oder **reduziert** werden, die nicht wesentlich zum Wert des Modells beitragen.

Hinweis: Der C5.0-Knoten kann nur ein kategoriales Ziel vorhersagen. Bei der Analyse von Daten mit kategorialen Feldern (nominal oder ordinal) fasst der Knoten mit größerer Wahrscheinlichkeit Kategorien zu einer Gruppe zusammen als C5.0-Versionen vor Version 11.0.

C5.0 kann zwei Arten von Modellen erstellen. Ein **Entscheidungsbaum** ist eine einfache Beschreibung der vom Algorithmus gefundenen Aufteilungen. Jeder Endknoten (oder Blattknoten) beschreibt ein bestimmtes Subset der Trainingsdaten. Und jeder in den Trainingsdaten vorhandene Fall gehört zu genau einem im Baum vorhandenen Endknoten. Somit ist für jeden in einem Entscheidungsbaum vorhandenen Datensatz genau eine Vorhersage möglich.

Ein **Regelset** ist dagegen eine Menge von Regeln, mit der versucht wird, Vorhersagen für einzelne Datensätze zu erstellen. Regelsets werden aus Entscheidungsbäumen abgeleitet und stellen eine vereinfachte oder konzentrierte Version der im Entscheidungsbaum gefundenen Informationen dar. Regelsets enthalten meist die wichtigsten Informationen eines gesamten Entscheidungsbaums, allerdings mit einem weniger komplexen Modell. Regelsets arbeiten anders als Entscheidungsbäume und besitzen daher nicht dieselben Eigenschaften. Der wichtigste Unterschied besteht darin, dass es bei einem Regelset möglich ist, dass für einen bestimmten Datensatz mehr als eine oder aber überhaupt keine Regel gilt. Wenn mehrere Regeln gelten, dann wird jeder Regel ein gewichtetes "Votum" zugeordnet, das auf der dieser Regel zugeordneten Konfidenz basiert, und die endgültige Vorhersage ergibt sich aus der Kombination der gewichteten Voten aller für den fraglichen Datensatz geltenden Regeln. Wenn keine Regel gilt, wird dem Datensatz eine Standardvorhersage zugeordnet.

Beispiel. Ein Medizinforscher hat Daten über eine Gruppe von Patienten zusammengetragen, die alle an der gleichen Krankheit leiden. Im Behandlungsverlauf sprach jeder Patient auf eines von fünf Medikamenten an. Sie können ein C5.0-Modell in Verbindung mit anderen Knoten verwenden, um herauszufinden, welches Medikament für einen zukünftigen Patienten mit derselben Krankheit geeignet sein könnte.

Anforderungen. Um ein C5.0-Modell zu trainieren, muss genau ein kategoriales (d. h. vom Typ "Nominal" oder "Ordinal") *Ziel*-Feld und mindestens ein *Eingabe*-Feld beliebigen Typs vorliegen. Felder, die auf *Beides* oder *Keine* gesetzt sind, werden ignoriert. Bei den im Modell verwendeten Feldern müssen die Typen vollständig instanziiert sein. Außerdem kann ein Gewichtungsfeld angegeben werden.

Stärken. C5.0-Modelle verhalten sich bei Problemen mit fehlenden Daten und einer großen Anzahl von Eingabefelder sehr robust. Sie benötigen für die Schätzung in der Regel keine langen Trainingsphasen. Darüber hinaus sind C5.0-Modelle tendenziell leichter verständlich als andere Modelltypen, da sich die aus dem Modell abgeleiteten Regeln sehr direkt interpretieren lassen. C5.0 bietet außerdem die leistungsstarke Methode des **Boosting**, mit der die Genauigkeit der Klassifizierung gesteigert wird.

Hinweis: Die Geschwindigkeit für die Erstellung von C5.0-Modellen kann durch die Aktivierung der parallelen Verarbeitung verbessert werden.

Modelloptionen für C5.0-Knoten

Modellname. Geben Sie den Namen des zu erstellenden Modells an.

- **Auto.** Wenn diese Option ausgewählt ist, wird der Modellname automatisch auf der Grundlage der Namen der Zielfelder generiert. Dies ist die Standardeinstellung.
- **Benutzerdefiniert.** Wählen Sie diese Option, um für das durch diesen Knoten erstellte Modellnugget einen eigenen Namen anzugeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Ausgabetyp. Legen Sie hier fest, ob das resultierende Modellnugget ein **Entscheidungsbaum** oder ein **Regelset** sein soll.

Symbolische Werte gruppieren. Wenn diese Option ausgewählt ist, versucht C5.0 symbolische Werte zu gruppieren, die in Bezug auf das Ausgabefeld ähnliche Muster aufweisen. Wenn diese Option nicht ausgewählt ist, erzeugt C5.0 für jeden Wert des symbolischen Felds einen untergeordneten Knoten, der zum Aufteilen des übergeordneten Knotens verwendet wird. C5.0 teilt beispielsweise das Feld *FARBE* (mit den Werten *ROT*, *GRÜN* und *BLAU*) standardmäßig in drei Teile. Wenn diese Option jedoch ausgewählt ist und die Datensätze mit *FARBE = ROT* sehr ähnlich aussehen wie Datensätze mit *FARBE = BLAU*, dann wird eine Zweiteilung durchgeführt, bei der alle Datensätze mit *GRÜN* eine Gruppe bilden und alle mit *BLAU* und *ROT* gemeinsam eine zweite.

Boosting verwenden. Der C5.0-Algorithmus verwendet eine spezielle Methode zur Verbesserung der Genauigkeitsrate, die als **Boosting** bezeichnet wird. Sie arbeitet so, dass sie mehrere Modelle in einer Folge erstellt. Das erste Modell wird auf die übliche Weise erstellt. Anschließend wird ein zweites Modell erstellt, bei dem besonders die Datensätze berücksichtigt werden, bei denen es im ersten Modell zu Fehlklassifizierungen kam. Das dritte Modell wird in Bezug auf die im zweiten Modell enthaltenen Fehler erstellt usw. Zum Schluss werden die Fälle klassifiziert, indem der gesamte Modellsatz auf ihnen angewendet wird, wobei ein gewichtetes Voting-Verfahren genutzt wird, um die einzelnen Vorhersagen zu einer Gesamtvorhersage zu kombinieren. Das Boosting kann die Genauigkeit eines C5.0-Modells signifikant verbessern, macht aber auch ein längeres Training notwendig. Mit der Option **Anzahl der Versuche** können Sie steuern, wie viele Modelle für das verstärkte Modell verwendet werden. Diese Funktion basiert auf der Forschung von Freund & Schapire, mit einigen gesetzlich geschützten Verbesserungen für die Behandlung verrauschter Daten.

Kreuzvalidieren. Mit dieser Option verwendet C5.0 einen Satz von Modellen, die aus einem Subset der Trainingsdaten erstellt werden, um die Genauigkeit eines aus dem gesamten Dataset erstellten Modells zu schätzen. Dies ist nützlich, wenn das Dataset zu klein ist, um in herkömmliche Trainings- und Testsets aufgeteilt zu werden. Die Kreuzvalidierungsmodelle werden nach der Berechnung der Genauigkeitsschätzung entfernt. Sie können auch die **Anzahl der Aufteilungen** oder die Anzahl der für die Kreuzvalidierung verwendeten Modelle festlegen. Beachten Sie, dass die Modellbildung und die Kreuzvalidierung in früheren IBM SPSS Modeler-Versionen zwei separate Vorgänge darstellten. In der aktuellen Version ist kein gesonderter Modellbildungsschritt erforderlich. Modellbildung und Kreuzvalidierung werden gleichzeitig durchgeführt.

Modalwert. Für ein Training vom Typ **Einfach** werden die meisten C5.0-Parameter automatisch eingestellt. Ein Training vom Typ **Experten** bietet Ihnen eine direktere Kontrolle über die Trainingsparameter.

Optionen des Modus "Einfach"

Vorselektion. Standardmäßig versucht C5.0 den genauest möglichen Baum zu erstellen. In einigen Fällen kann dies zu einer Überanpassung führen, die eine schwache Leistung bedingt, sobald das Modell auf neue Daten angewendet wird. Wählen Sie die Option **Allgemeingültigkeit**, um Algorithmuseinstellungen zu verwenden, mit denen dieses Problem weniger häufig auftritt.

Hinweis: Es kann nicht garantiert werden, dass mit der Option **Allgemeingültigkeit** erstellte Modelle besser verallgemeinert werden können als andere Modelle. Wenn die Allgemeingültigkeit ein wichtiger Punkt ist, müssen Sie Ihr Modell immer gegen eine zurückgehaltene Teststichprobe validieren.

Erwartetes Rauschen (%). Geben Sie den erwarteten Anteil der im Trainingsset enthaltenen verrauschten oder fehlerhaften Daten an.

Optionen des Expertenmodus

Reduzierungsgrad. Legt fest, in welchem Umfang der Entscheidungsbaum bzw. das Regelset reduziert werden. Wenn Sie diesen Wert erhöhen, erhalten Sie einen kleineren und prägnanteren Baum. Geben Sie einen niedrigeren Wert an, wenn Sie einen genaueren Baum erhalten wollen. Diese Einstellung wirkt sich ausschließlich auf die lokale Reduzierung aus (siehe unten "Globale Reduzierung verwenden").

Minimale Anzahl der Datensätze pro untergeordneter Verzweigung. Anhand der Größe der Untergruppen kann die Anzahl der in jeder Verzweigung des Baums durchgeführten Aufteilungen eingeschränkt werden. Eine Verzweigung wird nur dann aufgeteilt, wenn zwei oder mehr der daraus entstehenden Unterverzweigungen mindestens so viele Datensätze des Trainingssets enthalten. Der Standardwert ist 2. Erhöhen Sie den Wert, um ein **Übertrainieren** mit verrauschten Daten zu verhindern.

Globale Reduzierung verwenden. Bäume werden in zwei Phasen reduziert: Zuerst in einer lokalen Reduzierungsphase, die untergeordnete Bäume prüft und Verzweigungen ausblendet, um die Genauigkeit des Modells zu steigern. Die zweite, globale Reduzierungsphase berücksichtigt den Baum als Ganzes und reduziert schwache untergeordnete Bäume. Die globale Reduzierung wird standardmäßig durchgeführt. Wenn Sie die globale Reduzierungsphase auslassen wollen, müssen Sie diese Option inaktivieren.

Vorselektion. Mit dieser Option untersucht C5.0 die Nützlichkeit der Prädiktoren, bevor die Modellbildung gestartet wird. Als irrelevant eingestufte Prädiktoren werden dann aus dem Modellbildungsvorgang ausgeschlossen. Diese Option ist oft bei Modellen mit vielen Prädiktorfeldern hilfreich und kann eine Überanpassung verhindern.

Hinweis: Die Geschwindigkeit für die Erstellung von C5.0-Modellen kann durch die Aktivierung der parallelen Verarbeitung verbessert werden.

Entscheidungsbaummodellnuggets

Modellnuggets vom Typ "Entscheidungsbaum" stellen die Baumstrukturen für die Vorhersage eines bestimmten Ausgabefelds dar, das von einem der Knoten für die Entscheidungsbaummodellierung ("C&R-Baum", "CHAID", "QUEST", "C5.0"). Die Baummodelle können direkt aus dem Baumerstellungsknoten oder indirekt aus dem interaktiven Tree Builder generiert werden. Weitere Informationen finden Sie im Thema „Interactive Tree Builder” auf Seite 83.

Scoring von Baummodellen

Wenn Sie einen Stream ausführen, der Baummodellnugget enthält, hängt das jeweils erzielte Ergebnis vom Baumtyp ab.

- Bei Klassifizierungsbäumen (kategoriales Ziel) werden zwei neue Felder, die den vorhergesagten Wert und die Konfidenz für die einzelnen Datensätze enthalten, zu den Daten hinzugefügt. Die Vorhersage

beruht auf der häufigsten Kategorie für den Endknoten, dem der Datensatz zugewiesen ist; wenn die Mehrzahl der Antworten in einem bestimmten Knoten *ja* lautet, ist die Vorhersage für alle diesem Knoten zugewiesenen Datensätzen "Ja".

- Bei Regressionsbäumen werden lediglich vorhergesagte Werte generiert, es werden keine Konfidenzen zugewiesen.
- Optional kann für Modelle vom Typ "CHAID, "QUEST", und "C&R-Baum" ein weiteres Feld hinzugefügt werden, das die ID für den Knoten angibt, dem die einzelnen Datensätze zugewiesen werden.

Die neuen Feldnamen werden durch Hinzufügen von Präfixen aus dem Modellnamen abgeleitet. Bei "C &R-Baum", "CHAID" und "QUEST" lautet das Präfix *$R-* für das Vorhersagefeld *$RC-* für das Konfidenzfeld und *$RI-* für das Knoten-ID-Feld. Bei C5.0-Bäumen lautet das Präfix *$C-* für das Vorhersagefeld und *$CC-* für das Konfidenzfeld. Bei mehreren Baummodellknoten enthalten die neuen Feldnamen Zahlen im *Präfix*, um sie gegebenenfalls voneinander zu unterscheiden. Beispiel: *$R1-* und *$RC1-* und *$R2-*.

Arbeiten mit Modellnuggets vom Typ "Entscheidungsbaum"

Sie können zum Modell gehörige Informationen auf verschiedene Weise speichern bzw. exportieren.

Hinweis: Viele dieser Optionen sind auch im Tree Builder-Fenster verfügbar.

Über den Tree Builder oder ein Baummodellnugget können Sie Folgendees ausführen:

- Generieren Sie einen Filter oder wählen Sie basierend auf dem aktuellen Baum einen Knoten aus. Weitere Informationen finden Sie im Thema „Generieren von Filter- und Auswahlknoten" auf Seite 93.
- Generieren Sie ein Regelsetnugget, das die Baumstruktur als Set von Regeln darstellt, das die Endverzweigungen des Baums definiert. Weitere Informationen finden Sie im Thema „Generieren eines Regelsets aus einem Entscheidungsbaum" auf Seite 93.
- Bei Baummodellnuggets können Sie außerdem das Modell im PMML-Format exportieren. Weitere Informationen finden Sie im Thema „Modellpalette" auf Seite 40. Wenn das Modell benutzerdefinierte Aufteilungen beinhaltet, werden diese Informationen nicht in der exportierten PMML gespeichert. (Die Aufteilung wird beibehalten, die Tatsache, dass sie benutzerdefiniert und nicht vom Algorithmus gewählt ist, jedoch nicht.)
- Generieren Sie ein Diagramm auf der Basis des ausgewählten Teils des aktuellen Baums. *Hinweis*: Dies funktioniert bei einem Nugget nur, wenn es an andere Knoten in einem Stream angehängt ist. Weitere Informationen finden Sie im Thema „Erzeugen von Diagrammen" auf Seite 114.
- Nur bei verstärkten C5.0-Modellen: Auswählen von **Einzelner Entscheidungsbaum (Erstellungsbereich)** oder **Einzelner Entscheidungsbaum (Palette der generierten Modelle)**, um ein neues einzelnes Regelset zu erstellen, die aus dem aktuell ausgewählten Regelset abgeleitet ist. Weitere Informationen finden Sie im Thema „Verbesserte C5.0-Modelle" auf Seite 113.

Hinweis: Obwohl der Regelerstellungsknoten durch den C&R-Baumknoten ersetzt wurde, funktionieren Entscheidungsbaumknoten in vorhandenen Streams, die ursprünglich über einen Regelerstellungsbaum erstellt wurden, weiterhin ordnungsgemäß.

Modellnuggets bei einzelnen Bäumen

Wenn Sie am Modellierungsknoten **Einzelnen Baum aufbauen** als Hauptziel auswählen, enthält das daraus resultierende Modellnugget die folgenden Registerkarten:

Tabelle 7. Registerkarten für Nugget bei einzelnen Bäumen

Registerkarte	Beschreibung	Weitere Informationen
Modell	Zeigt die Regeln an, die das Modell definieren.	Weitere Informationen finden Sie im Thema „Regeln für Entscheidungsbaummodelle" auf Seite 110.

Tabelle 7. Registerkarten für Nugget bei einzelnen Bäumen (Forts.)

Registerkarte	Beschreibung	Weitere Informationen
Viewer	Zeigt die Baumansicht des Modells an.	Weitere Informationen finden Sie im Thema „Viewer für Entscheidungsbaummodelle" auf Seite 112.
Zusammenfassung	Zeigt Informationen über die Felder, die Aufbaueinstellungen und die Modellschätzung an.	Weitere Informationen finden Sie im Thema „Modellnuggets - Übersicht/Informationen" auf Seite 43.
Einstellungen	Hier können Sie Optionen für Konfidenzen und für die SQL-Generierung während des Modellscorings angeben.	Weitere Informationen finden Sie im Thema „Einstellungen für Modellnuggets vom Typ "Entscheidungsbaum"/"Regelset"" auf Seite 112.
Anmerkung	Hier können Sie beschreibende Anmerkungen hinzufügen, einen benutzerdefinierten Namen angeben, QuickInfo-Text hinzufügen und Suchwörter für das Modell angeben.	

Regeln für Entscheidungsbaummodelle

Die Registerkarte "Modell" für ein Entscheidungsbaumnugget zeigt die Regeln an, die das Modell definieren. Optional können auch ein Diagramm für den Prädiktoreinfluss und ein drittes Fenster mit Informationen zu Verlauf, Häufigkeiten und Ersatztrennern angezeigt werden.

Hinweis: Wenn Sie die Option **Modell für sehr umfangreiche Datasets erstellen** auf der Registerkarte "Erstellungsoptionen" des CHAID-Knotens (Fenster "Ziel") auswählen, zeigt die Registerkarte "Modell" drei Regeldetails an.

Baumregeln

Im linken Fensterbereich wird eine Liste mit Bedingungen angezeigt, die die Partitionierung der durch den Algorithmus ermittelten Daten definieren. Im Wesentlichen handelt es sich hierbei um eine Reihe von Regeln, mit denen einzelne Datensätze untergeordneten Knoten basierend auf den Werten verschiedener Prädiktoren zugeordnet werden können.

Entscheidungsbäume funktionieren durch die rekursive Aufteilung der Daten basierend auf den Werten der Eingabefelder. Die Datenaufteilungen werden als **Verzweigungen** bezeichnet. Die erste Verzweigung (manchmal als **Stamm** bezeichnet) umfasst alle Datensätze. Der Stamm ist in Subsets bzw. **untergeordnete Verzweigungen** aufgeteilt, basierend auf dem Wert eines bestimmten Eingabefelds. Jede untergeordnete Verzweigung kann in weitere Verzweigungen aufgeteilt werden, die wiederum aufgeteilt werden können usw. Auf der niedrigsten Ebene des Baums befinden sich Verzweigungen ohne weitere Aufteilungen. Diese Verzweigungen heißen **Endverzweigungen** (oder **Blätter**).

Baumregeldetails

Der Regelbrowser zeigt die Eingabewerte, die jede Partition oder Verzweigung definieren, sowie eine Übersicht über die Werte der Ausgabefelder für die Datensätze dieser Aufteilung. Allgemeine Informationen zum Verwenden des Modellbrowsers finden Sie unter „Durchsuchen von Modellnuggets" auf Seite 42.

Bei Aufteilungen, die auf numerischen Feldern basieren, wird die Verzweigung durch eine Zeile ähnlich der folgenden dargestellt:

```
Feldname Beziehungswert [Übersicht]
```

Dabei stellt *Beziehung* eine numerische Beziehung dar. Eine Verzweigung, die z. B. durch Werte größer als 100 für das Feld *Einkünfte* definiert ist, würde wie folgt dargestellt:

```
Einkünfte > 100 [Übersicht].
```

Bei Aufteilungen, die auf symbolischen Feldern basieren, wird die Verzweigung durch eine Zeile ähnlich der folgenden dargestellt:

```
Feldname = Wert [Übersicht] oder Feldname in [Werte] [Übersicht]
```

Dabei stellt *Werte* die Feldwerte dar, die von der Verzweigung definiert werden. Eine Verzweigung, die beispielsweise Datensätze enthält, bei denen der Wert *Region Norden*, *Westen* oder *Süden* sein kann, würde wie folgt dargestellt werden:

```
Region in ["Norden" "Westen" "Süden"] [Übersicht]
```

Für Endverzweigungen wird ebenfalls eine Vorhersage ausgegeben, indem am Ende der Regelbedingung ein Pfeil und der vorhergesagte Wert hinzugefügt werden. Ein Blatt, das z. B. als *Einkünfte* > 100 definiert ist und den Wert *hoch* für das Ausgabefeld vorhersagt, würde wie folgt dargestellt:

```
Einkünfte > 100 [Modus: hoch] → hoch
```

Die **Übersicht** für die Verzweigung ist für symbolische und numerische Ausgabefelder unterschiedlich definiert. Für Bäume mit numerischen Ausgabefeldern stellt die Übersicht den **durchschnittlichen** Wert für die Verzweigung dar und die **Standardabweichung** der Verzweigung ist die Abweichung zwischen dem Durchschnitt für die Verzweigung und dem Durchschnitt der übergeordneten Verzweigung. Im Falle von Bäumen mit symbolischen Ausgabefeldern ist die Übersicht der **Modus** bzw. der häufigste Wert für Datensätze in der Verzweigung.

Um eine Verzweigung vollständig zu beschreiben, müssen Sie die Bedingung einschließen, die die Verzweigung definiert sowie die Bedingungen, die die Aufteilungen weiter oben im Baum definieren. Beispiel: Im Baum

```
Einkünfte > 100
  Region = "Norden"
  Region in ["Süden" "Osten" "Westen"]
    Einkünfte <= 200
```

wird die durch die zweite Zeile dargestellte Verzweigung definiert durch die Bedingungen *Einkünfte > 100* und *Region = "Norden"*.

Wenn Sie auf die Schaltfläche **Instanzen/Konfidenz anzeigen** in der Symbolleiste klicken, werden bei jeder Regel auch die Informationen darüber angezeigt, für wie viele Datensätze die Regel gilt (**Instanzen**), sowie der Anteil der Datensätze, für die die gesamte Regel wahr ist (**Konfidenz**).

Bedeutung des Prädiktors

Optional kann auf der Registerkarte "Modell" auch ein Diagramm, das den relative Einfluss der einzelnen Prädiktoren für die Schätzung des Modells angibt, angezeigt werden. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und die Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass dieses Diagramm nur verfügbar ist, wenn vor dem Generieren des Modells **Prädiktoreinfluss berechnen** auf der Registerkarte "Analysieren" ausgewählt wurde. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Zusätzliche Modellinformationen

Wenn Sie in der Symbolleiste auf **Fenster mit weiteren Informationen** klicken, wird unten im Fenster ein Teilfenster mit detaillierten Informationen über die ausgewählte Regel angezeigt. Das Informationsfenster enthält drei Registerkarten.

Verlauf. Auf dieser Registerkarte werden die Aufteilungsbedingungen vom Stammknoten abwärts zum ausgewählten Knoten verfolgt. Auf diese Weise erhalten Sie eine Liste mit Bedingungen, die festlegen, wann ein Datensatz dem ausgewählten Knoten zugewiesen wird. Datensätze, für die alle Bedingungen wahr sind, werden diesem Knoten zugewiesen.

Häufigkeiten. Für Modelle mit symbolischen Zielfeldern zeigt diese Registerkarte für jeden möglichen Zielwert die Anzahl der Datensätze an, die diesem Knoten (in den Trainingsdaten) zugewiesen sind und diesen Zielwert aufweisen. Die Zahl für die Häufigkeit, ausgedrückt als Prozentwert (bis maximal drei Dezimalstellen) wird ebenfalls angezeigt. Für Modelle mit numerischen Zielwerten bleibt diese Registerkarte leer.

Ersatzfelder. Wo zutreffend, werden für den ausgewählten Knoten alle Ersatzfelder für das primäre Aufteilungsfeld angezeigt. Ersatzfelder sind alternative Felder, die verwendet werden, wenn der primäre Prädiktorwert für einen bestimmten Datensatz fehlt. Die maximale Anzahl an Ersatzfeldern, die für eine bestimmte Aufteilung erlaubt ist, ist im Baumerstellungsknoten angegeben, die tatsächliche Anzahl richtet sich jedoch nach den Trainingsdaten. Im Allgemeinen gilt: Je mehr fehlende Daten, desto mehr Ersatzfelder werden wahrscheinlich verwendet. Für andere Entscheidungsbaummodelle bleibt diese Registerkarte leer.

Hinweis: Damit Ersatzfelder im Modell berücksichtigt werden, müssen sie während der Trainingsphase ermittelt werden. Wenn die Trainingsstichprobe keine fehlenden Werte enthält, werden keine Ersatzfelder ermittelt. Alle Datensätze mit fehlenden Werten, die während des Testens oder der Bewertung gefunden werden, fallen automatisch in den untergeordneten Knoten mit der größten Anzahl an Datensätzen. Wenn während des Testens oder Bewertens fehlende Werte erwartet werden, können Sie sicher sein, dass auch in den Trainingsstichproben Werte fehlen. Für CHAID-Bäume sind keine Ersatzfelder verfügbar.

Viewer für Entscheidungsbaummodelle

Die Registerkarte "Viewer" eines Modellnuggets vom Typ "Entscheidungsbaum" ähnelt der Anzeige im Tree Builder. Der Hauptunterschied besteht darin, dass der Baum beim Modellnugget nicht erweitert oder bearbeitet werden kann. Andere Optionen für die Betrachtung und Anpassung der Anzeige sind zwischen den beiden Komponenten ähnlich. Weitere Informationen finden Sie im Thema „Anpassen der Baumansicht” auf Seite 86.

Hinweis: Die Registerkarte "Viewer" wird für CHAID-Modellnuggets nicht angezeigt, die beim Auswählen der Option **Modell für sehr umfangreiche Datasets erstellen** auf der Registerkarte "Erstellungsoptionen" (Fenster "Ziel") erstellt werden.

Bei der Anzeige von Aufteilungsregeln auf der Registerkarte "Viewer" bedeuten eckige Klammern, dass der angegebene Wert im Bereich enthalten ist, während runde Klammern anzeigen, dass der Wert aus dem Bereich ausgeschlossen ist. Der Audruck (23,37] bedeutet somit von 23 (ausgeschlossen) bis einschließlich 37, also von etwas über 23 bis 37. Auf der Registerkarte "Modell" wird dieselbe Bedingung wie folgt angezeigt:

```
Age > 23 and Age <= 37
```

Einstellungen für Modellnuggets vom Typ "Entscheidungsbaum"/"Regelset"

Auf der Registerkarte "Einstellungen" für ein Modellnugget vom Typ "Entscheidungsbaum" oder "Regelset" können Sie während des Modellscorings Optionen für Konfidenzen und zur SQL-Generierung angeben. Diese Registerkarte ist erst verfügbar, nachdem das Modellnugget einem Stream hinzugefügt wurde.

Konfidenzen berechnen. Wählen Sie diese Option aus, um Konfidenzen in die Scoring-Operationen aufzunehmen. Beim Scoren von Modellen in der Datenbank können Sie durch Ausschließen der Konfidenzen eine effizientere SQL erzeugen. Bei Regressionsbäumen werden keine Konfidenzen zugewiesen.

Hinweis: Wenn Sie die Option **Modell für sehr umfangreiche Datasets erstellen** auf der Registerkarte "Erstellungsoptionen" des CHAID-Modells (Fenster "Methoden") auswählen, ist dieses Kontrollkästchen nur in den Modellnuggets für die kategorialen Ziele "nominal" oder "Flag" verfügbar.

Raw-Propensity-Scores berechnen. Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Diese Werte werden zusätzlich zu den anderen Vorhersage- und Konfidenzwerten angegeben, die gegebenenfalls während des Scorings erstellt werden.

Hinweis: Wenn Sie die Option **Modell für sehr umfangreiche Datasets erstellen** auf der Registerkarte "Erstellungsoptionen" des CHAID-Modells (Fenster "Methoden") auswählen, ist dieses Kontrollkästchen nur in Modellnuggets mit einem kategorialen Ziel vom Typ "Flag" verfügbar.

Adjusted-Propensity-Scores berechnen. Raw-Propensity-Scores beruhen ausschließlich auf den Trainingsdaten und sind aufgrund der Neigung vieler Modelle zur übermäßigen Anpassung an die Trainingsdaten möglicherweise zu optimistisch. Bei Adjusted Propensitys wird versucht, dies durch Evaluation der Modellleistung anhand einer Test- bzw. Validierungspartition zu kompensieren. Bei dieser Option muss im Stream ein Partitionsfeld definiert sein und die Adjusted-Propensity-Scores müssen im Modellierungsknoten aktiviert werden, bevor das Modell generiert wird.

Hinweis: Adjusted-Propensity-Scores sind für aufgewertete Baum- und Regelsetmodelle nicht verfügbar. Weitere Informationen finden Sie im Thema „Verbesserte C5.0-Modelle”.

Regel-ID. Bei Modellen vom Typ "CHAID", "QUEST" und "C&R-Baum" fügt diese Option ein Feld in der Scoring-Ausgabe hinzu, das die ID für den Endknoten angibt, dem der jeweilige Datensatz zugewiesen ist.

Hinweis: Wenn diese Option ausgewählt ist, ist die SQL-Generierung nicht verfügbar.

SQL für dieses Modell generieren. Bei Verwendung von Daten aus einer Datenbank kann SQL-Code per Pushback zur Ausführung an die Datenbank zurückübertragen werden. Dadurch kann bei vielen Operationen eine bessere Leistung erzielt werden.

Wählen Sie eine der folgenden Optionen aus, um festzulegen, wie der SQL-Code generiert wird.

- **Standard: Mithilfe des Server-Scoring-Adapters (falls installiert), ansonsten bei der Verarbeitung scoren.** Bei Verbindung mit einer Datenbank mit installiertem Scoring-Adapter wird der SQL-Code mit dem Scoring-Adapter generiert. Andernfalls wird der SQL-Code prozessintern generiert SPSS Modeler.
- **Ohne Unterstützung für fehlende Werte generieren.** Wählen Sie diese Option, um die SQL-Generierung ohne den Aufwand für den Umgang mit fehlenden Werten zu aktivieren. Bei dieser Option wird die Vorhersage einfach auf null gesetzt (`$null$`), wenn beim Scoring eines Falles ein fehlender Wert gefunden wird.

  Hinweis: Diese Option ist für CHAID-Modelle nicht verfügbar. Bei anderen Modelltypen ist sie nur für Entscheidungsbäume (nicht für Regelsets) verfügbar.
- **Mit Unterstützung für fehlende Werte generieren.** Bei Modellen vom Typ "CHAID", "QUEST" und "C&R-Baum" können Sie die SQL-Generierung mit vollständiger Unterstützung für fehlende Werte aktivieren. Dabei wird SQL so generiert, dass die fehlenden Werte als im Modell angegeben behandelt werden. Bei C&R-Bäumen werden beispielsweise Ersatzregeln und ein Rückgriff auf das größte untergeordnete Element verwendet.

  Hinweis: Bei C5.0-Modellen ist diese Option nur für Regelsets verfügbar (nicht für Entscheidungsbäume).

Verbesserte C5.0-Modelle

Hinweis: Diese Funktion ist in SPSS Modeler Professional und SPSS Modeler Premium verfügbar.

Beim Erstellen eines verstärkten C5.0-Modells (entweder Regelset oder Entscheidungsbaum) wird eigentlich ein Set aus verwandten Modellen erstellt. Der Modellregelbrowser für ein verstärktes C5.0-Modell zeigt die Liste der Modelle auf der obersten Ebene der Hierarchie an. Außerdem wird die geschätzte Genauigkeit der einzelnen Modelle und die Gesamtgenauigkeit des Ensembles der verstärkten Modelle an-

gezeigt. Um die Regeln oder Aufteilungen für ein bestimmtes Modell zu untersuchen, wählen Sie das betreffende Modell aus und erweitern Sie es, wie bei einer Regel oder Verzweigung in einem einzelnen Modell.

Außerdem können Sie ein bestimmtes Modell aus dem Set der verstärkten Modelle extrahieren und ein Modellnugget vom Typ "Regelset" erstellen, das nur dieses eine Modell enthält. Um ein neues Regelset aus einem verstärkten C5.0-Modell zu erstellen, wählen Sie das gewünschte Regelset bzw. den gewünschten Baum aus und wählen Sie im Menü "Generieren" entweder die Option **Einzelner Entscheidungsbaum (Palette der generierten Modelle)** oder **Einzelner Entscheidungsbaum (Erstellungsbereich)**.

Erzeugen von Diagrammen

Die Baumknoten bieten eine Menge an Informationen, jedoch sind diese nicht unbedingt immer in einem leicht zugänglichen Format für Fachanwender. Zur Bereitstellung der Daten auf eine Art, die problemlos in Geschäftsberichte, Präsentationen u.s.w. integriert werden kann, können aus ausgewählten Daten Diagramme erstellt werden. Beispielsweise können Sie von der Registerkarte "Modell" oder "Viewer" eines Modellnuggets oder der Registerkarte "Viewer" eines interaktiven Baums ein Diagramm für den ausgewählten Teil eines Baums generieren und damit nur ein Diagramm für die Fälle im ausgewählten Baum oder Verzweigungsknoten erstellen.

Hinweis: Sie können ein Diagramm nur aus einem Nugget erstellen, wenn es an andere Knoten in einem Stream angehängt ist.

Generieren eines Diagramms

Wählen Sie als ersten Schritt die Informationen aus, die im Diagramm gezeigt werden sollen:

- Erweitern Sie auf der Registerkarte "Modell" eines Nuggets die Liste der Bedingungen und Regeln im rechten Fensterbereich und wählen Sie die gewünschte aus.
- Erweitern Sie auf der Registerkarte "Viewer" eines Nuggets die Liste der Verzweigungen und wählen Sie den gewünschten Knoten aus.
- Erweitern Sie auf der Registerkarte "Viewer" eines interaktiven Baums die Liste der Verzweigungen und wählen Sie den gewünschten Knoten aus.

  Hinweis: Sie können den ersten Knoten in keiner Registerkarte vom Typ "Viewer" auswählen.

Unabhängig von den anzuzeigenden Daten erstellen Sie ein Diagramm immer auf die gleiche Weise:

1. Wählen Sie aus dem Menü "Generieren" die Option **Diagramm (von Auswahl)**. Oder klicken Sie alternativ auf der Registerkarte "Viewer" auf die Schaltfläche **Diagramm (von Auswahl)** in der unteren linken Ecke. Die Registerkarte "Einfach" der Diagrammtafel wird angezeigt.

   Hinweis: Wenn Sie die Diagrammtafel auf diese Art anzeigen, sind nur die Registerkarten "Basis" und "Details" verfügbar.
2. Mithilfe der Einstellungen auf den Registerkarten "Basis" oder "Details" können Sie die Details angeben, die auf dem Diagramm angezeigt werden sollen.
3. Klicken Sie auf "OK", um das Diagram zu erstellen.

Die Überschrift des Diagramms identifiziert die Knoten oder Regeln, die im Diagramm berücksichtigt werden.

Modellnuggets für Boosting, Bagging und sehr große Datasets

Wenn Sie am Modellierungsknoten **Modellgenauigkeit erhöhen (Boosting)**, **Modellstabilität steigern (Bagging)** oder **Modell für sehr große Datasets erstellen** als Hauptziel auswählen, erstellt IBM SPSS Modeler ein Ensemble aus mehreren Modellen. Weitere Informationen finden Sie im Thema „Modelle für Ensembles” auf Seite 45.

Das daraus resultierende Modellnugget enthält die folgenden Registerkarten. Die Registerkarte "Modell" bietet mehrere unterschiedliche Modellansichten.

Tabelle 8. In Modellnugget verfügbare Registerkarten

Registerkarte	Ansicht	Beschreibung	Weitere Informationen
Modell	Modellübersicht	Zeigt eine Übersicht der Ensemble-Qualität und (mit Ausnahme von verbesserten Modellen und stetigen Zielen) -Vielfältigkeit an, ein Maß dafür, wie stark die Vorhersagen unter den verschiedenen Modellen abweichen.	Weitere Informationen finden Sie im Thema „Modellzusammenfassung" auf Seite 45.
	Bedeutung des Prädiktors	Zeigt eine Tabelle an, die die relative Wichtigkeit der einzelnen Prädiktoren (Eingabefeld) für die Schätzung des Modells angibt.	Weitere Informationen finden Sie im Thema „Prädiktoreinfluss" auf Seite 46.
	Prädiktorhäufigkeit	Zeigt eine Tabelle an, die die relative Häufigkeit angibt, mit der jeder Prädiktor in dem Modellset verwendet wird.	Weitere Informationen finden Sie im Thema „Prädiktorhäufigkeit" auf Seite 46.
	Komponentenmodell-genauigkeit	Erstellt ein Diagramm der Vorhersagegenauigkeit für jedes Modell im Ensemble.	
	Komponentenmodelldetails	Zeigt Informationen über jedes Modell im Ensemble an.	Weitere Informationen finden Sie im Thema „Komponentenmodelldetails" auf Seite 46.
	Information	Zeigt Informationen über die Felder, die Aufbaueinstellungen und die Modellschätzung an.	Weitere Informationen finden Sie im Thema „Modellnuggets - Übersicht/Informationen" auf Seite 43.
Einstellungen		Hier können Sie Konfidenzen in die Scoring-Operationen aufnehmen.	Weitere Informationen finden Sie im Thema „Einstellungen für Modellnuggets vom Typ "Entscheidungsbaum"/"Regelset"" auf Seite 112.
Anmerkung		Hier können Sie beschreibende Anmerkungen hinzufügen, einen benutzerdefinierten Namen angeben, QuickInfo-Text hinzufügen und Suchwörter für das Modell angeben.	

Regelsetmodellnuggets

Ein Regelsetmodellnugget stellt die Regeln zum Vorhersagen eines bestimmten Ausgabefelds dar, das von dem Modellierungsknoten für Assoziationsregeln (Apriori) oder von einem der drei Baumerstellungsknoten (C&R-Baum, CHAID, QUEST oder C5.0) erkannt wurde. Bei Assoziationsregeln muss das Regelset aus einem Nugget für nicht verfeinerte Regeln generiert werden. Bei Bäumen können Regelsets über den Tree Builder, aus einem C.50-Modellerstellungsknoten oder aus einem beliebigen Baummmodellnugget generiert werden. Im Gegensatz zu Nuggets vom Typ "Nicht verfeinerte Regel" können Nuggets vom Typ "Regelset" in Streams platziert werden, um Vorhersagen zu generieren.

Bei der Ausführung eines Streams, der einen Regelsetnugget enthält, werden zwei neue Felder, die den vorhergesagten Wert und die Konfidenz für die einzelnen Datensätze enthalten, zum Stream hinzugefügt. Die neuen Feldnamen werden durch Hinzufügen von Präfixen aus dem Modellnamen abgeleitet. Bei As-

soziationsregelsets lautet das Präfix *$A-* für das Vorhersagefeld und *$AC-* für das Konfidenzfeld. Bei C5.0-Regelsets lautet das Präfix *$C-* für das Vorhersagefeld und *$CC-* für das Konfidenzfeld. Bei C&R-Baumregelsets lautet das Präfix *$R-* für das Vorhersagefeld und *$RC-* für das Konfidenzfeld. In einem Stream mit mehreren Regelsetnuggets in einer Reihe, die dieselben Ausgabefelder vorhersagen, enthalten die neuen Feldnamen Zahlen im *Präfix*, damit sie auseinander gehalten werden können. Beim ersten Assoziations-Regelsetknoten im Stream werden die üblichen Namen verwendet, beim zweiten Knoten Namen, die mit *$A1-* und *$AC1-* beginnen, beim dritten Knoten Namen mit *$A2-* und *$AC2-* usw.

Anwendung der Regeln. Aus Assoziationsregeln erstellte Regelsets sind anders als andere Modellnuggets, da für jeden Datensatz mehrere Vorhersagen generiert werden können und diese Vorhersagen nicht unbedingt alle übereinstimmen. Es gibt zwei Methoden zum Generieren von Vorhersagen aus Regelsets:

Hinweis: Regelsets, die aus Entscheidungsbäumen erstellt wurden, geben unabhängig von der verwendeten Methode dieselben Ergebnisse zurück, da die von einem Entscheidungsbaum abgeleiteten Regeln sich gegenseitig ausschließen.

- **Voting.** Bei dieser Methode wird versucht, die Vorhersagen aller Regeln, die für den Datensatz gelten, zu kombinieren. Bei jedem Datensatz werden alle Regeln untersucht, und jede Regel, die für den Datensatz gilt, wird verwendet, um eine Vorhersage und eine zugehörige Konfidenz zu generieren. Für jeden Ausgabewert werden die Zahlen für die Summe der Konfidenz berechnet und der Wert mit der größten Konfidenzsumme wird als endgültige Vorhersage ausgewählt. Die Konfidenz für die endgültige Vorhersage ist die Konfidenzsumme für diesen Wert dividiert durch die Anzahl der Regeln, die für diesen Datensatz ausgelöst wurden.
- **Erster Treffer.** Diese Methode testet einfach die Regeln in der Reihenfolge und die erste Regel, die für den Datensatz gilt, wird zur Generierung der Vorhersage verwendet.

Die verwendete Methode kann in den Streamoptionen festgelegt werden.

Generieren von Knoten. Mit dem Menü "Generieren" können Sie neue Knoten basierend auf dem Regelset erstellen.

- **Filterknoten.** Erstellt einen neuen Filterknoten zum Filtern von Feldern, die nicht von den Regeln im Regelset verwendet werden.
- **Auswahlknoten.** Erstellt einen neuen Auswahlknoten zur Auswahl von Datensätzen, für die die ausgewählte Regel gilt. Der generierte Knoten wählt Datensätze aus, für die alle Antezedenzien der Regel wahr sind. Für diese Option muss eine Regel ausgewählt sein.
- **Regelverfolgungsknoten.** Erstellt einen neuen Superknoten und berechnet ein Feld, das angibt, welche Regel zur Erstellung der Vorhersage für die einzelnen Datensätze verwendet wurde. Bei der Auswertung eines Regelsets mithilfe der Methode "Erster Treffer" ist dies einfach ein Symbol, das die erste Regel angibt, die ausgelöst werden würde. Bei der Auswertung des Regelsets mithilfe der Methode "Voting" ist dies eine komplexere Zeichenfolge, die die Eingabe in den Voting-Mechanismus anzeigt.
- **Einzelner Entscheidungsbaum (Erstellungsbereich)/Einzelner Entscheidungsbaum (Palette der generierten Modelle).** Erstellt ein einzelnes Regelset, das aus der aktuell ausgewählten Regel abgeleitet wurde. Nur für **verstärkte** C5.0-Modelle verfügbar. Weitere Informationen finden Sie im Thema „Verbesserte C5.0-Modelle” auf Seite 113.
- **Modell in Palette.** Gibt das Modell an die Modellpalette zurück. Das ist nützlich, wenn Sie von einem Kollegen einen Stream, der das Modell enthält, jedoch nicht das Modell selbst erhalten.

Hinweis: Die Registerkarten "Einstellungen" und "Übersicht" im Regelsetnugget stimmen mit den für Entscheidungsbaummodelle verwendeten überein.

Regelset - Registerkarte "Modell"

Auf der Registerkarte "Modell" für ein Regelsetnugget wird eine Liste der Regeln angezeigt, die der Algorithmus aus den Daten extrahiert hat.

Die Regeln werden nach Sukzedens (vorhergesagte Kategorie) aufgeschlüsselt und in folgendem Format angezeigt:

```
if Antezedent_1
and Antezedent_2
...
and Antezedent_n
then vorhergesagter_Wert
```

Dabei handelt es sich bei Sukzedens und Antezedens_1 bis Antezedens_n jeweils um Bedingungen. Die Regel wird wie folgt interpretiert: "Bei Datensätzen, bei denen Antezedens_1 bis Antezedens_n alle wahr sind, ist wahrscheinlich auch Sukzedens wahr." Wenn Sie auf die Schalfläche **Instanzen/Konfidenz anzeigen** in der Symbolleiste klicken, werden bei jeder Regel auch die Informationen darüber angezeigt, für wie viele Datensätze die Regel gilt, also für wie viele Datensätze die Antezedenzien wahr sind (**Instanzen**), sowie der Anteil der Datensätze, für die die gesamte Regel wahr ist (**Konfidenz**).

Beachten Sie, dass die Berechnungsmethode für die Konfidenz bei C5.0-Regelsets ein wenig abweicht. Bei C5.0 wird folgende Formel zur Berechnung der Konfidenz einer Regel verwendet:

```
(1 + Anzahl der Datensätze, bei denen die Regel richtig ist)
/
(2 + Anzahl der Datensätze, bei denen die Antezedenzien der Regel wahr sind)
```

Diese Berechnung der Konfidenzschätzung wird für den Prozess der Verallgemeinerung von Regeln aus einem Entscheidungsbaum (das Verfahren, mit dem bei C5.0 Regelsets erstellt werden) angepasst.

Projekte aus AnswerTree 3.0 importieren

IBM SPSS Modeler kann in AnswerTree 3.0 oder 3.1 gespeicherte Modelle über das Standarddialogfeld "Datei" > "Öffnen" importieren:

1. Wählen Sie in den IBM SPSS Modeler-Menüs folgende Befehlsfolge:

   Datei > **Stream öffnen**
2. Wählen Sie in der Dropdown-Liste "Dateityp" die Option **AT Project Files (*.atp, *.ats)**.

   Jedes importierte Projekt wird in einen IBM SPSS Modeler-Stream konvertiert und enthält folgende Knoten:
   - einen Quellenknoten, der die verwendete Datenquelle definiert (z. B. eine IBM SPSS Statistics-Datendatei oder eine Datenbankquelle).
   - Für jedem im Projekt vorhandenen Baum (es kann mehrere enthalten) wird ein Typknoten erstellt, der die Eigenschaften aller Felder (Variablen) definiert, mit Typ, Rolle (Eingabe oder Prädiktorfeld <> Ausgabe oder vorhergesagtes Feld), fehlenden Werten und anderen Optionen.
   - Für jeden im Projekt enthaltenen Baum wird ein Partitionsknoten erstellt, der die Daten für eine Trainings- und eine Teststichprobe partitioniert. Ein Baumerstellungsknoten wird erstellt, der die Parameter definiert, mit denen der Baum generiert wird (Knoten vom Typ "C&R-Baum", "QUEST" oder "CHAID").
3. Führen Sie den Stream aus, um den oder die generierten Bäume anzuzeigen.

Kommentare

- In IBM SPSS Modeler generierte Entscheidungsbäume können nicht in AnswerTree exportiert werden. Der Import von AnswerTree in IBM SPSS Modeler ist nicht umkehrbar.
- In AnswerTree definierte Profite gehen beim Importieren des Projekts in IBM SPSS Modeler verloren.

Kapitel 7. Bayes-Netzmodelle

Bayes-Netzknoten

Mithilfe des Bayes-Netzknotens können Sie ein Wahrscheinlichkeitsmodell erstellen, indem Sie beobachtete und aufgezeichnete Hinweise mit "gesundem Menschenverstand" kombinieren, um die Wahrscheinlichkeit des Vorkommens unter Verwendung scheinbar nicht miteinander verknüpfter Attribute zu ermitteln. Der Knoten ist speziell für Netze vom Typ "Tree Augmented Naïve Bayes" (TAN) und "Markov-Decke" gedacht, die in erster Linie zur Klassifizierung verwendet werden.

Bayes-Netze dienen zur Erstellung von Vorhersagen in vielen verschiedenen Situationen. Hier einige Beispiele:

- Ermitteln von Antragstellern für Kredite mit geringem Risiko der Zahlungsunfähigkeit.
- Abschätzung, zu welchem Zeitpunkt für Geräte Wartungsarbeiten, Ersatzteile oder ein Austausch erforderlich ist, basierend auf Sensoreingaben und bestehenden Datensätzen.
- Lösen von Kundenproblemen mithilfe von Online-Tools zur Fehlerbehebung.
- Diagnostizierung und Fehlerbehebung bei Mobiltelefonnetzen in Echtzeit.
- Abschätzen des potenziellen Risikos und Nutzens von Forschungs- und Entwicklungsprojekten mit dem Ziel, die Ressourcen auf die aussichtsreichsten Möglichkeiten zu konzentrieren.

Ein Bayes-Netz ist ein grafisches Modell, das Variablen (häufig als **Knoten** bezeichnet) in einem Dataset und die probabilistischen bzw. bedingten Unabhängigkeiten zwischen den Variablen anzeigt. Durch Bayes-Netze können kausale Beziehungen zwischen Knoten angezeigt werden; die Verbindungen in den Netzen (auch als **arcs** (Bögen) bezeichnet) stehen jedoch nicht unbedingt für ein direktes Verhältnis von Ursache und Wirkung. Mithilfe eines Bayes-Netzes kann beispielsweise die Wahrscheinlichkeit berechnet werden, dass ein Patient unter einer bestimmten Krankheit leidet. Diese Berechnung beruht auf dem Vorliegen bzw. Fehlen bestimmter Symptome sowie anderen relevanten Daten und setzt voraus, dass sich die probabilistischen Unabhängigkeiten zwischen Symptomen und Krankheiten, die im Diagramm angezeigt werden, als wahr erweisen. Netze sind bei fehlenden Informationen sehr robust und führen zu der bestmöglichen Vorhersage unter Nutzung aller vorhandenen Informationen.

Ein typisches, einfaches Beispiel eines Bayes-Netzes wurde von Lauritzen und Spiegelhalter (1988) erstellt. Es wird häufig als das "Asien-Modell" bezeichnet und ist eine vereinfachte Version eines Netzes, das zur Diagnostizierung neuer Patienten eines Arztes verwendet werden kann. Die Richtung der Verbindungen entspricht in etwa der Kausalität. Jeder Knoten steht für eine Facette, die mit dem Zustand des Patienten in Beziehung stehen könnte. So zeigt "Smoking" (Rauchen) an, dass sie Raucher sind und "VisitAsia" (BesuchAsien) zeigt an, ob sie sich in letzter Zeit in Asien aufgehalten haben. Wahrscheinlichkeitsbeziehungen werden durch die Verbindungen zwischen Knoten angezeigt. So erhöht das Rauchen beispielsweise sowohl die Wahrscheinlichkeit, dass ein Patient Bronchitis entwickelt, als auch die Wahrscheinlichkeit, dass er Lungenkrebs entwickelt, während das Alter nur mit der Möglichkeit der Entwicklung von Lungenkrebs verknüpft zu sein scheint. Ebenso gilt, dass Anomalien auf einem Röntgenbild der Lungen entweder durch Tuberkulose oder Lungenkrebs verursacht werden können, während die Wahrscheinlichkeit, dass ein Patient an Atemnot (Dyspnoe) leidet, erhöht ist, wenn der Patient auch an Bronchitis oder Lungenkrebs leidet.

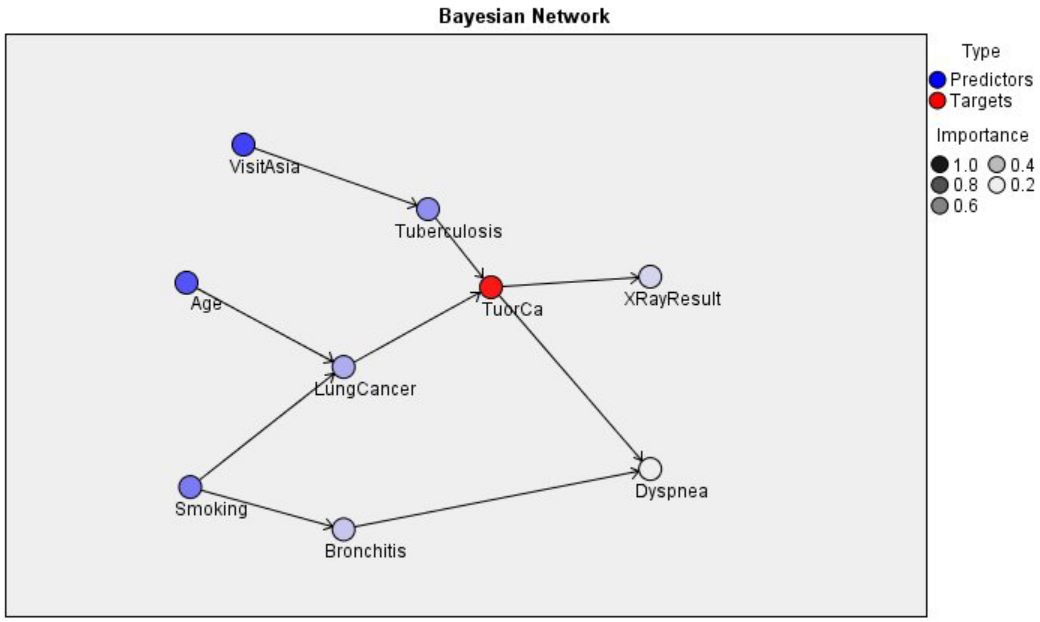


Abbildung 29. Asien-Netzbeispiel von Lauritzen und Spiegelhalter

Es gibt mehrere Gründe, die für die Verwendung eines Bayes-Netzes sprechen können:

- Es bietet Einblicke in Kausalbeziehungen. Ausgehend davon trägt es zum Verständnis eines Problembereichs bei und ermöglicht es, die Folgen von Eingriffen vorauszusagen.
- Das Netz bietet einen effizienten Ansatz zur Vermeidung einer Überanpassung an die Daten.
- Es bietet eine deutliche Visualisierung der beteiligten Beziehungen.

Voraussetzungen. Die Zielfelder müssen kategorial sein und können das Messniveau *Nominal*, *Ordinal* oder *Flag* aufweisen. Als Eingaben kommen Felder jedes Typs infrage. Stetige Eingabefelder (numerischer Bereich) werden automatisch klassiert. Bei verzerrten Verteilungen erzielen Sie jedoch möglicherweise bessere Ergebnisse, wenn Sie die Felder vor dem Bayes-Netzknoten manuell mithilfe eines Klassierknotens klassieren. Verwenden Sie beispielsweise "Optimales Klassieren", wenn das **Supervisorfeld** mit dem Feld **Ziel** des Bayes-Netzknotens übereinstimmt.

Beispiel. Ein Analyst einer Bank möchte in der Lage sein vorherzusagen, welche Kunden bzw. potenzielle Kunden mit hoher Wahrscheinlichkeit mit ihren Kreditrückzahlungen in Verzug geraten. Mithilfe eines Bayes-Netzmodells können Sie die Merkmale von Kunden ermitteln, die mit hoher Wahrscheinlichkeit zahlungsunfähig werden, und mehrere Modellarten erstellen, um herauszufinden, welches Modell die potenziell zahlungsunfähigen Personen am besten vorhersagt.

Beispiel. Ein Telekommunikationsbetreiber möchte die Anzahl der Kunden, die abwandern, verringern und das Modell monatlich mit den Daten des jeweiligen Vormonats aktualisieren. Mithilfe eines Bayes-Netzmodells können Sie die Merkmale von Kunden ermitteln, die mit hoher Wahrscheinlichkeit abwandern werden, und das Modell weiterhin jeden Monat mit neuen Daten trainieren.

Modelloptionen für Bayes-Netzknoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Modell für jede Aufteilung erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Partition. In diesem Feld können Sie ein Feld angeben, mit dem die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphasen der Modellerstellung aufgeteilt werden. Indem Sie das Modell mit einer Stichprobe erstellen und dann mit einer anderen Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln. Wenn mehrere Partitionsfelder mithilfe von Typ- oder Partitionsknoten erstellt wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.) Beachten Sie außerdem, dass die Partitionierung auch auf der Registerkarte "Modelloptionen" des Knotens aktiviert sein muss, wenn die ausgewählte Partitionierung in Ihrer Analyse angewendet werden soll. (Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung ohne Änderung der Feldeinstellungen inaktiviert werden.)

Aufteilungen. Wählen Sie für Aufteilungsmodelle das Aufteilungsfeld bzw. die Aufteilungsfelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Aufteilung* festlegen. Sie können nur Felder mit einem Messniveau **Flag**, **Nominal**, **Ordinal** oder **Stetig** als Aufteilungsfelder festlegen. Als Aufteilungsfelder gewählte Felder können nicht als Ziel-, Prädiktor-, Partitions-, Häufigkeits oder Gewichtungsfelder verwendet werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Training des bestehenden Modells fortsetzen. Wenn Sie diese Option auswählen, werden die auf der Registerkarte "Modell" des Modellnuggets angezeigten Ergebnisse bei jeder Modellausführung erneut generiert und aktualisiert. Diese Vorgehensweise wird beispielsweise verwendet, wenn eine neue oder aktualisierte Datenquelle zu einem bestehenden Modell hinzugefügt wurde.

Hinweis: Damit ist lediglich eine Aktualisierung des vorhandenen Netzes möglich. Weder Knoten noch Verbindungen können hinzugefügt bzw. entfernt werden. Bei jedem erneuten Trainieren des Modells behält das Netz dieselbe Form bei. Es ändern sich lediglich die bedingten Wahrscheinlichkeiten und der Prädiktoreinfluss. Wenn die neuen Daten im Großen und Ganzen den alten Daten ähneln, ist dies nicht von Bedeutung, da davon auszugehen ist, dass dieselben Elemente von Bedeutung sind. Wenn Sie jedoch überprüfen bzw. aktualisieren möchten, *was* von Bedeutung ist (im Gegensatz dazu, wie bedeutsam es ist), müssen Sie ein neues Modell, also ein neues Netz, erstellen.

Strukturtyp. Dient zur Auswahl der beim Erstellen des Bayes-Netzes zu verwendenden Struktur:

- **TAN.** Das TAN-Modell (Tree Augmented Naïve Bayes) erstellt ein einfaches Bayes-Netzmodell, das eine Verbesserung gegenüber dem standardmäßigen Naïve-Bayes-Modell darstellt. Dies liegt daran, dass dabei jeder Prädiktor neben der Zielvariablen auch von einem anderen Prädiktor abhängig sein kann, wodurch die Klassifizierungsgenauigkeit erhöht wird.
- **Markov-Decke.** Dient zum Auswählen des Sets von Knoten im Dataset, das die übergeordneten Elemente der Zielvariable, die zugehörigen untergeordneten Elemente und die Elemente enthält, die den untergeordneten Elementen übergeordnet sind. Im Wesentlichen identifiziert eine Markov-Decke alle Variablen im Netz, die zur Vorhersage der Zielvariablen erforderlich sind. Die Methode zur Erstellung eines Netzes gilt als genauer. Bei großen Datasets kann sie jedoch aufgrund der großen Anzahl an Variablen den Nachteil einer längeren Verarbeitungsdauer mit sich bringen. Um den Verarbeitungsaufwand zu verringern, können Sie mithilfe der Optionen **Merkmalauswahl** auf der Registerkarte "Experten" die Variablen auswählen, die in einer signifikanten Beziehung zur Zielvariablen stehen.

Vorbereitenden Merkmalauswahlschritt einschließen. Durch Auswahl dieses Kontrollkästchens können Sie die Optionen zur **Merkmalauswahl** auf der Registerkarte "Experten" nutzen.

Parameterlernmethode. Die Parameter von Bayes-Netzen beziehen sich auf die bedingten Wahrscheinlichkeiten für die einzelnen Knoten in Abhängigkeit von den Werten seiner jeweiligen übergeordneten

Elemente. Es gibt zwei verschiedene Auswahlmöglichkeiten, mit denen Sie die Aufgabe zur Schätzung der Tabellen zur bedingten Wahrscheinlichkeit zwischen Knoten beeinflussen können, wenn die Werte der übergeordneten Elemente bekannt sind:

- **Maximum Likelihood.** Aktivieren Sie dieses Kontrollkästchen, wenn Sie ein großes Dataset verwenden. Dies ist die Standardauswahl.
- **Bayes-Anpassung für kleine Anzahl in den Zellen.** Bei kleineren Datasets besteht die Gefahr einer Überanpassung des Modells sowie die Möglichkeit einer hohen Anzahl von Zellen mit einer Zellenhäufigkeit von 0. Mit dieser Option können Sie diese Probleme verringern, indem Sie Glättung zur Reduzierung des Effekts von Zellen mit einer Häufigkeit von 0 und etwaiger unzuverlässiger Schätzungen anwenden.

Expertenoptionen für Bayes-Netzknoten

Mit dem Knoten "Expertenoptionen" können Sie die Feinabstimmung der Modellerstellung vornehmen. Für den Zugriff auf die Expertenoptionen setzen Sie den Modus auf der Registerkarte "Experten" auf **Experten**.

Fehlende Werte. Standardmäßig verwendet IBM SPSS Modeler nur Datensätze mit gültigen Werten für alle im Modell verwendeten Felder. (Dies wird zuweilen als **listenweiser Ausschluss** fehlender Werte bezeichnet.) Wenn sehr viele fehlende Daten vorliegen, werden mit diesem Ansatz möglicherweise zu viele Datensätze entfernt, sodass nicht mehr genügend Daten zu Erstellung eines guten Modells vorhanden sind. In solchen Fällen können Sie die Auswahl der Option **Nur vollständige Datensätze verwenden** aufheben. IBM SPSS Modeler versucht anschließend, so viele Informationen wie möglich zu verwenden, um das Modell zu schätzen. Hierzu zählen auch Datensätze, bei denen bei einigen Feldern Werte fehlen. (Dies wird zuweilen als **paarweiser Ausschluss** fehlender Werte bezeichnet.) In einigen Situationen jedoch kann eine derartige Verwendung unvollständiger Datensätze zu Berechnungsproblemen bei der Schätzung des Modells führen.

Alle Wahrscheinlichkeiten anhängen. Gibt an, ob die Wahrscheinlichkeiten für die einzelnen Kategorien des Ausgabefelds zu den einzelnen vom Knoten verarbeiteten Datensätzen hinzugefügt werden. Wenn diese Option nicht ausgewählt wurde, wird nur die Wahrscheinlichkeit der vorhergesagten Kategorie hinzugefügt.

Unabhängigkeitstest. Ein Test auf Unabhängigkeit dient zur Einschätzung, ob paarige Beobachtungen bei zwei Variablen voneinander unabhängig sind. Wählen Sie den Typ des zu verwendenden Tests aus. Folgende Optionen sind verfügbar:

- **Likelihood-Quotient.** Testet auf Unabhängigkeit zwischen Ziel und Prädiktor durch Berechnung des Verhältnisses zwischen der maximalen Wahrscheinlichkeit eines Ergebnisses unter zwei verschiedenen Hypothesen.
- **Pearson-Chi-Quadrat.** Testet auf Unabhängigkeit zwischen Ziel und Prädiktor unter Verwendung der Nullhypothese, dass die relativen Häufigkeiten des Eintretens beobachteter Ereignisse einer angegebenen Häufigkeitsverteilung folgen.

Bayes-Netzmodelle führen bedingte Tests auf Unabhängigkeit durch, bei denen über die getesteten Paare hinaus zusätzliche Variablen verwendet werden. Außerdem untersuchen die Modelle nicht nur die Beziehungen zwischen Ziel und Prädiktoren, sondern auch die Beziehungen zwischen den Prädiktoren selbst.

Hinweis: Die Optionen für Unabhängigkeitstests sind nur verfügbar, wenn auf der Registerkarte "Modell" **Vorbereitenden Merkmalauswahlschritt einschließen** oder den Strukturtyp **Markov-Decke** ausgewählt wurde.

Signifikanzniveau. In Verbindung mit den Einstellungen für den Unabhängigkeitstest können Sie mit dieser Einstellung einen Trennwert festlegen, der bei der Durchführung der Tests verwendet werden soll. Je niedriger der Wert, desto weniger Verbindungen verbleiben im Netz; das Standardniveau ist 0,01.

Hinweis: Die Option ist nur verfügbar, wenn auf der Registerkarte "Modell" **Vorbereitenden Merkmalauswahlschritt einschließen** oder der Strukturtyp **Markov-Decke** ausgewählt wurde.

Maximale Größe des Konditionierungssets. Der Algorithmus zum Erstellen einer Struktur vom Typ "Markov-Decke" verwendet Konditionierungssets mit zunehmender Größe, um Unabhängigkeitstests durchzuführen und unnötige Verbindungen aus dem Netz zu entfernen. Da die Tests mit einer höheren Anzahl von Konditionierungsvariablen einen größeren Zeit- und Speicherbedarf für die Verarbeitung aufweisen, können Sie die Anzahl der aufzunehmenden Variablen begrenzen. Dies kann insbesondere bei der Verarbeitung von Daten mit starken Abhängigkeiten zwischen vielen Variablen nützlich sein. Beachten Sie jedoch, dass das so entstehende Netz einige überflüssige Verbindungen enthalten kann.

Geben Sie die Maximalzahl der für die Unabhängigkeitstests zu verwendenden Konditionierungsvariablen an. Die Standardeinstellung ist 5.

Hinweis: Die Option ist nur verfügbar, wenn auf der Registerkarte "Modell" **Vorbereitenden Merkmalauswahlschritt einschließen** oder der Strukturtyp **Markov-Decke** ausgewählt wurde.

Merkmalauswahl. Mit dieser Option können Sie die Anzahl der bei der Verarbeitung des Modells verwendeten Eingaben beschränken, um die Modellerstellung zu beschleunigen. Dies ist aufgrund der möglicherweise großen Anzahl potenzieller Eingaben besonders nützlich bei Erstellung einer Struktur vom Typ "Markov-Decke", da Sie auf diese Weise die Eingaben auswählen können, die in einer signifikanten Beziehung zur Zielvariablen stehen.

Hinweis: Die Optionen für die Merkmalauswahl sind nur verfügbar, wenn Sie **Vorbereitenden Merkmalauswahlschritt einschließen** auf der Registerkarte "Modell" auswählen.

- **Eingaben immer ausgewählt** Mit der Feldauswahlschaltfläche rechts neben dem Textfeld können Sie Felder aus dem Dataset auswählen, die immer bei Erstellung des Bayes-Netzmodells verwendet werden sollen. Beachten Sie, dass das Zielfeld immer ausgewählt ist.
- **Maximale Anzahl an Eingaben.** Geben Sie die Gesamtzahl an Eingaben aus dem Dataset an, die beim Erstellen des Bayes-Netzmodells verwendet werden sollen. Als Höchstwert kann die Gesamtzahl der Eingaben im Dataset eingegeben werden.

  Hinweis: Wenn die Anzahl der in **Immer ausgewählte Eingaben** ausgewählten Felder den Wert von **Maximale Anzahl an Eingaben** überschreitet, wird eine Fehlernachricht angezeigt.

Modellnuggets vom Typ "Bayes-Netz"

Hinweis: Wenn Sie die Option zum Fortsetzen des Trainings der vorhandenen Parameter auf der Registerkarte "Modell" des Modellierungsknotens ausgewählt haben, werden die auf der Registerkarte "Modell" des Modellnuggets angezeigten Informationen bei jeder erneuten Generierung des Modells aktualisiert.

Die Registerkarte "Modell" des Modellnuggets gliedert sich in zwei Bereiche:

Linker Bereich

Basis. Diese Ansicht enthält ein Netzdiagramm mit Knoten, das die Beziehung zwischen dem Ziel und seinen wichtigsten Prädiktoren sowie die Beziehung zwischen den Prädiktoren anzeigt. Der Einfluss der einzelnen Prädiktoren wird durch die Farbdichte angezeigt. Eine vollere Farbe zeigt einen bedeutsamen Prädiktor an und umgekehrt.

Die Klassenwerte für Knoten, die für einen Bereich stehen, werden in einer Popup-QuickInfo angezeigt, wenn Sie mit dem Mauszeiger über den Knoten fahren.

Mit den Diagrammtools von IBM SPSS Modeler können Sie in das Diagramm eingreifen, es bearbeiten und speichern. Beispielsweise zur Verwendung in anderen Programmen wie MS Word.

Tipp: Wenn das Netz sehr viele Knoten enthält, können Sie auf einen Knoten klicken und ihn an eine andere Stelle ziehen, um die Lesbarkeit des Diagramms zu verbessern.

Verteilung. In dieser Ansicht werden die bedingten Wahrscheinlichkeiten für die einzelnen Knoten im Netz als Minidiagramme angezeigt. Bewegen Sie den Mauszeiger über ein Diagramm, um dessen Werte in einer Popup-QuickInfo anzuzeigen.

Rechter Bereich

Prädiktoreinfluss. Damit wird eine Tabelle angezeigt, die die relative Wichtigkeit der einzelnen Prädiktoren für die Schätzung des Modells angibt. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Bedingte Wahrscheinlichkeiten. Wenn Sie im linken Bereich einen Knoten oder ein Miniaturverteilungsdiagramm auswählen, wird im rechten Bereich die zugehörige Tabelle mit bedingten Wahrscheinlichkeiten angezeigt. Diese Tabelle enthält den Wert der bedingten Wahrscheinlichkeit für die einzelnen Knotenwerte und die einzelnen Kombinationen von Werten in ihren übergeordneten Knoten. Außerdem beinhaltet sie die Anzahl der für die einzelnen Datensatzwerte beobachteten Datensätze und die einzelnen Kombinationen von Werten in den übergeordneten Knoten.

Einstellungen im Bayes-Netzmodell

Auf der Registerkarte "Einstellungen" für ein Modellnugget vom Typ "Bayes-Netz" werden Optionen zum Ändern des erstellten Modells angegeben. Beispielsweise können Sie mit dem Bayes-Netzknoten unter Verwendung derselben Daten und Einstellungen mehrere verschiedene Modelle erstellen und anschließend diese Registerkarte bei jedem Modell verwenden, um die Einstellungen leicht abzuändern und zu ermitteln, welche Auswirkungen dies auf die Ergebnisse hat.

Hinweis: Diese Registerkarte ist erst verfügbar, nachdem das Modellnugget einem Stream hinzugefügt wurde.

Raw-Propensity-Scores berechnen. Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Diese Werte werden zusätzlich zu den anderen Vorhersage- und Konfidenzwerten angegeben, die gegebenenfalls während des Scorings erstellt werden.

Adjusted-Propensity-Scores berechnen. Raw-Propensity-Scores beruhen ausschließlich auf den Trainingsdaten und sind aufgrund der Neigung vieler Modelle zur übermäßigen Anpassung an die Trainingsdaten möglicherweise zu optimistisch. Bei Adjusted Propensitys wird versucht, dies durch Evaluation der Modellleistung anhand einer Test- bzw. Validierungspartition zu kompensieren. Bei dieser Option muss im Stream ein Partitionsfeld definiert sein und die Adjusted-Propensity-Scores müssen im Modellierungsknoten aktiviert werden, bevor das Modell generiert wird.

Alle Wahrscheinlichkeiten ausgeben. Gibt an, ob die Wahrscheinlichkeiten für die einzelnen Kategorien des Ausgabefelds zu den einzelnen vom Knoten verarbeiteten Datensätzen hinzugefügt werden. Wenn diese Option nicht ausgewählt wurde, wird nur die Wahrscheinlichkeit der vorhergesagten Kategorie hinzugefügt.

Die Standardeinstellung dieses Kontrollkästchens richtet sich nach dem entsprechenden Kontrollkästchen auf der Registerkarte "Experten" des Modellierungsknotens. Weitere Informationen finden Sie im Thema „Expertenoptionen für Bayes-Netzknoten" auf Seite 122.

Bayes-Netzmodell - Übersicht

Auf der Registerkarte "Übersicht" eines Modellnuggets finden Sie Informationen zum Modell selbst (*Analyse*), den im Modell verwendeten Feldern (*Felder*), den beim Erstellen des Modells verwendeten Einstellungen (*Aufbaueinstellungen*) und zum Modelltraining (*Trainingsübersicht*).

Beim ersten Durchsuchen des Knotens sind die Ergebnisse der Registerkarte "Übersicht" reduziert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie sie mithilfe des Erweiterungssteuerelements links neben dem betreffenden Element oder klicken Sie auf die Schaltfläche **Alles anzeigen**, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement die gewünschten Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche **Alles ausblenden** alle Ergebnisse ausblenden.

Analyse. Zeigt Informationen zum jeweiligen Modell an.

Felder. Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden.

Erstellungseinstellungen. Enthält Informationen zu den beim Aufbau des Modells verwendeten Einstellungen.

Trainingsübersicht. In diesem Abschnitt werden folgende Informationen angezeigt: der Typ des Modells, der für seine Erstellung verwendete Stream, der Benutzer, der ihn erstellt hat, der Erstellungszeitpunkt und die für den Aufbau des Modells benötigte Zeit.

Kapitel 8. Neuronale Netze

Mit einem **neuronalen Netz** können Näherungswerte für eine große Bandbreite an Vorhersagemodellen mit minimalen Anforderungen an die Modellstruktur und minimalen Annahmen für das Modell berechnet werden. Die Form der Beziehungen wird während des Lernprozesses bestimmt. Wenn sich eine lineare Beziehung zwischen Ziel und Prädiktoren eignet, sollten sich die Ergebnisses des neuronalen Netzes denen eines klassischen linearen Modells stark annähern. Ist eine nicht lineare Beziehung besser geeignet, erstellt das neuronale Netz automatisch eine Näherung für die "korrekte" Modellstruktur.

Der Nachteil dieser Flexibilität besteht darin, dass das neuronale Netz schwer zu interpretieren ist. Sollten Sie versuchen, einen Prozess zu erklären, der den Beziehungen zwischen Ziel und Prädiktoren zugrunde liegt, ist ein klassisches statistisches Modell besser geeignet. Wenn die Interpretierbarkeit des Modells jedoch keine große Rolle spielt, können Sie mit einem neuronalen Netz gute Vorhersagen erzielen.

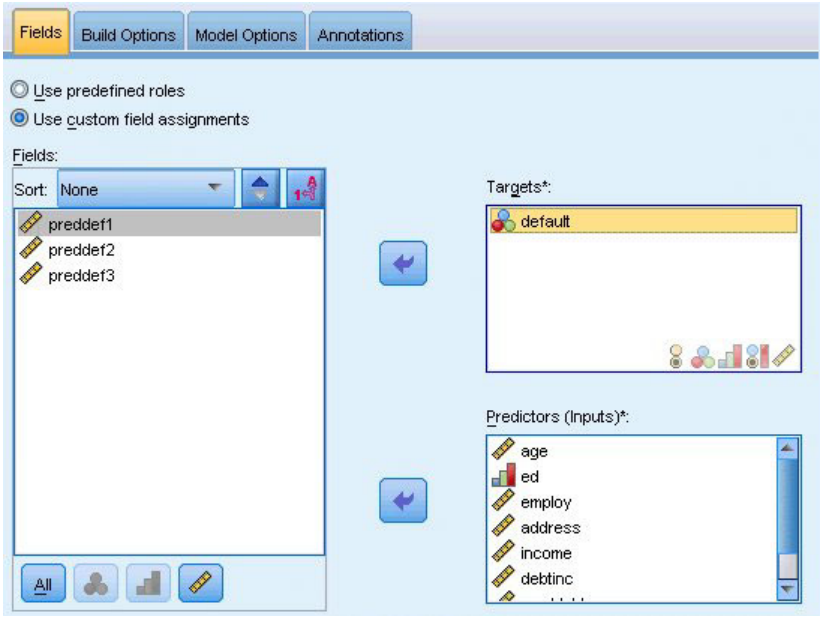


Abbildung 30. Registerkarte "Felder"

Feldanforderungen. Es müssen mindestens ein Ziel und eine Eingabe vorhanden sein. Felder, die auf Beides oder Keine gesetzt sind, werden ignoriert. Es gibt keine Messniveaubeschränkungen bei Zielen oder Prädiktoren (Eingaben). Weitere Informationen finden Sie im Thema „Feldoptionen der Modellierungsknoten” auf Seite 31.

Neuronales Netzmodell

Neuronale Netze sind einfache Modelle der Funktionsweise des Nervensystems. Die Grundeinheiten sind **Neuronen**, die in der Regel in **Schichten** organisiert sind, wie in der folgenden Abbildung dargestellt.

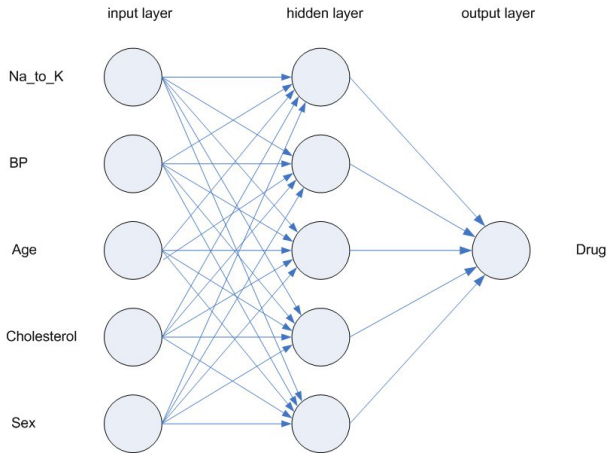


Abbildung 31. Struktur eines neuronalen Netzes

Ein **neuronales Netz** ist ein vereinfachtes Modell der Art und Weise, wie ein menschliches Gehirn Informationen verarbeitet. Es funktioniert, indem eine große Anzahl miteinander verbundener Verarbeitungseinheiten simuliert wird, die abstrakten Versionen von Neuronen ähnlich sind.

Die Verarbeitungseinheiten sind in Schichten angeordnet. Für gewöhnlich gibt es drei Schichten in einem neuronalen Netz: eine **Eingabeschicht** mit Einheiten, die die Eingabefelder darstellen, mindestens eine **verborgene Schicht** und eine **Ausgabeschicht** mit einer Einheit bzw. Einheiten, die das(die) Zielfeld(er) darstellen. Die Einheiten sind mit verschiedenen Verbindungsstärken (**Gewichtungen**) verbunden. Die Eingangsdaten werden der ersten Schicht präsentiert und die Werte von jedem Neuron an jedes Neuron in der nächsten Schicht weitergeleitet. Schließlich gibt die Ausgabeschicht ein Ergebnis aus.

Das Netz lernt durch Prüfen einzelner Datensätze und generiert eine Vorhersage für jeden Datensatz. Außerdem nimmt es Änderungen der Gewichtungen vor, sobald eine falsche Vorhersage erfolgt. Dieser Vorgang wird viele Male wiederholt. Das Netz verbessert seine Vorhersagen so lange, bis mindestens eines der Stoppkriterien erfüllt ist.

Ursprünglich sind alle Gewichtungen zufällig und die Antworten, die vom Netz stammen, sind wahrscheinlich unsinnig. Das Netz lernt durch **Training**. Beispiele, für die die Ausgabe bekannt ist, werden dem Netz wiederholt präsentiert und die Antworten werden mit den bekannten Ausgaben verglichen. Die Informationen aus diesem Vergleich werden durch das Netz geleitet und die Gewichtungen schrittweise geändert. Je weiter das Training fortschreitet, desto genauer wird das Netz bei der Replizierung der bekannten Ergebnisse. Sobald das Netz trainiert ist, kann es auf zukünftige Fälle angewendet werden, bei denen das Ergebnis unbekannt ist.

Verwenden von neuronalen Netzen mit traditionellen Streams

In Version 14 von IBM SPSS Modeler ist jetzt ein neuer Netzknoten verfügbar, der Techniken für Boosting und Bagging und Optimierung für sehr große Datensätze unterstützt. Vorhandene Streams, die den alten Knoten enthalten, werden in dieser Ausgabe weiterhin Modelle erstellen und scoren. Diese Unterstützung wird jedoch in zukünftigen Versionen wegfallen, sodass wir von nun an die Verwendung der neuen Version empfehlen.

Ab Version 13 werden Felder mit unbekannten Werten (also Werte, die nicht in den Trainingsdaten vorhanden sind) nicht mehr automatisch als fehlende Werte behandelt und mit dem Wert `$null$` gescort. Wenn Sie also Felder mit unbekannten Werten mit einer älteren Version des neuronalen Netzmodells (vor Version 13) in Version 13 oder höher als Nicht-Null-Wert scoren möchten, sollten Sie unbekannte Werte als fehlende Werte markieren (beispielsweise mithilfe des Typknotens).

Beachten Sie, dass traditionelle Streams, die immer noch den alten Knoten enthalten, aus Gründen der Kompatibilität möglicherweise immer noch die Option *Dimension der Variablen begrenzen* unter **Tools > Streameigenschaften > Optionen** verwenden. Diese Option gilt nur für Kohonen-Netze und *K*-Means-Knoten ab Version 14.

Ziele

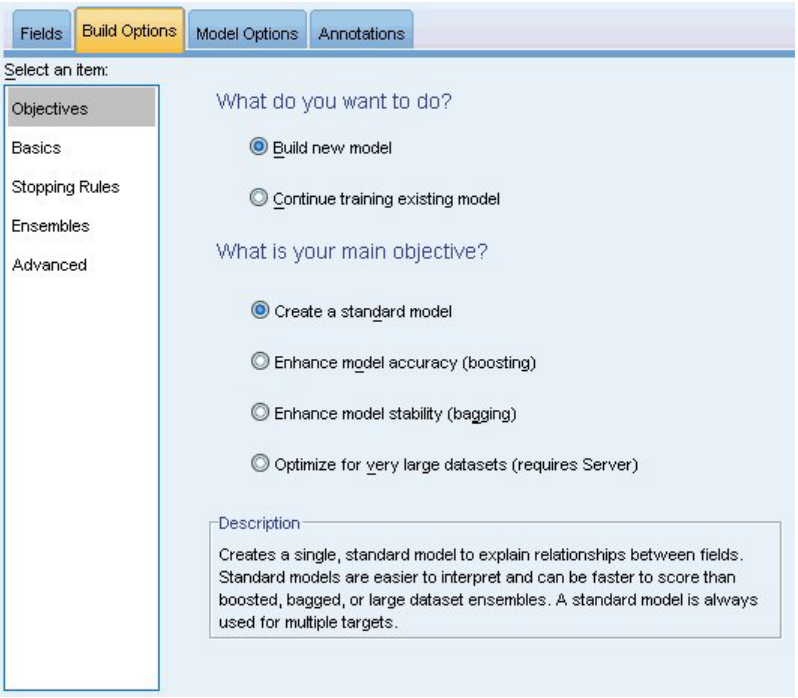


Abbildung 32. Ziele - Einstellungen

Was möchten Sie tun?

- **Neues Modell erstellen.** Ein vollständig neues Modell aufbauen. Dies ist die übliche Wirkungsweise des Knotens.
- **Training des bestehenden Modells fortsetzen.** Das Training wird mit dem letzten vom Knoten erfolgreich erstellten Modell fortgesetzt. Dies ermöglicht die Aktualisierung eines bestehenden Modells, ohne dass ein Zugriff auf die ursprünglichen Daten erforderlich ist, und kann zu einer wesentlich schnelleren Leistung führen, da ausschließlich die neuen bzw. aktualisierten Datensätze in den Stream eingespeist werden. Details zum vorherigen Modell werden zusammen mit dem Modellierungsknoten gespeichert, wodurch diese Option auch dann verwendet werden kann, wenn das vorherige Modellnugget nicht mehr im Stream oder in der Modellpalette verfügbar ist.

  Hinweis: Wenn diese Option aktiviert ist, werden alle anderen Steuerelemente auf den Registerkarten "Felder" und "Erstellungsoptionen" inaktiviert.

Was ist Ihr Hauptziel? Wählen Sie das geeignete Ziel aus.

- **Standardmodell erstellen.** Bei der Methode wird ein einziges Modell erstellt, um das Ziel unter Verwendung der Prädiktoren vorherzusagen. In der Regel gilt, dass Standardmodelle einfacher interpretiert und schneller gescort werden können, als verbesserte, verstärkte oder große Dataset-Ensembles.
- **Modellgenauigkeit verbessern (Boosting).** Bei der Methode wird mittels Boosting ein Ensemblemodell erstellt. Dabei wird eine Modellsequenz erzeugt, um genauere Vorhersagen zu erhalten. Möglicherweise nimmt das Erstellen und Scoren bei Ensembles mehr Zeit in Anspruch als bei Standardmodellen.

  Durch Boosting wird eine Reihe von "Komponentenmodellen" erstellt, von denen jede einzelne Komponente auf dem gesamten Dataset beruht. Vor dem Erstellen jedes aufeinander folgenden Komponenten-

modells werden die Datensätze basierend auf den Residuen des vorangegangenen Komponentenmodells gewichtet. Größere Residuen erhalten eine höhere Analysegewichtung, sodass beim nächsten Komponentenmodell das Augenmerk auf einer hochwertigen Vorhersage dieser Datensätze liegt. Zusammen bilden diese Komponentenmodelle ein Ensemblemodell. Das Ensemblemodell scort basierend auf einer Kombinationssregel neue Datensätze. Die verfügbaren Regeln hängen vom Messniveau des Ziels ab.

- **Modellstabilität verbessern (Bagging).** Bei der Methode wird mittels Bagging (Bootstrap-Aggregation) ein Ensemblemodell erstellt. Dabei werden mehrere Modelle erzeugt, um zuverlässigere Vorhersagen zu erhalten. Möglicherweise nimmt das Erstellen und Scoren bei Ensembles mehr Zeit in Anspruch als bei Standardmodellen.

  Bei der Bootstrap-Aggregation (Bagging) werden Reproduktionen des Trainingsdataset erstellt, indem von der Ersetzung aus dem ursprünglichen Dataset Stichproben genommen werden. Dadurch werden Bootstrap-Stichproben mit der gleichen Größe wie beim ursprünglichen Dataset erstellt. Dann wird von jeder Reproduktion ein "Komponentenmodell" erstellt. Zusammen bilden diese Komponentenmodelle ein Ensemblemodell. Das Ensemblemodell scort basierend auf einer Kombinationssregel neue Datensätze. Die verfügbaren Regeln hängen vom Messniveau des Ziels ab.
- **Modell für sehr umfangreiche Datasets erstellen (erfordert IBM SPSS Modeler Server).** Bei dieser Methode wird ein Ensemblemodell durch Aufteilen des Datasets in separate Datenblöcke erstellt. Diese Option ist empfehlenswert, wenn Ihr Dataset zu groß für die Erstellung eines der oben erwähnten Modelle oder die inkrementelle Modellerstellung ist. Unter Umständen kann das Modell mit dieser Option schneller als ein Standardmodell erstellt werden, das Scoren dauert jedoch evtl. länger als bei einem Standardmodell. Für diese Option ist eine Verbindung zuIBM SPSS Modeler Server erforderlich.

Wenn es mehrere Ziele gibt, wird bei dieser Methode nur ein Standardmodell erstellt, unabhängig vom ausgewählten Ziel.

Grundeinstellungen

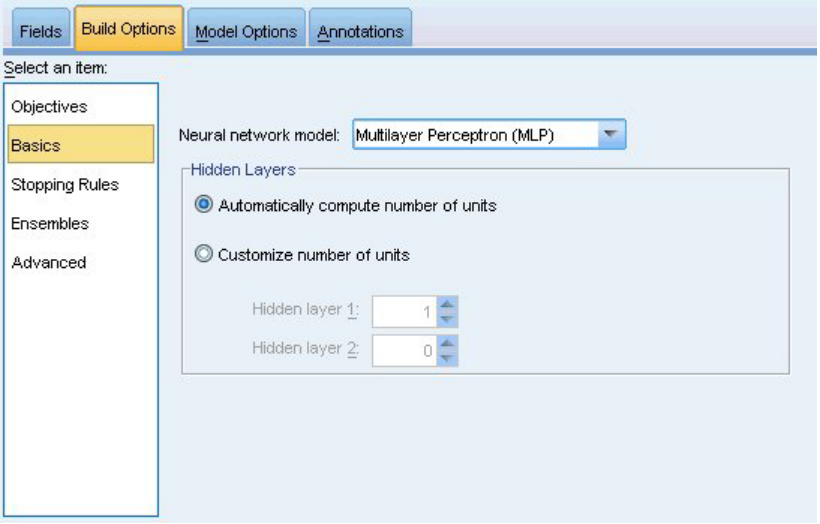


Abbildung 33. Grundeinstellungen

Neuronales Netzmodell. Der Modelltyp bestimmt, wie das Netz die Prädiktoren durch die verborgene(n) Schichte(n) verbindet. Das **Mehrschicht-Perzeptron (MLP)** ermöglicht komplexere Beziehungen, was sich jedoch unter Umständen auf längere Trainings- und Scoringzeiten auswirkt. Die **radiale Basisfunktion (RBF)** kann die Trainings- und Scoringdauer verringern, was jedoch unter Umständen Einbußen bei der Vorhersagekraft im Vergleich zu MLP mit sich bringt.

Verborgene Schicht. Die verdeckte(n) Schicht(en) eines neuronalen Netzes enthält nicht sichtbare Einheiten. Der Wert jeder verdeckten Einheit ist eine Funktion der Prädiktoren; die exakte Form der Funktion hängt teilweise vom Netztyp ab. Ein Mehrschicht-Perzeptron kann eine oder zwei verdeckte Schichten besitzen; ein radiales Basisfunktionsnetz kann nur eine verdeckte Schicht besitzen.

- **Anzahl der Einheiten automatisch berechnen.** Mit dieser Option wird ein Netz mit einer verdeckten Schicht erstellt und die "beste" Anzahl an Einheiten in der verdeckten Schicht berechnet.
- **Anzahl der Einheiten anpassen.** Mit dieser Option können Sie die Anzahl der Einheiten in jeder verdeckten Schicht eingeben. Die erste verdeckte Schicht muss mindestens eine Einheit aufweisen. Durch Eingabe von 0 Einheiten in der zweiten verdeckten Schicht wird ein Mehrschicht-Perzeptron mit einer einzigen verdeckten Schicht erstellt.

  Hinweis: Sie sollten Werte so auswählen, dass die Anzahl der Knoten die Anzahl der stetigen Prädiktoren plus die Gesamtzahl der Kategorien in allen kategorialen Prädiktoren (Flag, nominal und ordinal) nicht überschreitet.

Stoppregeln

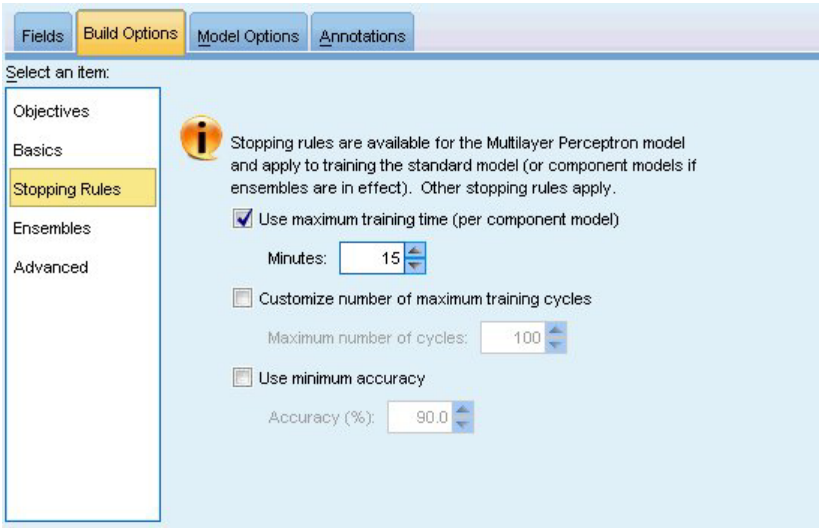


Abbildung 34. Stoppregeln - Einstellungen

Mit diesen Regeln wird festgelegt, wann das Training von Mehrschicht-Perzeptron-Netzen gestoppt werden soll; diese Einstellungen werden ignoriert, wenn der radiale Basisfunktionsalgorithmus verwendet wird. Das Training durchläuft mindestens einen Zyklus (Datendurchlauf) und kann dann entsprechend den folgenden Kriterien gestoppt werden:

Maximale Trainingsdauer verwenden (pro Komponentenmodell). Sie können wählen, ob sie eine maximale Minutenanzahl für die Ausführung des Algorithmus angeben wollen. Geben Sie eine Zahl größer 0 ein. Wenn ein Ensemblemodell erstellt wird, ist das die zulässige Trainingszeit für jedes Komponentenmodell des Ensembles. Das Training kann ein wenig länger dauern, als die angegebene Zeit, da der aktuelle Zyklus abgeschlossen wird.

Anzahl maximaler Trainingszyklen anpassen. Die maximale Anzahl der zulässigen Trainingszyklen. Wenn die maximale Anzahl an Zyklen überschritten wird, stoppt das Training. Geben Sie eine ganze Zahl größer 0 ein.

Mindestgenauigkeit verwenden. Bei dieser Option wird das Training fortgesetzt, bis die angegebene Genauigkeit erreicht ist. Es kann vorkommen, dass dieser Zustand überhaupt nicht erreicht wird; sie können jedoch das Training jederzeit unterbrechen und das Netz mit der besten Genauigkeit speichern, die bisher erzielt wurde.

Der Trainingsalgorithmus stoppt auch dann, wenn der Fehler im Set zur Verhinderung übermäßiger Anpassung nach den einzelnen Zyklen nicht abnimmt, wenn die relative Änderung im Trainingsfehler gering ist, oder wenn die Quote des aktuellen Trainingsfehlers gering im Vergleich zum Anfangsfehler ist.

Ensembles

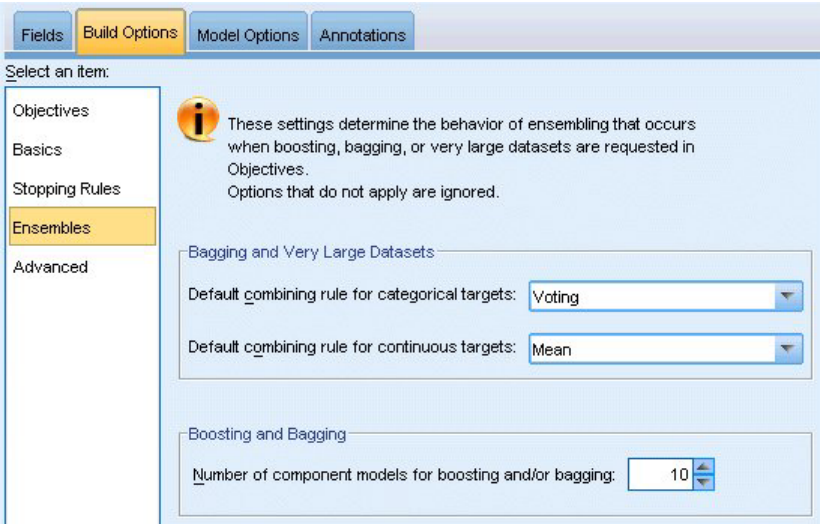


Abbildung 35. Ensemble-Einstellungen

Diese Einstellungen legen das Verhalten der Ensemblebildung fest, die erfolgt, wenn auf der Registerkarte "Ziele" die Option "Boosting", "Bagging" oder "Sehr große Datasets" ausgewählt ist. Optionen, die für das ausgewählte Ziel nicht gelten, werden ignoriert.

Bagging und sehr umfangreiche Datasets. Beim Scoren eines Ensembles wird diese Regel angewendet, um die vorhergesagten Werte aus den Basismodellen für die Berechnung des Scorewerts für das Ensemble zu kombinieren.

- **Standardkombinationsregel für kategoriale Ziele.** Ensemblevorhersagewerte für kategoriale Ziele können mithilfe von "Voting", "höchster Wahrscheinlichkeit" oder "höchste mittlere Wahrscheinlichkeit" kombiniert werden. Mit **Voting** wird die Kategorie gewählt, die in allen Basismodellen am häufigsten die höchste Wahrscheinlichkeit erreicht. Mit **Höchste Wahrscheinlichkeit** wird die Kategorie gewählt, die in allen Basismodellen den höchsten Einzelwert bei der höchsten Wahrscheinlichkeit erzielt. Mit **Höchste mittlere Wahrscheinlichkeit** wird die Kategorie mit dem höchsten Wert ausgewählt, wenn der Mittelwert der Kategoriewahrscheinlichkeiten aus der Menge aller Basismodelle berechnet wird.
- **Standardkombinationsregel für stetige Ziele.** Ensemblevorhersagewerte für stetige Ziele können unter Verwendung des Mittelwerts oder Medians der Vorhersagewerte aus den Basismodellen kombiniert werden.

Hinweis: Wenn als Ziel die Verbesserung der Modellgenauigkeit ausgewählt wurde, wird die Auswahl zum Kombinieren der Regeln ignoriert. Beim Boosting wird für das Scoring der kategorialen Ziele stets eine gewichtete Mehrheit verwendet und für das Scoring stetiger Ziele ein gewichteter Median.

Boosting und Bagging. Geben Sie die Anzahl der zu erstellenden Basismodelle an, wenn als Ziel die Verbesserung der Modellgenauigkeit oder -stabilität angegeben ist. Im Falle des Bagging ist das die Anzahl der Bootstrap-Stichproben. Muss eine positive ganze Zahl sein.

Erweitert

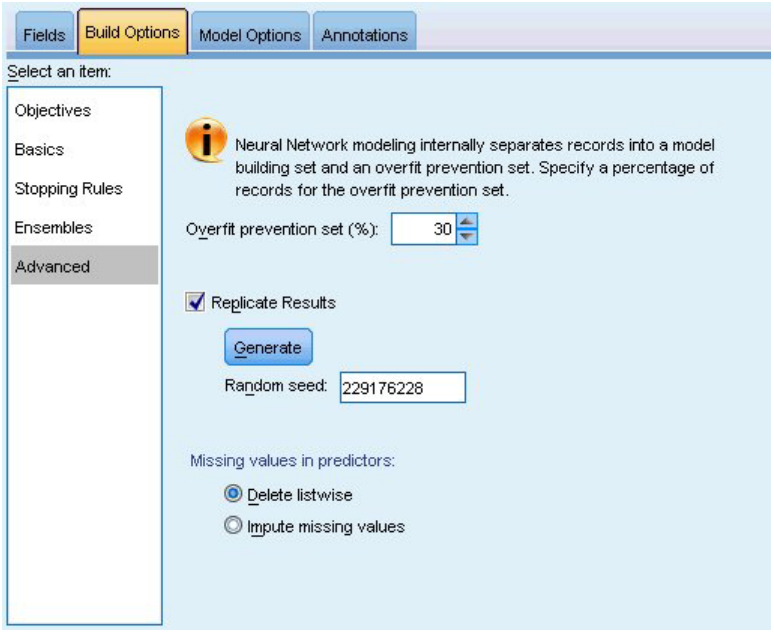


Abbildung 36. Erweiterte Einstellungen

Mit den erweiterten Einstellungen lassen sich Optionen steuern, die nicht wirklich in andere Einstellungsgruppen passen.

Set zur Verhinderung übermäßiger Anpassung. Bei der neuronalen Netzmethode werden Datensätze intern in ein Modellerstellungsset und ein Set zur Verhinderung übermäßiger Anpassung aufgeteilt. Zweiteres ist ein unabhängiges Set an Datensätzen, das dazu verwendet wird, Fehler während des Trainings zu erfassen. So kann verhindert werden, dass die Methode zufällige Variationen in den Daten modelliert. Geben Sie einen Prozentsatz an Datensätzen an. Der Standardwert ist 30.

Ergebnisse replizieren. Durch Einstellen eines Startwerts für den Zufallsgenerator können Analysen reproduziert werden. Geben Sie eine ganze Zahl ein oder klicken Sie auf **Generieren**. Dadurch wird eine pseudozufällige Ganzzahl zwischen 1 und 2147483647 (einschließlich) erzeugt. Standardmäßig werden Analysen mit dem Startwert 229176228 reproduziert.

Fehlende Werte in Prädiktoren. Hier wird festgelegt, wie fehlende Werte behandelt werden sollen. **Listenweise löschen** entfernt Datensätze mit fehlenden Werten bei Prädiktoren aus der Modellerstellung. Mit **Fehlende Werte imputieren** werden fehlende Werte in Prädiktoren ersetzt und diese Datensätze in der Analyse verwendet. Stetige Felder imputieren den Mittelwert der minimalen und maximalen beobachteten Werte; kategoriale Felder imputieren die am häufigsten auftretende Kategorie. Hinweis: Datensätze mit fehlenden Werten in einem anderen auf der Registerkarte "Felder" angegebenen Feld werden stets aus der Modellerstellung ausgeschlossen.

Modelloptionen

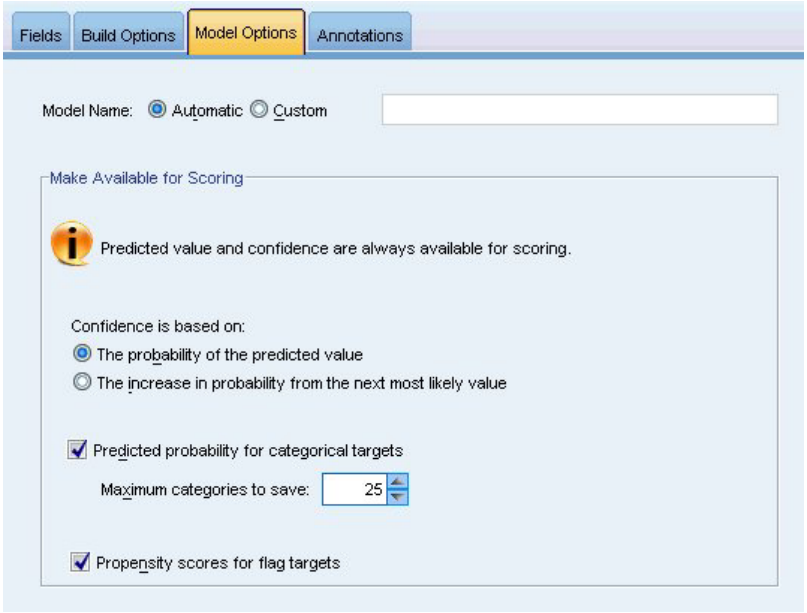


Abbildung 37. Registerkarte "Modelloptionen"

Modellname. Sie können den Modellnamen automatisch basierend auf den Zielfeldern generieren, oder einen benutzerdefinierten Namen angeben. Der automatisch generierte Name ist der Zielfeldname. Bei mehreren Zielen besteht der Modellname aus den Feldnamen, die der Reihe nach durch Und-Zeichen verbunden werden. Wenn beispielsweise*Feld1 Feld2 Feld3* Ziele sind, lautet der Modellname *Feld1 & Feld2 & Feld3*.

Für Scoring bereitstellen. Beim Scoren des Modells sollten die ausgewählten Elemente in dieser Gruppe erzeugt werden. Wenn das Modell gescort wird, werden stets der vorhergesagte Wert (für alle Ziele) und die Konfidenz (für stetige Ziele) berechnet. Die berechnete Konfidenz kann auf der Wahrscheinlichkeit des vorhergesagten Werts (die höchste vorhergesagte Wahrscheinlichkeit) oder der Differenz zwischen der höchsten vorhergesagten Wahrscheinlichkeit und der zweithöchsten vorhergesagten Wahrscheinlichkeit basieren.

- **Vorhergesagte Wahrscheinlichkeit für kategoriale Ziele.** Mit dieser Option werden die vorhergesagten Wahrscheinlichkeiten für kategoriale Ziele produziert. Für jede Kategorie wird ein Feld erstellt.
- **Propensity-Scores für Flagziele.** Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Das Modell erzeugt Raw-Propensity-Scores. Bei aktiven Partitionen erzeugt das Modell außerdem Adjusted-Propensitys-Scores anhand der Testpartition.

Modellübersicht

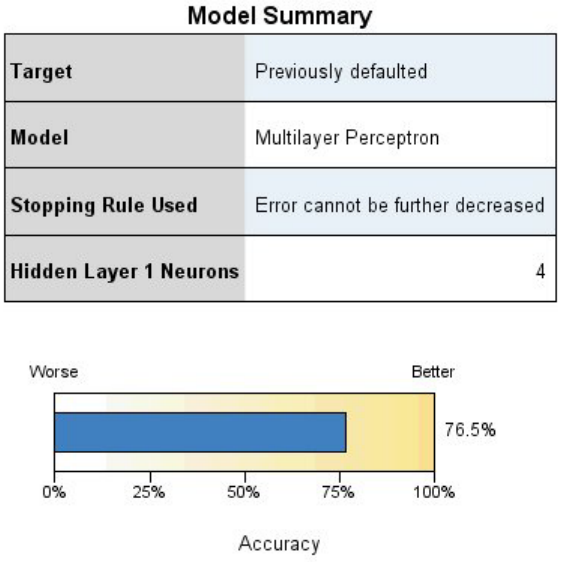


Abbildung 38. Ansicht "Neuronales Netzmodell - Übersicht"

Die Modellübersicht ist eine Momentaufnahme, eine Übersicht auf einen Blick über die Vorhersage- oder Klassifizierungsgenauigkeit des neuronalen Netzes.

Modellübersicht. Die Tabelle ermittelt das Ziel, den Typ des trainierten neuronalen Netzes, die Stoppregel, die das Training gestoppt hat (wird angezeigt, wenn ein Mehrschicht-Perzeptron-Netz trainiert wurde), und die Anzahl der Neuronen in den einzelnen verborgenen Schichten des Netzes.

Neuronales Netz - Qualität. Das Diagramm zeigt die Genauigkeit des endgültigen Modells an, das nach dem Prinzip "größer ist besser" dargestellt wird. Bei einem kategorialen Ziel ist das einfach der Prozentsatz an Datensätzen, zu dem der vorhergesagte Wert dem beobachteten Wert entspricht. Bei stetigen Zielen ist das 1 minus das Verhältnis des mittleren absoluten Fehlers bei der Vorhersage (der Mittelwert der absoluten Werte aus vorhergesagten Werten minus beobachteten Werten) zum Bereich der vorhergesagten Werte (der höchste vorhergesagte Werte minus dem niedrigsten vorhergesagten Wert).

Mehrere Ziele. Bei mehreren Zielen werden die einzelnen Ziele in der Zeile **Ziel** der Tabelle angezeigt. Die im Diagramm angegebene Genauigkeit ist der Mittelwert der einzelnen Zielgenauigkeiten.

Prädiktoreinfluss

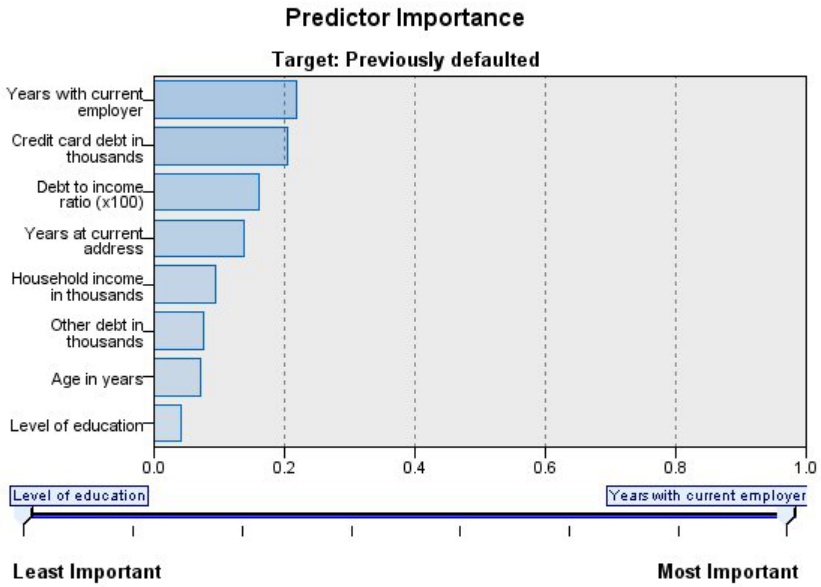


Abbildung 39. Ansicht "Prädiktoreinfluss"

In der Regel konzentriert man sich bei der Modellerstellung auf die Prädiktorfelder, die am wichtigsten sind, und vernachlässigt jene, die weniger wichtig sind. Dabei unterstützt Sie das Wichtigkeitsdiagramm für die Prädiktoren, da es die relative Wichtigkeit der einzelnen Prädiktoren für das Modell angibt. Da die Werte relativ sind, beträgt die Summe der Werte aller Prädiktoren im Diagramm 1,0. Die Wichtigkeit der Prädiktoren steht in keinem Bezug zur Genauigkeit des Modells. Sie bezieht sich lediglich auf die Wichtigkeit der einzelnen Prädiktoren für eine Vorhersage und nicht auf die Genauigkeit der Vorhersage.

Mehrere Ziele. Bei mehreren Zielen wird jedes Ziel in einem separaten Diagramm dargestellt. Zusätzlich ist die Dropdown-Liste **Ziel** vorhanden, aus der das anzuzeigende Ziel ausgewählt werden kann.

Vorhergesagt/Beobachtet

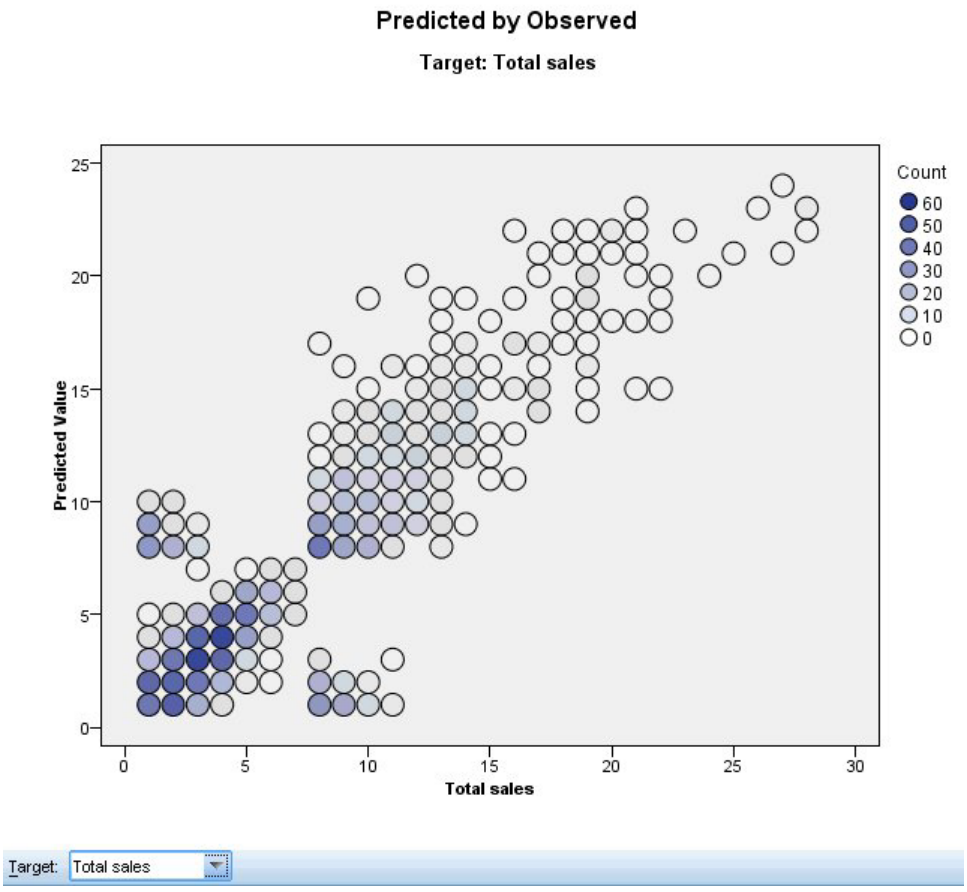


Abbildung 40. Ansicht "Vorhergesagt/Beobachtet"

Für stetige Ziele zeigt diese Ansicht ein Bin-Streudiagramm der vorhergesagten Werte auf der vertikalen Achse durch die beobachteten Werte auf der horizontalen Achse.

Mehrere Ziele. Bei mehreren stetigen Zielen wird jedes Ziel in einem separaten Diagramm angezeigt. Zudem ist die Dropdown-Liste **Ziel** vorhanden, aus der das anzuzeigende Ziel ausgewählt werden kann.

Klassifikation

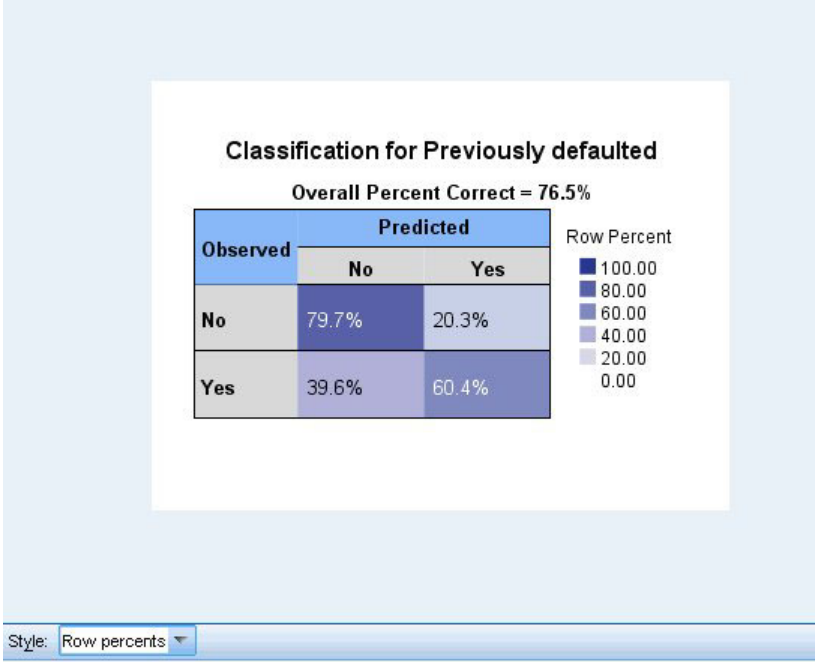


Abbildung 41. Ansicht "Klassifikation", Stil "Reihenprozente"

Bei kategorialen Zielen wird hiermit die Kreuzklassifikation der beobachteten Werte in Abhängigkeit von den vorhergesagten Werten in einer Heat-Map angezeigt, zuzüglich des Gesamtprozentsatzes der korrekten Werte.

Tabellenstile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Zeilenprozentwerte.** Hiermit werden die Zeilenprozentsätze (die Zellenhäufigkeiten ausgedrückt als Prozentsatz der Gesamtzeilenzahl) in den Zellen angezeigt. Dies ist die Standardeinstellung.
- **Zellenhäufigkeit.** Hiermit werden die Zellenhäufigkeiten in den Zellen angezeigt. Die Schattierung für die Heat-Map beruht weiterhin auf den Zeilenprozentsätzen.
- **Heat-Map.** Hiermit werden in den Zellen keine Werte, sondern nur die Schattierung angezeigt.
- **Komprimiert.** Hiermit werden keine Zeilen- oder Spaltenüberschriften oder Werte in den Zellen angezeigt. Diese Option kann nützlich sein, wenn das Ziel sehr viele Kategorien aufweist.

Fehlende Werte. Wenn Datensätze fehlende Werte im Ziel aufweisen, werden diese unter allen gültigen Werten in der Zeile **(Fehlend)** angezeigt. Datensätze mit fehlenden Werten werden im Wert von "Gesamtprozent korrekt" nicht berücksichtigt.

Mehrere Ziele. Bei mehreren kategorialen Zielen wird jedes Ziel in einer separaten Tabelle dargestellt. Dazu gibt es eine Dropdown-Liste **Ziel**, aus der das anzuzeigende Ziel ausgewählt werden kann.

Große Tabellen. Wenn das angezeigte Ziel mehr als 100 Kategorien enthält, wird keine Tabelle angezeigt.

Netz

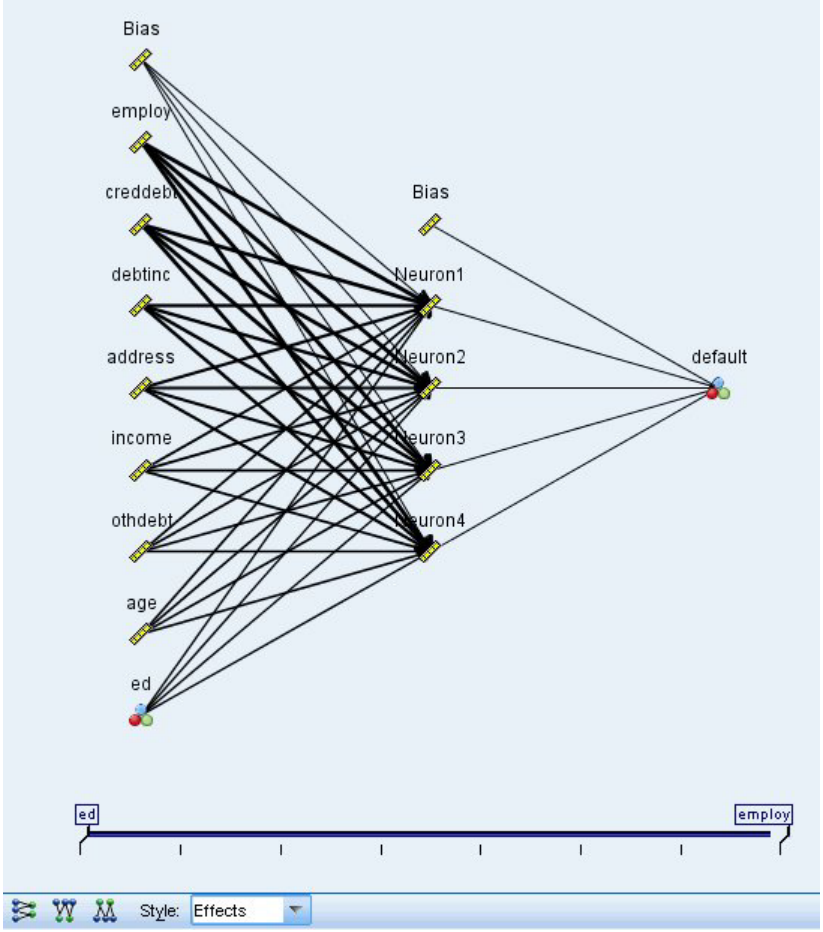


Abbildung 42. Ansicht "Netz", Eingaben links, Stil "Effekte"

Hier wird eine grafische Darstellung des neuronalen Netzes angezeigt.

Diagrammstile. Es sind zwei verschiedene Stile verfügbar, auf die Sie über die Dropdown-Liste **Stil** zugreifen können.

- **Effekte.** Hier werden die einzelnen Prädiktoren und Ziele als ein Knoten im Diagramm angezeigt, unabhängig davon, ob die Messskala stetig oder kategorial ist. Dies ist die Standardeinstellung.
- **Koeffizienten.** Hier werden mehrere Indikatorknoten für kategoriale Prädiktoren und Ziele angezeigt. Die Verbindungslinien im Diagramm im Koeffizienten-Stil sind farbig dargestellt, basierend auf dem geschätzten Wert der synaptischen Gewichtung.

Diagrammausrichtung. Das Netzdiagramm ist standardmäßig so angeordnet, dass die Eingaben links und die Ziele rechts dargestellt sind. Mithilfe der Steuerelemente in der Symbolleiste können Sie die Ausrichtung so ändern, dass die Eingaben oben und die Ziele unten oder die Eingaben unten und die Ziele oben dargestellt werden.
Prädiktoreinfluss. Verbindungslinien im Diagramm sind basierend auf dem Prädiktoreinfluss gewichtet, wobei eine größere Linienstärke einem größeren Einfluss entspricht. Für den Prädiktoreinfluss gibt es einen Schieberegler in der Symbolleiste, mit dem eingestellt wird, welche Prädiktoren im Netzdiagramm gezeigt werden. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Prädiktoren konzentrieren.
Mehrere Ziele. Bei mehreren Zielen werden alle Ziele im Diagramm angezeigt.

Einstellungen

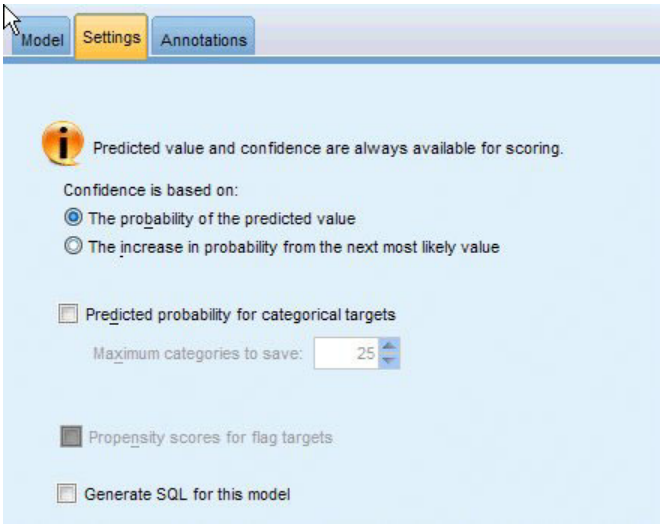


Abbildung 43. Registerkarte "Einstellungen"

Beim Scoring des Modells sollten die ausgewählten Objekte in dieser Registerkarte produziert werden. Wenn das Modell gescort wird, werden stets der vorhergesagte Wert (für alle Ziele) und die Konfidenz (für stetige Ziele) berechnet. Die berechnete Konfidenz kann auf der Wahrscheinlichkeit des vorhergesagten Werts (die höchste vorhergesagte Wahrscheinlichkeit) oder der Differenz zwischen der höchsten vorhergesagten Wahrscheinlichkeit und der zweithöchsten vorhergesagten Wahrscheinlichkeit basieren.

- **Vorhergesagte Wahrscheinlichkeit für kategoriale Ziele.** Mit dieser Option werden die vorhergesagten Wahrscheinlichkeiten für kategoriale Ziele produziert. Für jede Kategorie wird ein Feld erstellt.
- **Propensity-Scores für Flagziele.** Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Das Modell erzeugt Raw-Propensity-Scores. Bei aktiven Partitionen erzeugt das Modell außerdem Adjusted-Propensitys-Scores anhand der Testpartition.

SQL für dieses Modell generieren. Bei Verwendung von Daten aus einer Datenbank kann SQL-Code per Pushback zur Ausführung an die Datenbank zurückübertragen werden. Dadurch kann bei vielen Operationen eine bessere Leistung erzielt werden.

Durch Konvertierung in natives SQL scoren. Wenn ausgewählt, wird SQL-Code zum Scoren des Modells innerhalb der Anwendung generiert.

Kapitel 9. Entscheidungsliste

Entscheidungslistenmodelle kennzeichnen Untergruppen bzw. **Segmente**, die eine höhere oder geringere Wahrscheinlichkeit für ein binäres Ergebnis ("Ja" bzw. "Nein") aufweisen als die Gesamtstichprobe. Sie könnten beispielsweise nach Kunden suchen, deren Abwanderung besonders unwahrscheinlich ist oder die mit größter Wahrscheinlichkeit auf ein Angebot oder eine Kampagne ansprechen. Mit dem Entscheidungslistenviewer erhalten Sie die vollständige Kontrolle über das Modell. Sie können Segmente bearbeiten, Ihre persönlichen Geschäftsregeln hinzufügen, angeben, wie die einzelnen Segmente gescort werden sollen, und das Modell auf zahlreiche andere Weisen anpassen, um den Trefferanteil über alle Segmente hinweg zu optimieren. Er eignet sich besonders gut für die Erstellung von Mailing-Listen bzw. die anderweitige Ermittlung der Datensätze, die für eine bestimmte Kampagne gezielt betrachtet werden sollen. Außerdem können Sie mehrere **Mining-Aufgaben** verwenden, um Modellierungsansätze zu kombinieren, indem Sie beispielsweise Segmente mit hoher und niedriger Leistung in demselben Modell ermitteln und die einzelnen Segmente, je nach Bedarf, in die Scoringphase ein- oder von dieser ausschließen.

Segmente, Regeln und Bedingungen

Ein Modell besteht aus einer Liste von Segmenten, von denen jedes durch eine Regel definiert ist, die übereinstimmende Datensätze auswählt. Eine Regel kann jeweils mehrere Bedingungen aufweisen; Beispiel:

```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

Die Regeln werden in der angegebenen Reihenfolge angewendet. Dabei legt die erste zutreffende Regel das Ergebnis für einen bestimmten Datensatz fest. Für sich genommen können sich Regeln bzw. Bedingungen überschneiden, durch die Reihenfolge der Regeln wird die Mehrdeutigkeit jedoch aufgelöst. Wenn keine Regel zutrifft, wird der Datensatz der Restregel zugewiesen.

Vollständige Kontrolle über das Scoring

Mit dem Entscheidungslistenviewer können Sie Segmente anzeigen, ändern und reorganisieren. Sie können zudem auswählen, welche Segmente für das Scoring ein- oder ausgeschlossen werden sollen. Sie können beispielsweise eine Gruppe von Kunden von zukünftigen Angeboten aus- und andere einschließen und sofort ablesen, wie dies die Gesamttrefferquote beeinflusst. Modelle vom Typ "Entscheidungsliste" geben den Score *Ja* für eingeschlossene Segmente und *$null$* für alles andere, einschließlich des Rests, zurück. Durch diese unmittelbare Steuerung des Scorings sind Entscheidungsliste-Modelle ideal für das Erstellen von Mailing-Listen und sie sind - unter anderem in Callcenter- und Marketinganwendungen - weit verbreitet im Customer Relationship Management.

Mining-Aufgaben, Maße und Auswahlmöglichkeiten

Der Modellierungsvorgang wird durch **Mining-Aufgaben** gesteuert. Jede Mining-Aufgabe initiiert effektiv einen neuen Modellierungsdurchgang und gibt ein neues Set mit alternativen Modellen aus, aus denen Sie wählen können. Die Standardaufgabe beruht auf Ihren ursprünglichen Angaben im Entscheidungsliste-Knoten; Sie können jedoch jede beliebige Anzahl an benutzerdefinierten Aufgaben definieren. Sie können Aufgaben auch iterativ anwenden. Beispielsweise können Sie eine Suche vom Typ "Hohe Wahrscheinlichkeit" für das gesamte Trainingsset und anschließend eine Suche vom Typ "Geringe Wahrscheinlichkeit" für den Rest ausführen, um Segmente mit niedriger Leistung auszusondern.

Datenauswahl

Sie können Datenauswahlmöglichkeiten und benutzerdefinierte Modellmaße für Modellerstellung und -evaluation definieren. Sie können beispielsweise eine Datenauswahl in einer Mining-Aufgabe angeben,

um das Modell auf eine bestimmte Region zuzuschneiden, und ein benutzerdefiniertes Maß erstellen, um zu evaluieren, wie leistungsfähig das betreffende Modell für das gesamte Land ist. Im Gegensatz zu Mining-Aufgaben ändern Maße nicht das zugrunde liegende Modell, sondern bieten vielmehr einen anderen Fokus zur Einschätzung, wie leistungsfähig es ist.

Einbringen Ihres Fachwissens

Durch die Feinabstimmung oder Erweiterung der vom Algorithmus ermittelten Segmente können Sie mithilfe des Entscheidungslistenviewers Ihr Fachwissen direkt in das Modell integrieren. Sie können die vom Modell generierten Segmente bearbeiten oder auf der Grundlage der von Ihnen angegebenen Regeln weitere Segmente hinzufügen. Anschließend können Sie die Änderungen übernehmen und eine Vorschau der Ergebnisse anzeigen.

Um weitere Einblicke zu erhalten, können Sie über eine dynamische Verbindung mit Excel Ihre Daten in Excel exportieren, wo sie zum Erstellen von Präsentationsdiagrammen und zum Berechnen von benutzerdefinierten Maßen verwendet werden können, wie beispielsweise komplexe Profit- und ROI-Werte, die während der Modellerstellung im Entscheidungslistenviewer angezeigt werden können.

Beispiel. Die Marketingabteilung eines Finanzinstituts möchte in zukünftigen Kampagnen profitablere Ergebnisse erzielen, indem jedem Kunden ein speziell für ihn geeignetes Angebot unterbreitet wird. Mit einem Entscheidungslistenmodell können Sie basierend auf früheren Werbeaktionen die Merkmale von Kunden ermitteln, die mit der größten Wahrscheinlichkeit positiv antworten werden, und basierend auf den Ergebnissen eine Mailing-Liste erstellen.

Anforderungen. Ein einzelnes kategoriales Zielfeld mit einem Messniveau des Typs *Flag* oder *Nominal*, das das binäre Ergebnis angibt, das Sie vorhersagen möchten (Ja/Nein), sowie mindestens ein Eingabefeld. Wenn das Zielfeld den Typ *Nominal* aufweist, müssen Sie manuell einen einzelnen Wert auswählen, der als **Treffer** oder **Antwort** behandelt werden soll; alle anderen Werte werden als **kein Treffer** zusammengefasst. Außerdem kann ein optionales Häufigkeitsfeld angegeben werden. Stetige Datums-/Uhrzeitfelder werden ignoriert. Eingaben mit stetigem numerischen Bereich werden automatisch vom Algorithmus klassiert, wie auf der Registerkarte "Experten" des Modellierungsknotens angegeben. Eine detailliertere Kontrolle über das Klassieren erhalten Sie, wenn Sie weiter oben im Stream einen Klassierknoten einfügen und das klassierte Feld als Eingabe mit dem Messniveau *Ordinal* verwenden.

Entscheidungslistenmodell - Optionen

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen” auf Seite 28.

Modalwert. Legt fest, welche Methode für die Modellbildung verwendet wird.

- **Modell generieren.** Generiert beim Ausführen des Knotens automatisch ein Modell in der Modellpalette. Das so entstandene Modell kann zum Zwecke des Scorens zu Streams hinzugefügt werden; eine weitere Bearbeitung des Modells ist jedoch nicht möglich.
- **Interaktive Sitzung starten.** Öffnet das interaktive Fenster für die Modellierung (Ausgabe) des Entscheidungslistenviewers, in dem Sie eine Auswahl aus mehreren Alternativen treffen und wiederholt

den Algorithmus mit unterschiedlichen Einstellungen anwenden können, um das Modell progressiv zu erweitern oder zu ändern. Weitere Informationen finden Sie im Thema „Entscheidungslistenviewer" auf Seite 145.

- **Gespeicherte Informationen aus interaktiver Sitzung verwenden.** Startet eine interaktive Sitzung unter Verwendung von zuvor gespeicherten Einstellungen. Interaktive Sitzungen können im Entscheidungslistenviewer mithilfe des Menüs "Generieren" (zur Erstellung eines Modells oder Modellierungsknotens) oder des Menüs "Datei" (zur Aktualisierung des Knotens, von dem aus die Sitzung gestartet wurde) gespeichert werden.

Zielwert. Gibt den Wert des Zielfelds an, in dem das zu modellierende Ergebnis angegeben wird. Wenn beispielsweise das Zielfeld churn (Abwanderung) mit `0 = no` und `1 = yes` codiert ist, geben Sie `1` an, um Regeln festzulegen, mit denen die Datensätze für wahrscheinlich abwandernde Kunden identifiziert werden.

Segmente suchen mit. Gibt an, ob die Suche nach der Zielvariablen auf eine **Hohe Wahrscheinlichkeit** oder eine **Geringe Wahrscheinlichkeit** des Auftretens achten soll. Das Suchen und Ausschließen dieser Segmente kann eine geeignete Methode zum Verbessern des Modells darstellen und ist oft besonders nützlich, wenn das verbleibende Subset eine geringe Wahrscheinlichkeit aufweist.

Maximale Anzahl an Segmenten. Legt die maximale Anzahl von zurückzugebenden Segmenten fest. Die obersten *N* Segmente werden erstellt, wobei das beste Segment die höchste Wahrscheinlichkeit bzw., bei mehreren Modellen mit der gleichen Wahrscheinlichkeit, die höchste Abdeckung aufweist. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

Minimale Segmentgröße. Die beiden Einstellungen unten geben die minimale Segmentgröße vor. Der größere der beiden Werte wird ausgewählt. Wenn z. B. der Prozentwert einer höheren Zahl entspricht als der absolute Wert, wird die prozentuale Einstellung ausgewählt.

- **Als Prozentsatz des vorherigen Segments (%).** Legt die minimale Gruppengröße als Prozentsatz der Datensätze fest. Die zulässige Minimaleinstellung liegt bei 0, die zulässige Maximaleinstellung bei 99,9.
- **Als absoluter Wert (N).** Legt die minimale Gruppengröße als absolute Anzahl der Datensätze fest. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

Segmentregeln.

Maximale Anzahl an Attributen. Legt die maximale Anzahl von Bedingungen pro Segmentregel fest. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

- **Wiederverwendung von Attributen zulassen.** Bei Aktivierung dieser Option können in jedem Zyklus alle Attribute berücksichtigt werden, auch solche, die in vorherigen Zyklen verwendet wurden. Die Bedingungen für ein Segment sind in Zyklen aufgebaut, wobei in jedem Zyklus eine neue Bedingung hinzugefügt wird. Die Anzahl der Zyklen wird über die Einstellung **Maximale Anzahl an Attributen** festgelegt.

Konfidenzintervall für neue Bedingungen (%). Legt das Konfidenzniveau für das Testen der Signifikanz des Segments fest. Diese Einstellung spielt eine wichtige Rolle für die Anzahl der zurückgegebenen Segmente (sofern zutreffend) sowie für die Anzahl von Bedingungen pro Segmentregel. Je höher der Wert, desto kleiner das zurückgegebene Ergebnisset. Die zulässige Minimaleinstellung ist 50. Die zulässige Maximaleinstellung ist 99,9.

Entscheidungslistenknoten - Expertenoptionen

Mit Expertenoptionen können Sie den Modellerstellungsprozess optimieren.

Klassierungsmethode. Die Methode zum Klassieren stetiger Felder (gleiche Anzahl oder gleiche Breite).

Anzahl der Klassen. Die Anzahl der für stetige Felder zu erstellenden Klassen. Die zulässige Minimaleinstellung ist 2, eine Maximaleinstellung ist nicht vorhanden.

Modellsuchbreite: Die maximale Anzahl von Modellergebnissen pro Zyklus, die für den nächsten Zyklus verwendet werden können. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

Regelsuchbreite. Die maximale Anzahl von Regelergebnissen pro Zyklus, die für den nächsten Zyklus verwendet werden können. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

Klassenzusammenführungsfaktor. Der minimale Betrag, um den ein Segment beim Zusammenführen mit dem benachbarten Segment wachsen muss. Die zulässige Minimaleinstellung ist 1,01, eine Maximaleinstellung ist nicht vorhanden.

- **Fehlende Werte in Bedingungen zulassen.** `Wahr`, um den Test `IS MISSING` in Regeln zuzulassen.
- **Zwischenergebnisse verwerfen.** Wenn `Wahr`, werden nur die Endergebnisse des Suchvorgangs zurückgegeben. Ein Endergebnis ist ein Ergebnis, das im Suchvorgang nicht weiter verfeinert wird. Wenn `Falsch`, werden auch Zwischenergebnisse zurückgegeben.

  Maximale Anzahl an Alternativen. Gibt die maximale Anzahl von Alternativen an, die beim Ausführen der Mining-Aufgabe zurückgegeben werden. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

  Beachten Sie, dass die Mining-Aufgabe nur die tatsächliche Anzahl an Alternativen bis zum angegebenen Maximum zurückgibt. Falls als Maximum beispielsweise 100 angegeben wurde und nur 3 Alternativen gefunden werden, werden nur diese 3 angezeigt.

Modellnugget vom Typ "Entscheidungsliste"

Ein Modell besteht aus einer Liste von **Segmenten**, von denen jedes durch eine **Regel** definiert ist, die übereinstimmende Datensätze auswählt. Sie können die Segmente problemlos anzeigen oder ändern, bevor Sie das Modell erzeugen und auswählen, welche Segmente ein- bzw. ausgeschlossen werden sollen. Beim Scoring ergeben Entscheidungslistenmodelle den Wert *Ja* für eingeschlossene Segmente und *$null$* für alles andere, einschließlich des Rests. Durch diese unmittelbare Steuerung des Scorings sind Entscheidungslistenmodelle ideal für das Erstellen von Mailing-Listen und sie sind weit verbreitet im Customer Relationship Management, unter anderen in Callcenter- und Marketinganwendungen.

Wenn Sie einen Stream ausführen, der ein Entscheidungslistenmodell enthält, fügt der Knoten drei neue Felder hinzu; diese enthalten den Score, entweder *1* (d. h. *Ja*) für eingeschlossene Felder oder *$null* für ausgeschlossene Felder, die Wahrscheinlichkeit (Trefferquote) für das Segment, in das der Datensatz fällt, sowie die ID-Nummer des Segments. Die Namen der neuen Felder werden aus dem Namen des prognostizierten Ausgabefelds abgeleitet, dem *$D-* für den Score, *$DP-* für die Wahrscheinlichkeit und *$DI-* für die Segment-ID vorangestellt ist.

Das Modell wird auf der Grundlage des Zielwerts gescort, der zum Zeitpunkt der Modellerstellung angegeben wurde. Sie können Segmente manuell ausschließen, sodass sie beim Scoren den Wert *$null$* erhalten. Wenn Sie beispielsweise eine Suche des Typs "Geringe Wahrscheinlichkeit" ausführen, um Segmente mit unterdurchschnittlichen Trefferraten zu finden, erhalten diese "niedrigen" Segmente den Score *Ja*, sofern Sie sie nicht manuell ausschließen. Falls erforderlich, können Nullen mithilfe eines Ableitungs- oder Füllerknotens zu *Nein* umcodiert werden.

PMML

Ein Entscheidungslistenmodell kann als PMML-Regelsatzmodell (RuleSetModel) mit dem Auswahlkriterium "Erster Treffer" gespeichert werden. Es wird jedoch erwartet, dass alle Regeln denselben Score aufweisen. Um Änderungen im Zielfeld oder Zielwert zu berücksichtigen, können mehrere Regelsetmodelle in einer einzigen Datei gespeichert und nacheinander angewendet werden: Fälle die vom ersten Modell

nicht abgedeckt werden, werden an das zweite weitergeleitet usw. Der Algorithmusname *DecisionList* gibt dieses nicht dem Standard entsprechende Verhalten an und nur Regelsetmodelle mit diesem Namen werden als Entscheidungslistenmodelle erkannt und als solche gescort.

Entscheidungslistenmodellnugget - Einstellungen

Über die Registerkarte "Einstellungen" für ein Modellnugget vom Typ "Entscheidungsliste" können Sie Propensity-Scores ermitteln und die SQL-Optimierung aktivieren oder inaktivieren. Diese Registerkarte ist erst verfügbar, nachdem das Modellnugget zu einem Stream hinzugefügt wurde.

Raw-Propensity-Scores berechnen. Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Diese Werte werden zusätzlich zu den anderen Vorhersage- und Konfidenzwerten angegeben, die gegebenenfalls während des Scorings erstellt werden.

Adjusted-Propensity-Scores berechnen. Raw-Propensity-Scores beruhen ausschließlich auf den Trainingsdaten und sind aufgrund der Neigung vieler Modelle zur übermäßigen Anpassung an die Trainingsdaten möglicherweise zu optimistisch. Bei Adjusted Propensitys wird versucht, dies durch Evaluation der Modellleistung anhand einer Test- bzw. Validierungspartition zu kompensieren. Bei dieser Option muss im Stream ein Partitionsfeld definiert sein und die Adjusted-Propensity-Scores müssen im Modellierungsknoten aktiviert werden, bevor das Modell generiert wird.

Durch Konvertierung in natives SQL scoren. Wenn ausgewählt, wird SQL-Code zum Scoren des Modells innerhalb der Anwendung generiert.

Entscheidungslistenviewer

Mit der einfach bedienbaren, aufgabenbezogenen grafischen Schnittstelle des Entscheidungslistenviewers entfällt die Komplexität der Modellerstellung, da Sie sich nicht mit Details der unteren Ebene der Data-Mining-Verfahren befassen müssen. Sie können Ihre gesamte Aufmerksamkeit auf die Teile der Analyse richten, die einen Benutzereingriff erfordern, wie zum Beispiel das Festlegen von Zielen, das Auswählen von Zielgruppen, das Analysieren der Ergebnisse und das Auswählen des optimalen Modells.

Arbeitsmodellbereich

Der Arbeitsmodellbereich zeigt das aktuelle Modell an, einschließlich der Mining-Aufgaben und anderer Aktionen für das Arbeitsmodell.

ID. Legt die Reihenfolge der Segmente fest. Modellsegmente werden entsprechend der Reihenfolge ihrer ID-Nummer berechnet.

Segmentregeln. Gibt den Segmentnamen und die definierten Segmentbedingungen an. Beim Segmentnamen handelt es sich standardmäßig um den Feldnamen oder um aneinandergereihte Feldnamen, die in Bedingungen verwendet werden und durch Kommas getrennt sind.

Score. Steht für das vorherzusagende Feld, dessen Werte vermutlich mit den Werten anderer Felder (den Prädiktoren) in Beziehung stehen.

Hinweis: Die folgenden Optionen können zur Anzeige im Dialogfeld „Organisieren von Modellmaßen" auf Seite 156 ausgewählt werden.

Abdeckung. Das Kreisdiagramm stellt die Abdeckung der einzelnen Segmente in Bezug zur gesamten Abdeckung visuell dar.

Abdeckung (n). Liste der Abdeckung der einzelnen Segmente in Bezug zur gesamten Abdeckung.

Häufigkeit. Liste der Anzahl der Treffer in Bezug zur Abdeckung. Beispiel: Wenn die Abdeckung bei 79 liegt und die Häufigkeit bei 50, dann bedeutet dies, dass für das ausgewählte Segment 50 von 79 geantwortet haben.

Wahrscheinlichkeit. Zeigt die Wahrscheinlichkeit des Segments an. Beispiel: Wenn die Abdeckung bei 79 liegt und die Häufigkeit bei 50, dann bedeutet dies, dass die Wahrscheinlichkeit für das Segment 63,29 % ist (50 geteilt durch 79).

Fehler. Zeigt den Segmentfehler an.

Die am unteren Rand des Bereichs eingeblendeten Informationen zeigen die Abdeckung, die Häufigkeit und die Wahrscheinlichkeit des gesamten Modells an.

Symbolleiste des Arbeitsmodells

Der Arbeitsmodellbereich bietet über die Symbolleiste die folgenden Funktionen.

Hinweis: Einige Funktionen sind auch verfügbar, wenn Sie mit der rechten Maustaste auf ein Modellsegment klicken.

Tabelle 9. Symbolleistenschaltflächen für das Arbeitsmodell.

Schaltfläche auf der Symbolleiste	**Beschreibung**
	Öffnet das Dialogfeld Neues Modell erstellen, das Optionen zum Erstellen eines neuen Modellnugget bereitstellt.
	Speichert den aktuellen Status der interativen Sitzung. Der Modellierungsknoten der Entscheidungsliste mit den aktuellen Einstellungen wird aktualisiert, darunter Mining-Aufgaben, Modellmomentaufnahmen, Datenauswahl und benutzerdefinierte Maße. Zum Wiederherstellen einer Sitzung in diesem Status aktivieren Sie das Kästchen **Gespeicherte Sitzungsinformationen verwenden** auf der Registerkarte "Modell" des Modellierungsknotens und klicken Sie auf **Ausführen**.
	Zeigt das Dialogfeld "Modellmaße organisieren" an. Weitere Informationen finden Sie im Thema „Organisieren von Modellmaßen” auf Seite 156.
	Zeigt das Dialogfeld "Datenauswahl organisieren" an. Weitere Informationen finden Sie im Thema „Datenauswahl organisieren” auf Seite 151.
	Zeigt die Registerkarte "Momentaufnahmen" an. Weitere Informationen finden Sie im Thema „Registerkarte "Momentaufnahmen"” auf Seite 147.
	Zeigt die Registerkarte "Alternativen" an. Weitere Informationen finden Sie im Thema „Registerkarte "Alternativen"” auf Seite 147.
	Erstellt eine Momentaufnahme der aktuellen Modellstruktur. Momentaufnahmen werden auf der Registerkarte "Momentaufnahmen" angezeigt und in der Regel zum Zweck des Modellvergleichs verwendet.
	Öffnet das Dialogfeld Einfügen von Segmenten, das Optionen zum Erstellen neuer Modellsegmente bereitstellt.
	Öffnet das Dialogfeld Bearbeiten von Segmentregeln, das Optionen zum Hinzufügen von Bedingungen zu Modellsegmenten oder zum Ändern zuvor definierter Modellsegmentbedingungen bereitstellt.
	Verschiebt das ausgewählte Segment in der Modellhierarchie nach oben.

Tabelle 9. Symbolleistenschaltflächen für das Arbeitsmodell (Forts.).

	Verschiebt das ausgewählte Segment in der Modellhierarchie nach unten.
	Löscht das ausgewählte Segment.
	Schließt das ausgewählte Segment entweder in das Modell ein oder davon aus. Wenn es ausgeschlossen wird, werden die Ergebnisse zum Rest hinzugefügt. Dies unterscheidet sich dahin gehend vom Löschen eines Segments, dass Sie die Option haben, es wieder zu aktivieren.

Registerkarte "Alternativen"

Die Registerkarte "Alternativen" wird generiert, wenn Sie auf **Segmente suchen** klicken. Auf ihr werden alle alternativen Mining-Ergebnisse für das im Arbeitsmodellfenster ausgewählte Modell oder Segment aufgeführt.

Wenn eine Alternative in ein Arbeitsmodell umgewandelt werden soll, markieren Sie die entsprechende Alternative und klicken Sie auf **Laden**; das alternative Modell wird im Arbeitsmodellfenster angezeigt.

Hinweis: Die Registerkarte "Alternativen" wird nur angezeigt, wenn Sie die Option **Maximale Anzahl an Alternativen** auf der Registerkarte "Experten" des Modellierungsknotens "Entscheidungsliste" für die Erstellung mehrerer Alternativen eingestellt haben.

Alle generierten Modellalternativen zeigen bestimmte Modellinformationen an:

Name. Jede Alternative wird fortlaufend nummeriert. Die erste Alternative enthält in der Regel die besten Ergebnisse.

Ziel. Gibt den Zielwert an. Beispielsweise 1, was dem Wert "true" entspricht.

Anzahl der Segmente. Die Anzahl der Segmentregeln, die im alternativen Modell verwendet werden.

Abdeckung. Die Abdeckung des alternativen Modells.

Häufigkeit. Anzahl der Treffer in Bezug zur Abdeckung.

Wahrscheinlichkeit. Gibt die prozentuale Wahrscheinlichkeit des alternativen Modells an.

Hinweis: Alternative Ergebnisse werden nicht im Modell gespeichert. Ergebnisse sind nur während der aktiven Sitzung gültig.

Registerkarte "Momentaufnahmen"

Eine Momentaufnahme ist eine Ansicht eines Modells zu einem bestimmten Zeitpunkt. Sie können beispielsweise eine Modellmomentaufnahme erstellen, wenn Sie ein anderes alternatives Modell in das Arbeitsmodellfenster laden möchten und die am aktuellen Modell durchgeführten Arbeiten nicht verlieren möchten. Über die Registerkarte "Momentaufnahmen" werden alle manuell erstellten Momentaufnahmen für eine beliebige Anzahl von Modellzuständen aufgeführt.

Hinweis: Momentaufnahmen werden im Modell gespeichert. Sie sollten eine Momentaufnahme erzeugen, wenn Sie das erste Modell laden. Diese Momentaufnahme nimmt die Originalstruktur des Modells auf und bietet Ihnen die Möglichkeit, jederzeit zum ursprünglichen Modellzustand zurückzukehren. Der Name der generierten Momentaufnahme wird als Zeitmarke angezeigt, die dem Generierungszeitpunkt entspricht.

Erstellen einer Modellmomentaufnahme

1. Wählen Sie für die Anzeige im Arbeitsmodellfenster ein geeignetes Modell oder eine Alternative aus.
2. Nehmen Sie am Arbeitsmodell alle erforderlichen Änderungen vor.
3. Klicken Sie auf **Momentaufnahme erstellen**. Über die Registerkarte "Momentaufnahmen" wird eine neue Momentaufnahme angezeigt.

   Name. Der Name der Momentaufnahme. Sie können den Namen einer Momentaufnahme ändern, indem Sie auf den Namen doppelklicken.

   Ziel. Gibt den Zielwert an. Beispielsweise 1, was dem Wert "true" entspricht.

   Anzahl der Segmente. Die Anzahl der Segmentregeln, die im Modell verwendet werden.

   Abdeckung. Die Abdeckung des Modells.

   Häufigkeit. Anzahl der Treffer in Bezug zur Abdeckung.

   Wahrscheinlichkeit. Gibt die prozentuale Wahrscheinlichkeit des Modells an.
4. Wenn eine Momentaufnahme in ein Arbeitsmodell umgewandelt werden soll, markieren Sie die entsprechende Momentaufnahme und klicken Sie auf **Laden**; die Momentaufnahme wird im Arbeitsmodellfenster angezeigt.
5. Um eine Momentaufnahme zu löschen, klicken Sie auf **Löschen** oder klicken Sie mit der rechten Maustaste auf die Momentaufnahme und wählen Sie im Menü **Löschen**.

Arbeiten mit dem Entscheidungslistenviewer

Ein Modell, das die Reaktionen und das Verhalten von Kunden am besten vorhersagt, wird in mehreren Phasen erstellt. Wenn der Entscheidungslistenviewer gestartet wird, wird das Arbeitsmodell mit den definierten Modellsegmenten und Maßen gefüllt. Anschließend können Mining-Aufgaben ausgeführt, die Segmente/Maße geändert und ein neues Modell oder ein Modellbildungsknoten generiert werden.

Sie können eine oder mehrere Segmentregeln hinzufügen, um ein zufriedenstellendes Modell zu entwickeln. Sie können Segmentregeln zum Modell hinzufügen, indem Sie Mining-Aufgaben ausführen oder indem Sie die Funktion **Segmentregel bearbeiten** anwenden.

Während des Modellbildungsprozesses können Sie die Leistung des Modells bewerten, indem Sie das Modell mit den Maßdaten vergleichen, das Modell in einem Diagramm visuell darstellen oder benutzerdefinierte Excel-Maße erstellen.

Wenn Sie sich hinsichtlich der Qualität des Modells sicher sind, können Sie ein neues Modell erzeugen und es in den IBM SPSS Modeler-Erstellungsbereich oder in die Modellpalette einfügen.

Mining-Aufgaben

Eine **Mining-Aufgabe** ist eine Sammlung von Parametern, die festlegen, wie neue Regeln generiert werden. Einige dieser Parameter können ausgewählt werden, um Ihnen die Möglichkeit zu bieten, das Modell an neue Situationen anzupassen. Eine Aufgabe besteht aus einer Aufgabenvorlage (Typ), einer Zielgröße und einer Erstellungsauswahl (Mining-Dataset).

Im folgenden Abschnitt werden die verschiedenen Vorgänge von Mining-Aufgaben beschrieben:

- „Ausführen von Mining-Aufgaben”
- „Erstellen und Bearbeiten einer Mining-Aufgabe” auf Seite 149
- „Datenauswahl organisieren” auf Seite 151

Ausführen von Mining-Aufgaben: Mit dem Entscheidungslistenviewer können Sie Segmentregeln manuell zu einem Modell hinzufügen, indem Sie Mining-Aufgaben ausführen oder Segmentregeln zwischen Modelle kopieren und einfügen. Eine Mining-Aufgabe enthält Informationen für das Erstellen neuer Segmentregeln (die Data-Mining-Parametereinstellungen, wie die Suchstrategie, Quellenattribute, Breite der Suche, Konfidenzniveau usw.), das vorherzusagende Kundenverhalten und die zu untersuchenden Daten. Das Ziel einer Mining-Aufgabe besteht darin, die bestmöglichen Segmentregeln zu finden.

So erstellen Sie eine Modellsegmentregel, indem Sie eine Mining-Aufgabe ausführen:

1. Klicken Sie auf die Zeile **Rest**. Wenn im Arbeitsmodellfenster bereits Modelle angezeigt werden, können Sie auch eines der Segmente auswählen, um auf der Grundlage des ausgewählten Segments nach neuen Regeln zu suchen. Verwenden Sie nach der Auswahl des Rests oder eines Segments eine der folgenden Methoden, um das Modell oder alternative Modelle zu generieren:

- Wählen Sie im Menü "Extras" die Option **Segmente suchen** aus.
- Klicken Sie mit der rechten Maustaste auf die Zeile **Rest**/das Segment und wählen Sie **Segmente suchen**.
- Klicken Sie im Arbeitsmodellfenster auf die Schaltfläche **Segmente suchen**.

Während die Aufgabe ausgeführt wird, wird der Fortschritt am unteren Rand des Arbeitsbereichs angezeigt. Dort sehen Sie, wenn die Aufgabe abgeschlossen ist. Wie lange eine Aufgabe genau dauert, hängt von der Komplexität der Mining-Aufgabe und der Größe des Datasets ab. Wenn das Ergebnis nur ein einziges Modell umfasst, wird dieses im Arbeitsmodellfenster angezeigt, sobald die Aufgabe erledigt ist. Wenn das Ergebnis jedoch mehrere Modelle enthält, werden diese auf der Registerkarte "Alternativen" angezeigt.

Hinweis: Ein Aufgabenergebnis schließt entweder mit Modellen, ohne Modelle oder mit einem Fehler ab.

Der Vorgang zum Finden neuer Segmentregeln kann so lange wiederholt werden, bis dem Modell keine neuen Regeln mehr hinzugefügt werden. Dies bedeutet, dass alle signifikanten Gruppen oder Kunden gefunden wurden.

Eine Mining-Aufgabe kann auf einem vorhandenen Modellsegment ausgeführt werden. Wenn die Aufgabe nicht das gesuchte Ergebnis liefert, können Sie auf demselben Segment eine andere Mining-Aufgabe ausführen. So können Sie auf der Grundlage des ausgewählten Segments zusätzliche Regeln finden. Segmente, die sich unterhalb des ausgewählten Segments befinden (d. h., die nach dem ausgewählten Segment zum Modell hinzugefügt wurden), werden durch die neuen Segmente ersetzt, da jedes Segment von seinen Vorgängern abhängt.

Erstellen und Bearbeiten einer Mining-Aufgabe: Eine Mining-Aufgabe ist der Mechanismus, der nach der Sammlung von Regeln sucht, die einem Datenmodell zugrunde liegen. Neben den in der ausgewählten Vorlage definierten Suchkriterien definiert eine Aufgabe außerdem das Ziel (die Frage, die der Analyse zugrunde liegt, z. B., wie viele Kunden wahrscheinlich auf ein Mailing reagieren werden) und identifiziert die zu verwendenden Datasets. Das Ziel einer Mining-Aufgabe besteht darin, die bestmöglichen Modelle zu finden.

Erstellen einer Mining-Aufgabe

So erstellen Sie eine Mining-Aufgabe:

1. Wählen Sie das Segment, aus dem Sie zusätzliche Segmentbedingungen ermitteln möchten.
2. Klicken Sie auf **Einstellungen**. Das Dialogfeld "Mining-Aufgabe erstellen/bearbeiten" wird geöffnet. Dieses Dialogfeld bietet Optionen für die Definition der Mining-Aufgabe.
3. Nehmen Sie die erforderlichen Änderungen vor und klicken Sie auf **OK**, um in das Arbeitsmodellfenster zurückzukehren. Der Entscheidungslistenviewer verwendet die Einstellungen als Standardeinstellungen, die für jede Aufgabe ausgeführt werden, bis eine alternative Aufgabe oder Einstellung gewählt wird.
4. Klicken Sie auf **Segmente suchen**, um die Mining-Aufgabe auf dem ausgewählten Segment zu starten.

Bearbeiten einer Mining-Aufgabe

Das Dialogfeld "Mining-Aufgabe erstellen/bearbeiten" bietet Optionen zum Definieren einer neuen Mining-Aufgabe oder zum Bearbeiten einer vorhandenen Mining-Aufgabe.

Die meisten für Mining-Aufgaben verfügbaren Parameter entsprechen denen für den Entscheidungslistenknoten. Die Ausnahmen werden unten gezeigt. Weitere Informationen finden Sie im Thema „Entscheidungslistenmodell - Optionen" auf Seite 142.

Einstellungen laden: Wenn Sie mehrere Mining-Aufgaben erstellt haben, wählen Sie die erforderliche Aufgabe aus.

Neu... Klicken Sie auf diese Option, um eine neue Mining-Aufgabe basierend auf den Einstellungen der derzeit angezeigten Aufgabe zu erstellen.

Ziel

Zielfeld: Steht für das Feld, das Sie vorhersagen möchten, dessen Wert vermutlich mit den Werten anderer Felder (den Prädiktoren) in Beziehung steht.

Zielwert. Gibt den Wert des Zielfelds an, in dem das zu modellierende Ergebnis angegeben wird. Wenn beispielsweise das Zielfeld churn (Abwanderung) mit `0 = no` und `1 = yes` codiert ist, geben Sie `1` an, um Regeln festzulegen, mit denen die Datensätze für wahrscheinlich abwandernde Kunden identifiziert werden.

Einfache Einstellungen (SimpleSettings)

Maximale Anzahl an Alternativen. Gibt die Anzahl von Alternativen an, die beim Ausführen der Mining-Aufgabe angezeigt werden. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

Experteneinstellungen (ExpertSettings)

Bearbeiten... Öffnet das Dialogfeld **Erweiterte Parameter bearbeiten**, in dem Sie die erweiterten Einstellungen definieren können. Weitere Informationen finden Sie im Thema „Erweiterte Parameter bearbeiten".

Data

Erstellungsauswahl. Bietet Optionen zur Angabe des Evaluierungsmaßes, das der Entscheidungslistenviewer analysieren soll, um neue Regeln zu finden. Die aufgeführten Evaluierungsmaße werden im Dialogfeld "Datenauswahl organisieren" erstellt/bearbeitet.

Verfügbare Felder. Bietet Optionen zur Anzeige aller Felder oder zur manuellen Auswahl der anzuzeigenden Felder.

Bearbeiten... Wenn die Option **Benutzerdefiniert** ausgewählt ist, wird das Dialogfeld **Verfügbare Felder anpassen** geöffnet, in dem Sie auswählen können, welche Felder als Segmentattribute verfügbar sind, die von der Mining-Aufgabe gefunden wurden. Weitere Informationen finden Sie im Thema „Verfügbare Felder anpassen" auf Seite 151.

Erweiterte Parameter bearbeiten: Das Dialogfeld "Erweiterte Parameter bearbeiten" bietet die folgenden Konfigurationsoptionen.

Klassierungsmethode. Die Methode zum Klassieren stetiger Felder (gleiche Anzahl oder gleiche Breite).

Anzahl der Klassen. Die Anzahl der für stetige Felder zu erstellenden Klassen. Die zulässige Minimaleinstellung ist 2, eine Maximaleinstellung ist nicht vorhanden.

Modellsuchbreite: Die maximale Anzahl von Modellergebnissen pro Zyklus, die für den nächsten Zyklus verwendet werden können. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

Regelsuchbreite. Die maximale Anzahl von Regelergebnissen pro Zyklus, die für den nächsten Zyklus verwendet werden können. Die zulässige Minimaleinstellung liegt bei 1, eine Maximaleinstellung ist nicht vorhanden.

Klassenzusammenführungsfaktor. Der minimale Betrag, um den ein Segment beim Zusammenführen mit dem benachbarten Segment wachsen muss. Die zulässige Minimaleinstellung ist 1,01, eine Maximaleinstellung ist nicht vorhanden.

- **Fehlende Werte in Bedingungen zulassen.** `Wahr`, um den Test `IS MISSING` in Regeln zuzulassen.
- **Zwischenergebnisse verwerfen.** Wenn `Wahr`, werden nur die Endergebnisse des Suchvorgangs zurückgegeben. Ein Endergebnis ist ein Ergebnis, das im Suchvorgang nicht weiter verfeinert wird. Wenn `Falsch`, werden auch Zwischenergebnisse zurückgegeben.

Verfügbare Felder anpassen: Im Dialogfeld "Verfügbare Felder anpassen" können Sie auswählen, welche Felder als Segmentattribute verfügbar sind, die von der Mining-Aufgabe gefunden wurden.

Verfügbar. Führt die Felder auf, die aktuell als Segmentattribute verfügbar sind. Um Felder aus der Liste zu entfernen, wählen Sie die entsprechenden Felder aus und klicken Sie auf **Entfernen>>**. Die ausgewählten Felder werden aus der Liste der verfügbaren Felder in die Liste der nicht verfügbaren Felder verschoben.

Nicht verfügbar. Führt die Felder auf, die nicht als Segmentattribute verfügbar sind. Um diese Felder in die Liste der verfügbaren Felder aufzunehmen, wählen Sie die entsprechenden Felder aus und klicken Sie auf **<< Hinzufügen**. Die ausgewählten Felder werden aus der Liste der nicht verfügbaren Felder in die Liste der verfügbaren Felder verschoben.

Datenauswahl organisieren: Durch das Organisieren einer Datenauswahl (einem Mining-Dataset) können Sie festlegen, welche Evaluierungsmaße der Entscheidungslistenviewer analysieren soll, um neue Regeln zu suchen, und welche Datenauswahl als Grundlage für die Maße verwendet wird.

So organisieren Sie eine Datenauswahl:

1. Wählen Sie im Menü "Extras" die Option **Datenauswahl organisieren** oder klicken Sie mit der rechten Maustaste auf ein Segment und wählen Sie die Option. Das Dialogfeld "Datenauswahl organisieren" wird geöffnet.

   Hinweis: Mit dem Dialogfeld "Datenauswahl organisieren" können Sie auch eine vorhandene Datenauswahl bearbeiten oder löschen.
2. Klicken Sie auf die Schaltfläche **Neue Datenauswahl hinzufügen**. Zur vorhandenen Tabelle wird ein neuer Datenauswahleintrag hinzugefügt.
3. Klicken Sie auf **Name** und geben Sie für die Auswahl einen geeigneten Namen ein.
4. Klicken Sie auf **Partition** und wählen Sie einen Partitionstyp aus.
5. Klicken Sie auf **Bedingung** und wählen Sie eine Bedingungsoption aus. Wenn **Angeben** ausgewählt ist, wird das Dialogfeld "Auswahlbedingung angeben" geöffnet, das Optionen zur Angabe bestimmter Feldbedingungen enthält.
6. Definieren Sie die entsprechende Bedingung und klicken Sie auf **OK**.

Die Datenauswahl ist im Dialogfeld "Mining-Aufgabe erstellen/bearbeiten" in der Dropdown-Liste "Erstellungsauswahl" verfügbar. In der Liste können Sie auswählen, welches Evaluierungsmaß für eine bestimmte Mining-Aufgabe verwendet wird.

Segmentregeln

Modellsegmentregeln finden Sie, indem Sie eine Mining-Aufgabe ausführen, die auf einer Aufgabenvorlage basiert. Sie können einem Modell manuell Regeln hinzufügen, wenn Sie die Funktionen "Segment einfügen" oder "Segmentregel bearbeiten" verwenden.

Wenn Sie neue Segmentregeln mithilfe einer Mining-Aufgabe suchen, werden die Ergebnisse (sofern vorhanden) auf der Registerkarte "Viewer" des Dialogfelds "Interaktive Liste" angezeigt. Sie können Ihr Modell schnell verfeinern, indem Sie eines der alternativen Ergebnisse aus dem Dialogfeld "Alben modellieren" auswählen und auf **Laden** klicken. Auf diese Art können Sie so lange mit verschiedenen Ergebnissen experimentieren, bis Sie ein Modell erstellen können, das Ihre optimale Zielgruppe genau beschreibt.

Einfügen von Segmenten: Sie können einem Modell manuell Regeln hinzufügen, wenn Sie die Funktion "Segment einfügen" verwenden.

So fügen Sie eine Segmentregelbedingung zu einem Modell hinzu:

1. Wählen Sie im Dialogfeld "Interaktive Liste" eine Position aus, an der Sie ein neues Segment hinzufügen möchten. Das neue Segment wird direkt über dem ausgewählten Segment eingefügt.
2. Wählen Sie im Menü "Bearbeiten" **Segment einfügen** aus oder greifen Sie darauf zu, indem Sie mit der rechten Maustaste auf ein Segment klicken.

   Das Dialogfeld "Segment einfügen" wird geöffnet, in dem Sie die neuen Segmentregelbedingungen einfügen können.
3. Klicken Sie auf **Einfügen**. Das Dialogfeld "Bedingung einfügen" wird geöffnet, in dem Sie die Attribute für die neue Regelbedingung definieren können.
4. Wählen Sie in den Dropdown-Listen ein Feld und einen Operator aus.

   Hinweis: Wenn Sie den Operator **Nicht in** auswählen, funktioniert die ausgewählte Bedingung als Ausschlussbedingung und wird im Dialogfeld "Regel einfügen" rot angezeigt. Wenn beispielsweise die Bedingung `Region = 'STADT'` in Rot angezeigt wird, bedeutet dies, dass `STADT` vom Ergebnisset ausgeschlossen ist.
5. Geben Sie mindestens einen Wert ein oder klicken Sie auf das Symbol **Wert einfügen**, um das Dialogfeld "Wert einfügen" anzuzeigen. In diesem Dialogfeld können Sie einen Wert auswählen, der für das ausgewählte Feld definiert ist. Das Eingabefeld **verheiratet** würde beispielsweise die Werte **ja** und **nein** liefern.
6. Klicken Sie auf **OK**, um zum Dialogfeld "Segment einfügen" zurückzukehren. Klicken Sie ein zweites Mal auf **OK**, um das erstellte Segment dem Modell hinzuzufügen.

Das neue Segment wird an der angegebenen Position im Modell angezeigt.

Bearbeiten von Segmentregeln: Mit der Funktion "Segmentregel bearbeiten" können Sie Segmentregelbedingungen hinzufügen, ändern oder löschen.

So ändern Sie eine Segmentregelbedingung:

1. Wählen Sie das Modellsegment aus, das Sie bearbeiten möchten.
2. Wählen Sie im Menü "Bearbeiten" die Option **Segmentregel bearbeiten** oder klicken Sie mit der rechten Maustaste auf die Regel, um auf diese Option zuzugreifen.

   Das Dialogfeld "Segmentregel bearbeiten" wird geöffnet.
3. Wählen Sie die entsprechende Bedingung aus und klicken Sie auf **Bearbeiten**.

   Das Dialogfeld "Bedingung bearbeiten" wird geöffnet, in dem Sie die Attribute für die ausgewählte Regelbedingung definieren können.
4. Wählen Sie in den Dropdown-Listen ein Feld und einen Operator aus.

Hinweis: Wenn Sie den Operator **Nicht in** auswählen, funktioniert die ausgewählte Bedingung als Ausschlussbedingung und wird im Dialogfeld "Segmentregel bearbeiten" rot angezeigt. Wenn beispielsweise die Bedingung `Region = 'STADT'` in Rot angezeigt wird, bedeutet dies, dass `STADT` vom Ergebnisset ausgeschlossen ist.

5. Geben Sie mindestens einen Wert ein oder klicken Sie auf die Schaltfläche **Wert einfügen**, um das Dialogfeld "Wert einfügen" anzuzeigen. In diesem Dialogfeld können Sie einen Wert auswählen, der für das ausgewählte Feld definiert ist. Das Eingabefeld **verheiratet** würde beispielsweise die Werte **ja** und **nein** liefern.
6. Klicken Sie auf **OK**, um zum Dialogfeld "Segmentregel bearbeiten" zurückzukehren. Klicken Sie ein zweites Mal auf **OK**, um zum Arbeitsmodell zurückzukehren.

Das ausgewählte Modell wird mit den aktualisierten Regelbedingungen angezeigt.

Löschen von Segmentregelbedingungen: **So löschen Sie eine Segmentregelbedingung:**

1. Wählen Sie das Modellsegment aus, das die Regelbedingungen enthält, die Sie löschen möchten.
2. Wählen Sie im Menü "Bearbeiten" die Option **Segmentregel bearbeiten** oder klicken Sie mit der rechten Maustaste auf das Segment, um auf diese Option zuzugreifen.

   Das Dialogfeld "Segmentregel bearbeiten" wird geöffnet, in dem Sie eine oder mehrere Segmentregelbedingungen löschen können.
3. Wählen Sie die entsprechende Regelbedingung aus und klicken Sie auf **Löschen**.
4. Klicken Sie auf **OK**.

Das Löschen einer oder mehrerer Segmentregelbedingungen bewirkt, dass die Maßmetriken im Arbeitsmodellbereich aktualisiert werden.

Kopieren von Segmenten: Der Entscheidungslistenviewer bietet ein bequemes Verfahren zum Kopieren von Modellsegmenten. Wenn Sie ein Segment eines Modells auf ein anderes Modell anwenden möchten, dann kopieren Sie das Segment einfach in einem Modell (oder schneiden Sie es aus) und fügen Sie es in ein anderes Modell ein. Sie können auch ein Segment eines Modells kopieren, das im Bereich "Alternative Vorschau" angezeigt wird, und es in das Modell einfügen, das im Arbeitsmodellfenster angezeigt wird. Diese Funktionen zum Ausschneiden, Kopieren und Einfügen verwenden zum Speichern und Abrufen temporärer Daten die Zwischenablage des Systems. Dies bedeutet, dass die Bedingungen und das Ziel in die Zwischenablage kopiert werden. Die Inhalte der Zwischenablage sind nicht für die Verwendung im Entscheidungslistenviewer reserviert und können auch in andere Anwendungen eingefügt werden. Wenn der Inhalt der Zwischenablage beispielsweise in einen Texteditor eingefügt wird, werden die Bedingungen und das Ziel im XML-Format eingefügt.

So kopieren Sie Modellsegmente oder schneiden diese aus:

1. Wählen Sie das Modellsegment aus, das Sie in einem anderen Modell verwenden möchten.
2. Wählen Sie im Menü "Bearbeiten" die Option **Kopieren** (oder **Ausschneiden**) oder klicken Sie mit der rechten Maustaste auf das Modellsegment und wählen Sie **Kopieren** oder **Ausschneiden**.
3. Öffnen Sie das entsprechende Modell (in das Sie das Modellsegment einfügen möchten).
4. Wählen Sie eines der Modellsegmente und klicken Sie auf **Einfügen**.

Hinweis: Anstatt der Befehle **Ausschneiden**, **Kopieren** und **Einfügen** können Sie auch die folgenden Tastenkombinationen verwenden: **Strg+X**, **Strg+C** und **Strg+V**.

Das kopierte (oder ausgeschnittene) Segment wird oberhalb des ausgewählten Modellsegments eingefügt. Die Maße des eingefügten Segments und der darunterliegenden Segmente werden (neu) berechnet.

Hinweis: Beide Modelle in dieser Prozedur müssen auf derselben Modellvorlage basieren und dasselbe Ziel enthalten. Andernfalls wird eine Fehlernachricht angezeigt.

Alternative Modelle: Wenn mehrere Ergebnisse vorhanden sind, zeigt die Registerkarte "Alternativen" die Ergebnisse der einzelnen Mining-Aufgaben an. Jedes Ergebnis besteht aus den Bedingungen der ausgewählten Daten, die am stärksten mit dem Ziel übereinstimmen, sowie allen Alternativen, die als ausreichend gut eingestuft werden. Die Gesamtzahl der angezeigten Alternativen hängt von den Suchkriterien ab, die im Analysevorgang verwendet werden.

So zeigen Sie alternative Modelle an:

1. Klicken Sie auf der Registerkarte "Alternativen" auf ein alternatives Modell. Im Fenster "Alternative Vorschau" werden die alternativen Modellsegmente angezeigt bzw. sie ersetzen die aktuellen Modellsegmente.
2. Um im Arbeitsmodellfenster mit einem alternativen Modell zu arbeiten, wählen Sie das Modell aus und klicken Sie im Fenster "Alternative Vorschau" auf **Laden** oder klicken Sie mit der rechten Maustaste auf der Registerkarte "Alternativen" auf den Namen einer Alternative und wählen Sie **Laden**.

Hinweis: Alternative Modelle werden beim Erstellen eines neuen Modells nicht gespeichert.

Anpassen eines Modells

Daten sind nicht statisch. Kunden ziehen um, heiraten und ändern ihren Arbeitsplatz. Produkte fallen aus dem Marktfokus und veralten.

Der Entscheidungslistenviewer bietet Fachanwendern die Flexibilität, Modelle einfach und schnell an neue Situationen anzupassen. Sie können ein Modell ändern, indem Sie es bearbeiten, mit Prioritäten versehen, löschen oder bestimmte Modellsegmente inaktivieren.

Prioritäten für Segmente zuweisen: Sie können Segmentregeln in beliebiger Reihenfolge eine Rangfolge zuweisen. Standardmäßig werden Modellsegmente in der Reihenfolge ihrer Priorität angezeigt, wobei das erste Segment die höchste Priorität besitzt. Wenn Sie einem oder mehreren Segmenten eine andere Priorität zuweisen, wird das Modell entsprechend geändert. Sie können das Modell an die Anforderungen anpassen, indem Sie Segmente in eine höhere oder niedrigere Prioritätsposition verschieben.

So weisen Sie Modellsegmenten Prioritäten zu:

1. Wählen Sie das Modellsegment aus, dem Sie eine andere Priorität zuweisen möchten.
2. Klicken Sie in der Symbolleiste des Arbeitsmodellfensters auf eine der beiden Pfeilschaltflächen, um das ausgewählte Modellsegment in der Liste nach oben oder nach unten zu verschieben.

Nach dem Zuweisen der Priorität werden alle vorherigen Bewertungsergebnisse neu berechnet und die neuen Werte werden angezeigt.

Löschen von Segmenten: So löschen Sie eines oder mehrere Segmente:

1. Wählen Sie ein Modellsegment aus.
2. Wählen Sie im Menü "Bearbeiten" die Option **Segment löschen** oder klicken Sie in der Symbolleiste des Arbeitsmodellfensters auf die Schaltfläche "Löschen".

Die Maße werden für das geänderte Modell neu berechnet und das Modell entsprechend geändert.

Ausschließen von Segmenten: Wenn Sie nach bestimmten Gruppen suchen, werden Sie Geschäftsaktionen wahrscheinlich auf der Grundlage einer Auswahl von Modellsegmenten entscheiden. Wenn Sie ein Modell bereitstellen, können Sie auswählen, Segmente innerhalb des Modells auszuschließen. Ausgeschlossene Segmente werden als Nullwerte gescort. Wenn ein Segment ausgeschlossen wird, bedeutet dies nicht, dass es nicht verwendet wird. Es bedeutet, dass alle Datensätze, die dieser Regel entsprechen, aus der Mailing-Liste ausgeschlossen werden. Die Regel wird weiterhin angewendet, aber in anderer Form.

So schließen Sie bestimmte Modellsegmente aus:

1. Wählen Sie im Arbeitsmodellfenster ein Segment aus.

2. Klicken Sie in der Symbolleiste des Arbeitsmodellfensters auf die Schaltfläche **Zwischen Segmentausschluss umschalten**. In der ausgewählten Spalte "Ziel" des ausgewählten Segments wird nun **Ausgeschlossen** angezeigt.

Hinweis: Im Gegensatz zu gelöschten Segmenten sind ausgeschlossene Segmente weiterhin für die Wiederverwendung im endgültigen Modell verfügbar. Ausgeschlossene Segmente wirken sich auf die Diagrammergebnisse aus.

Zielwert ändern: Im Dialogfeld "Zielwert ändern" können Sie den Zielwert des aktuellen Zielfelds ändern.

Momentaufnahmen und Sitzungsergebnisse mit einem anderen Zielwert als dem des Arbeitsmodells werden dahin gehend gekennzeichnet, dass der Tabellenhintergrund der entsprechenden Spalte gelb dargestellt wird. Damit wird angezeigt, dass das Momentaufnahme-/Sitzungsergebnis veraltet ist.

Im Dialogfeld **Mining-Aufgabe erstellen/bearbeiten** wird der Zielwert für das aktuelle Arbeitsmodell angezeigt. Der Zielwert wird nicht mit der Mining-Aufgabe gespeichert. Er wird stattdessen dem Wert des Arbeitsmodells entnommen.

Wenn Sie ein gespeichertes Modell als Arbeitsmodell übernehmen, das einen anderen Zielwert besitzt als das aktuelle Arbeitsmodell (indem Sie beispielsweise ein alternatives Ergebnis oder eine Kopie einer Momentaufnahme bearbeiten), wird der Zielwert des gespeicherten Modells dahingehend geändert, dass er mit dem des Arbeitsmodells übereinstimmt (der im Arbeitsmodellfenster angezeigte Zielwert ändert sich nicht). Die Modellmetriken werden mit dem neuen Ziel erneut bewertet.

Neues Modell erzeugen

Das Dialogfeld "Neues Modell erzeugen" bietet Optionen für die Benennung des Modells und zur Festlegung, wo der neue Knoten erstellt werden soll.

Modellname. Wählen Sie die Option **Benutzerdefiniert** aus, um den automatisch generierten Namen anzupassen oder um einen eindeutigen Namen für den Knoten zu erstellen, der im Streamerstellungsbereich angezeigt wird.

Knoten erstellen auf. Wenn Sie **Erstellungsbereich** auswählen, wird das neue Modell im Arbeitserstellungsbereich platziert. Wenn Sie **Generierte Modelle** auswählen, wird das neue Modell in der Modellpalette platziert. Wenn Sie **Beide** auswählen, wird das neue Modell sowohl im Erstellungsbereich als auch in der Modellpalette platziert.

Status der interaktiven Sitzung einbeziehen. Wenn diese Option aktiviert ist, wird der Status der interaktiven Sitzung im generierten Modell erhalten. Wenn Sie später aus dem Modell einen Modellbildungsknoten generieren, wird der Status übernommen und für die Initialisierung der interaktiven Sitzung verwendet. Unabhängig davon, ob die Option aktiviert ist, scort das Modell neue Daten identisch. Wenn die Option nicht ausgewählt ist, ist das Modell immer noch in der Lage, einen Erstellungsknoten zu erstellen, dabei wird es sich aber um einen allgemeineren Erstellungsknoten handeln, der eine neue interaktive Sitzung startet, anstatt die alte Sitzung dort wieder aufzunehmen, wo sie verlassen wurde. Wenn Sie die Knoteneinstellungen ändern und dann einen gespeicherten Status ausführen, werden Ihre geänderten Einstellungen aber ignoriert und die Einstellungen des gespeicherten Status angewendet.

Hinweis: Die Standardmetriken sind die einzigen Metriken, die im Modell gespeichert werden. Zusätzliche Metriken werden zusammen mit dem interaktiven Status gespeichert. Das generierte Modell repräsentiert nicht den gespeicherten Status der interaktiven Mining-Aufgabe. Nach dem Starten des Entscheidungslistenviewers werden die ursprünglich im Viewer vorgenommenen Einstellungen angezeigt.

Weitere Informationen finden Sie im Thema „Erneutes Erzeugen eines Modellierungsknotens” auf Seite 49.

Modellauswertung

Eine erfolgreiche Modellbildung erfordert vor der Implementierung des Modells in einer Produktionsumgebung eine sorgfältige Auswertung des Modells. Der Entscheidungslistenviewer bietet eine gewisse Anzahl von statistischen und betriebswirtschaftlichen Methoden, die für die Bewertung der Auswirkung des Modells in der Realität verwendet werden können. Dazu gehören Gewinndiagramme und eine vollständige Interoperabilität mit Excel, wodurch für die Bewertung der Auswirkung der Bereitstellung Kosten-/Nutzen-Szenarios simuliert werden können.

Modelle können auf folgende Arten ausgewertet werden:

- Mithilfe der im Entscheidungslistenviewer vorhandenen statistischen und betriebswirtschaftlichen Methoden (Wahrscheinlichkeit, Häufigkeit).
- Evaluierung von aus Microsoft Excel importierten Maßen.
- Visualisierung des Modells mithilfe eines Gewinndiagramms.

Organisieren von Modellmaßen: Der Entscheidungslistenviewer bietet Optionen für die Definition von Maßen, die als Spalten berechnet und angezeigt werden. Jedes Segment kann die standardmäßige Abdeckung, Häufigkeit, Wahrscheinlichkeit und Fehlermaße als Spalten enthalten. Sie können außerdem neue Maße erstellen, die als Spalten angezeigt werden.

Definieren von Modellmaßen

So fügen Sie zu Ihrem Modell ein Maß hinzu oder definieren ein vorhandenes Maß:

1. Wählen Sie im Menü "Extras" die Option **Modellmaße organisieren** oder klicken Sie mit der rechten Maustaste auf das Modell, um diese Option auszuwählen. Das Dialogfeld "Modellmaße organisieren" wird geöffnet.
2. Klicken Sie auf die Schaltfläche **Neues Modellmaß hinzufügen** (rechts von der Spalte "Anzeigen"). In der Tabelle wird ein neues Maß angezeigt.
3. Geben Sie einen Namen für das Maß sowie den entsprechenden Typ, die Anzeigeoption und die Auswahl ein. In der Spalte "Anzeigen" wird angegeben, ob das Maß für das Arbeitsmodell angezeigt wird. Wenn Sie ein vorhandenes Maß definieren, wählen Sie eine entsprechende Metrik und Auswahl und geben Sie an, ob das Maß für das Arbeitsmodell angezeigt wird.
4. Klicken Sie auf **OK**, um zum Arbeitsbereich des Entscheidungslistenviewers zurückzukehren. Wenn für das neue Maß die Spalte "Anzeigen" aktiviert wurde, wird das neue Maß für das Arbeitsmodell angezeigt.

Benutzerdefinierte Metriken in Excel

Weitere Informationen finden Sie im Thema „Auswertung in Excel”.

Aktualisieren von Maßen: In bestimmten Fällen kann es notwendig sein, die Modellmaße neu zu berechnen, beispielsweise wenn Sie ein vorhandenes Modell auf ein neues Kundenset anwenden.

So werden Modellmaße neu berechnet (aktualisiert):

Wählen Sie im Menü "Bearbeiten" die Option **Alle Maße aktualisieren**.

oder

Drücken Sie die Taste F5.

Alle Maße werden neu berechnet und die neuen Werte werden im Arbeitsmodell angezeigt.

Auswertung in Excel: Der Entscheidungslistenviewer kann in Microsoft Excel integriert werden, wodurch Sie Ihre eigenen Wertberechnungen und Gewinnformeln direkt im Modellerstellungsprozess ver-

wenden können, um Kosten-Nutzen-Szenarios zu simulieren. Über die Verknüpfung mit Excel können Sie Daten in Excel exportieren, wo diese verwendet werden können, um Präsentationsdiagramme zu erstellen, benutzerdefinierte Maße, wie beispielsweise komplexe Profit- und ROI-Maße, zu berechnen und sie während der Modellerstellung im Entscheidungslistenviewer anzuzeigen.

Hinweis: Damit Sie mit einem Excel-Arbeitsblatt arbeiten können, muss der Experte für analytisches CRM Konfigurationsinformationen für die Synchronisierung des Entscheidungslistenviewers mit Microsoft Excel definieren. Die Konfiguration befindet sich in der Tabelle einer Excel-Datei. Sie gibt an, welche Informationen vom Entscheidungslistenviewer an Excel und umgekehrt übertragen werden.

Die folgenden Schritte sind nur dann gültig, wenn MS Excel installiert ist. Wenn Excel nicht installiert ist, werden die Optionen für die Synchronisierung von Modellen mit Excel nicht angezeigt.

So synchronisieren Sie Modelle mit MS Excel:

1. Öffnen Sie das Modell, führen Sie eine interaktive Sitzung aus und wählen Sie im Menü "Extras" die Option **Modellmaße organisieren** aus.
2. Wählen Sie **Ja** für die Option **Benutzerdefinierte Metriken in Excel berechnen**. Das Feld **Arbeitsmappe** wird aktiviert, wodurch Sie eine vorkonfigurierte Excel-Arbeitsmappenvorlage auswählen können.
3. Klicken Sie auf die Schaltfläche **Verbindung mit Excel herstellen**. Das Dialogfeld "Öffnen" wird angezeigt, in dem Sie in Ihrem lokalen Dateisystem oder im Netzdateisystem zum Speicherort der vorkonfigurierten Vorlage navigieren können.
4. Wählen Sie die entsprechende Excel-Vorlage aus und klicken Sie auf **Öffnen**. Die ausgewählte Excel-Vorlage wird gestartet. Wechseln Sie mithilfe der Windows-Taskleiste (oder durch Drücken von Alt-Tab) zurück zum Dialogfeld "Eingaben für benutzerdefinierte Maße".
5. Wählen Sie die Zuordnungen zwischen den in der Excel-Vorlage definierten Metriknamen und den Metriknamen des Modells aus und klicken Sie auf **OK**.

Nachdem die Verknüpfung hergestellt ist, startet Excel mit der vorkonfigurierten Excel-Vorlage, die die Modellregeln in einer Tabelle anzeigt. Die in Excel berechneten Ergebnisse werden im Entscheidungslistenviewer als neue Spalten angezeigt.

Hinweis: Excel-Metriken werden nicht beim Speichern des Modells gespeichert. Die Metriken gelten nur während der aktiven Sitzung. Sie können jedoch Momentaufnahmen erstellen, die Excel-Metriken enthalten. Die in Momentaufnahme-Ansichten gespeicherten Excel-Metriken eignen sich für den historischen Vergleich und werden beim erneuten Öffnen nicht aktualisiert. Weitere Informationen finden Sie im Thema „Registerkarte "Momentaufnahmen"" auf Seite 147. Die Excel-Metriken werden erst in der Momentaufnahme angezeigt, wenn Sie erneut eine Verbindung zur Excel-Vorlage herstellen.

MS Excel-Integration - Setup: Die Integration des Entscheidungslistenviewers und Microsoft Excel erfolgt über vorkonfigurierte Excel-Tabellenvorlagen. Die Vorlage besteht aus drei Arbeitsblättern:

Modellmaße. Zeigt die importierten Entscheidungslistenviewer-Maße, die benutzerdefinierten Excel-Maße und die Summen der Berechnungen (die im Arbeitsblatt "Einstellungen" definiert sind).

Einstellungen. Enthält die Variablen, mit denen Berechnungen auf der Grundlage der importierten Entscheidungslistenviewer-Maße und der benutzerdefinierten Excel-Maße erstellt werden.

Konfiguration. Enthält Optionen, mit denen festgelegt wird, welche Maße aus dem Entscheidungslistenviewer importiert werden, und mit denen benutzerdefinierte Excel-Maße definiert werden.

WARNUNG: Die Struktur des Arbeitsblatt "Konfiguration" ist streng definiert. Bearbeiten Sie **KEINESFALLS** Zellen im grau schattierten Bereich.

- **Metriken aus Modell.** Gibt an, welche Entscheidungslistenviewer-Metriken in den Berechnungen verwendet werden.

- **Metriken an Modell.** Gibt an, welche von Excel generierten Metriken an den Entscheidungslistenviewer zurückgegeben werden. Die von Excel generierten Metriken werden im Entscheidungslistenviewer als neue Maßspalten angezeigt.

Hinweis: Excel-Metriken werden beim Erstellen eines neuen Modells nicht gespeichert. Die Metriken sind nur während der aktiven Sitzung gültig.

Ändern der Modellmaße: In den folgenden Beispielen werden verschiedene Möglichkeiten zum Ändern von Modellmaßen erläutert:

- Ändern eines bestehenden Maßes.
- Importieren eines weiteren Standardmaßes aus dem Modell.
- Exportieren eines weiteren Standardmaßes in das Modell.

Ändern eines bestehenden Maßes

1. Öffnen Sie die Vorlage und wählen Sie das Arbeitsblatt "Konfiguration" aus.
2. Bearbeiten Sie einen beliebigen Wert für **Name** oder **Beschreibung**, indem Sie ihn markieren und überschreiben.

   Beachten Sie: Um ein Maß zu ändern, beispielsweise, um den Benutzer statt zur Eingabe der Häufigkeit zur Eingabe der Wahrscheinlichkeit aufzufordern, müssen Sie lediglich den Namen und die Beschreibung unter **Metriken aus Modell** ändern. Die Änderung wird dann im Modell angezeigt und der Benutzer kann das entsprechende Maß für die Zuordnung auswählen.

Importieren eines weiteren Standardmaßes aus dem Modell

1. Öffnen Sie die Vorlage und wählen Sie das Arbeitsblatt "Konfiguration" aus.
2. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Tools > **Schutz** > **Schutz des Arbeitsblatts aufheben**
3. Wählen Sie Zelle A5, die gelb schattiert ist, und das Wort **End** enthält.
4. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Einfügen > **Zeilen**
5. Geben Sie **Name** und **Beschreibung** des neuen Maßes ein. Beispielsweise könnten Sie **Error** (Fehler) und **Error associated with segment** (Zum Segment gehöriger Fehler) eingeben.
6. Geben Sie in Zelle C5 die Formel **=COLUMN('Model Measures'!N3)** ein.
7. Geben Sie in Zelle D5 die Formel **=ROW('Model Measures'!N3)+1** ein.

   Durch diese Formeln wird das neue Maß in Spalte N des Arbeitsblatts "Model Measures" angezeigt, die derzeit leer ist.
8. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Tools > **Schutz** > **Arbeitsblatt schützen**
9. Klicken Sie auf **OK**.
10. Vergewissern Sie sich auf dem Arbeitsblatt "Model Measures", dass Zelle N3 **Error** als Titel für die neue Spalte aufweist.
11. Wählen Sie die gesamte Spalte N aus.
12. Wählen Sie die folgenden Befehle aus den Menüs aus:

    Format > **Zellen**
13. Standardmäßig haben alle Zellen die Zahlenkategorie **Allgemein**. Klicken Sie auf **Prozentsatz**, um die Darstellungsart der Zahlen zu ändern. Dadurch können Sie Ihre Zahlen besser in Excel überprüfen. Außerdem können Sie dadurch die Daten auf andere Weise nutzen, beispielsweise als Ausgabe für ein Diagramm.
14. Klicken Sie auf **OK**.

15. Speichern Sie das Arbeitsblatt als Excel 2003-Vorlage, mit einem eindeutigen Namen und der Dateierweiterung *.xlt*. Um die neue Vorlage leichter wieder auffinden zu können, sollten Sie sie am Speicherort der vorkonfigurierten Vorlage auf Ihrem lokalen System oder Netzdateisystem speichern.

Exportieren eines weiteren Standardmaßes in das Modell

1. Öffnen Sie die Vorlage, zu der Sie die Spalte "Error" (Fehler) im vorherigen Beispiel hinzugefügt haben; wählen Sie das Arbeitsblatt "Konfiguration" aus.
2. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Tools > **Schutz** > **Schutz des Arbeitsblatts aufheben**
3. Wählen Sie Zelle A14, die gelb schattiert ist, und das Wort **End** enthält.
4. Wählen Sie die folgenden Befehle aus den Menüs aus:

   Einfügen > **Zeilen**
5. Geben Sie **Name** und **Beschreibung** des neuen Maßes ein. Beispielsweise könnten Sie **Scaled Error** (Skalierter Fehler) und **Scaling applied to error from Excel** (Skalierung auf Fehler aus Excel angewendet) eingeben.
6. Geben Sie in Zelle C14 die Formel **=COLUMN('Model Measures'!O3)** ein.
7. Geben Sie in Zelle D14 die Formel **=ROW('Model Measures'!O3)+1** ein.

   Diese Formeln geben an, dass die Spalte O das neue Maß für das Modell liefert.
8. Wählen Sie das Arbeitsblatt "Einstellungen" aus.
9. Geben Sie in Zelle A17 die Beschreibung **'- Scaled Error** (Skalierter Fehler) ein.
10. Geben Sie in Zelle B17 den Skalierungsfaktor **10** ein.
11. Geben Sie auf dem Arbeitsblatt "Model Measures" (Modellmaße) die Beschreibung **Scaled Error** (Skalierter Fehler) in Zelle O3 als Titel für die neue Spalte ein.
12. Geben Sie in Zelle O4 die Formel **=N4*Settings!B17** ein.
13. Wählen Sie die Ecke von Zelle O4 aus und ziehen Sie sie nach unten auf Zelle O22, um die Formel in jede Zelle zu kopieren.
14. Wählen Sie die folgenden Befehle aus den Menüs aus:

    Tools > **Schutz** > **Arbeitsblatt schützen**
15. Klicken Sie auf **OK**.
16. Speichern Sie das Arbeitsblatt als Excel 2003-Vorlage, mit einem eindeutigen Namen und der Dateierweiterung *.xlt*. Um die neue Vorlage leichter wieder auffinden zu können, sollten Sie sie am Speicherort der vorkonfigurierten Vorlage auf Ihrem lokalen System oder Netzdateisystem speichern.

Wenn Sie über diese Vorlage eine Verbindung mit Excel herstellen, ist der Fehlerwert als neues benutzerdefiniertes Maß verfügbar.

Visualisieren von Modellen

Die Auswirkung eines Modells wird am deutlichsten, wenn es visuell dargestellt wird. Mithilfe eines Gewinndiagramms erhalten Sie einen wertvollen täglichen Einblick in den betriebswirtschaftlichen und technischen Nutzen Ihres Modells, indem Sie die Effekte mehrerer Alternativen in Echtzeit untersuchen. Im Abschnitt „Gewinndiagramm" wird der Vorteil eines Modells im Vergleich zu einer zufälligen Entscheidungsfindung erläutert. Sie können somit mehrere Diagramme miteinander vergleichen, wenn alternative Modelle vorhanden sind.

Gewinndiagramm: Das Gewinndiagramm bildet die Werte der Tabellenspalte *Gewinn %* ab. Gewinne sind als der Anteil der in jedem Inkrement enthaltenen Treffer im Verhältnis zur Gesamtzahl der im Baum enthaltenen Treffer definiert. Dabei kommt folgende Gleichung zum Einsatz:

(Treffer im Inkrement / Gesamtzahl Treffer) x 100 %

Gewinndiagramme illustrieren, wie weit Sie das Netz auswerfen müssen, um einen bestimmten Prozentsatz aller im Baum enthaltenen Treffer zu erzielen. Die diagonale Linie bildet die für die gesamte Stichprobe erwarteten Treffer ab, wenn das Modell nicht verwendet wird. In diesem Fall ist die Trefferrate konstant, da die Wahrscheinlichkeit eines Treffers für alle Personen gleich ist. Um das Ergebnis zu verdoppeln, müssen Sie doppelt so viele Personen ansprechen. Die gekrümmte Linie zeigt an, wie weit Sie Ihre Treffer verbessern können, wenn Sie nur die einschließen, deren Prozentsatz hinsichtlich des Gewinns höher ausfällt. Wenn Sie beispielsweise die obersten 50 % einschließen, erhalten Sie über 70 % der positiven Treffer. Je steiler die Kurve, desto höher ist der Gewinn.

So zeigen Sie ein Gewinndiagramm an:

1. Öffnen Sie einen Stream, der einen Entscheidungslistenknoten enthält, und starten Sie von diesem Knoten aus eine interaktive Sitzung.
2. Klicken Sie auf die Registerkarte **Gewinne**. Je nachdem, welche Partitionen angegeben sind, werden ein oder zwei Diagramme angezeigt (zwei Diagramme werden angezeigt, wenn für die Modellmaße beispielsweise die Trainings- und die Testpartition definiert sind).

Die Diagramme werden standardmäßig als Segmente angezeigt. Sie können die Anzeige der Diagramme in Quantile ändern, indem Sie **Quantile** und anschließend im Dropdown-Menü die entsprechende Quantilmethode auswählen.

Diagrammoptionen: Die Funktion "Diagrammoptionen" bietet Optionen zur Auswahl der Modelle und Momentaufnahmen, die als Diagramm dargestellt werden, welche Partitionen angezeigt werden und ob Segmentbeschriftungen angezeigt werden oder nicht.

Modelle für das Diagramm

Aktuelle Modelle. Mit dieser Option können Sie auswählen, welche Modelle als Diagramm dargestellt werden sollen. Sie können das Arbeitsmodell oder eines der erstellten Momentaufnahmemodelle auswählen.

Partitionen für das Diagramm

Partitionen für linkes Diagramm. Die Dropdown-Liste enthält Optionen für die Anzeige aller definierten Partitionen oder aller Daten.

Partitionen für rechtes Diagramm. Die Dropdown-Liste enthält Optionen für die Anzeige aller definierten Partitionen, aller Daten oder nur des linken Diagramms. Wenn **Diagramm nur links** ausgewählt ist, wird nur das linke Diagramm angezeigt.

Segmentbeschriftungen anzeigen. Wenn diese Option aktiviert ist, werden die Segmentbeschriftungen in den Diagrammen angezeigt.

Kapitel 10. Statistische Modelle

Statistische Modelle verwenden mathematische Gleichungen, um Informationen zu codieren, die aus den Daten extrahiert wurden. Mitunter können mithilfe statistischer Modellierungstechniken geeignete Modelle sehr schnell bereitgestellt werden. Selbst bei Problemen, bei denen flexiblere Techniken zum Maschinenlernen (z. B. neuronale Netze) letztendlich bessere Ergebnisse liefern, können Sie statistische Modelle als Basisvorhersagemodelle einsetzen, um die Leistung fortgeschrittener Techniken zu beurteilen.

Die folgenden Knoten für die statistische Modellierung sind verfügbar:

Bei linearen Regressionsmodellen wird ein stetiges Ziel auf der Basis linearer Beziehungen zwischen dem Ziel und einem oder mehreren Prädiktoren vorhergesagt.

Die logistische Regression ist ein statistisches Verfahren zur Klassifizierung von Datensätzen auf der Grundlage der Werte von Eingabefeldern. Sie ist analog zur linearen Regression, außer dass statt eines numerischen Bereichs ein kategoriales Zielfeld verwendet wird.

Der Faktor/PCA-Knoten bietet leistungsstarke Datenreduktionsverfahren zur Verringerung der Komplexität der Daten. Die Hauptkomponentenanalyse (PCA) findet lineare Kombinationen der Eingabefelder, die die Varianz im gesamten Set der Felder am besten erfassen, wenn die Komponenten orthogonal (senkrecht) zueinander sind. Mit der Faktorenanalyse wird versucht, die zugrunde liegenden Faktoren zu bestimmen, die die Korrelationsmuster innerhalb eines Sets beobachteter Felder erklären. Bei beiden Ansätzen besteht das Ziel darin, eine kleine Zahl abgeleiteter Felder zu finden, mit denen die Informationen im ursprünglichen Set der Felder effektiv zusammengefasst werden können.

Bei der Diskriminanzanalyse werden strengere Annahmen als bei der logistischen Regression verwendet, sie kann jedoch eine wertvolle Alternative oder Ergänzung zu einer logistischen Regressionsanalyse sein, wenn diese Annahmen erfüllt sind.

Das verallgemeinerte lineare Modell erweitert das allgemeine lineare Modell so, dass die abhängige Variable über eine angegebene Verknüpfungsfunktion in linearem Zusammenhang zu den Faktoren und Kovariaten steht. Außerdem ist es mit diesem Modell möglich, dass die abhängige Variable eine von der Normalverteilung abweichende Verteilung aufweist. Es deckt die Funktionen einer großen Bandbreite an Statistikmodellen ab, darunter lineare Regression, logistische Regression, loglineare Modelle für Häufigkeitsdaten und Überlebensmodelle mit Intervallzensierung.

Verallgemeinerte lineare gemischte Modelle (GLMM - Generalized Linear Mixed Models) erweitern lineare Modelle so, dass das Ziel nicht normalverteilt zu sein braucht und über eine angegebene Verknüpfungsfunktion in einer linearen Beziehung zu den Faktoren und Kovariaten steht und die Beobachtungen korreliert werden können. Verallgemeinerte lineare gemischte Modelle decken eine breite Palette verschiedener Modelle ab, von einfacher linearer Regression bis hin zu komplexen Mehrebenenmodellen für nicht normalverteilte Longitudinaldaten.

Der Knoten vom Typ "Cox-Regression" ermöglicht Ihnen auch bei zensierten Datensätzen die Erstellung eines Überlebensmodells für Daten über die Zeit bis zum Eintreten des Ereignisses. Das Modell erstellt eine Überlebensfunktion, die die Wahrscheinlichkeit vorhersagt, dass das untersuchte Ereignis für bestimmte Werte der Eingabevariablen zu einem bestimmten Zeitpunkt (*t*) eingetreten ist.

Linearknoten

Die lineare Regression ist ein verbreitetes statistisches Verfahren zur Klassifizierung von Datensätzen auf der Grundlage der Werte von numerische Eingabefeldern. Die lineare Regression entspricht einer geraden Linie oder Fläche, die die Diskrepanzen zwischen den vorhergesagten und den tatsächlichen Werten minimiert.

Anforderungen. In linearen Regressionsmodellen können nur numerische Felder verwendet werden. Es werden genau ein Zielfeld (mit der Rolle *Ziel*) und mindestens ein Prädiktor (mit der Rolle *Eingabe*) benötigt. Felder mit der Rolle *Beides* oder *Keine* werden, wie auch nicht numerische Felder, ignoriert. (Nicht numerische Felder können, falls erforderlich, mithilfe eines Ableitungsknotens umcodiert werden.)

Stärken. Lineare Regressionsmodelle sind relativ einfach und bieten eine leicht zu interpretierende mathematische Formel für das Generieren von Vorhersagen. Da die lineare Regressionsmodellierung ein seit langem etabliertes statistisches Verfahren ist, liegen umfassende Kenntnisse über die Eigenschaften dieser Modelle vor. Lineare Regressionsmodelle lassen sich üblicherweise sehr schnell trainieren. Der Linearknoten bietet Methoden für die automatische Feldauswahl zum Entfernen nicht signifikanter Eingabefelder aus der Gleichung.

Hinweis: In Fällen, in denen das Zielfeld keinen stetigen Bereich darstellt, sondern kategorial ist, wie beispielsweise *ja/nein* oder *Abwanderung/Keine Abwanderung*, kann die logistische Regression als Alternative verwendet werden. Die logistische Regression bietet außerdem Unterstützung für nicht numerische Eingaben, sodass eine Umcodierung dieser Felder nicht mehr erforderlich ist. Weitere Informationen finden Sie im Thema „Logistikknoten” auf Seite 169.

Lineare Modelle

Bei linearen Modellen wird ein stetiges Ziel auf der Basis linearer Beziehungen zwischen dem Ziel und einem oder mehreren Prädiktoren vorhergesagt.

Lineare Modelle sind relativ einfach und bieten eine leicht zu interpretierende mathematische Formel für das Scoring. Die Eigenschaften dieser Modelle sind umfassend bekannt und lassen sich üblicherweise im Vergleich zu anderen Modelltypen (wie neuronalen Netzmodellen oder Entscheidungsbäumen) in demselben Dataset sehr schnell erstellen.

Beispiel. Eine Versicherungsgesellschaft mit beschränkten Ressourcen für die Untersuchung der Versicherungsansprüche von Hauseigentümern möchte ein Modell zur Schätzung der Kosten von Ansprüchen erstellen. Durch die Bereitstellung dieses Modells in Service-Centern können zuständige Mitarbeiter bei einem Telefongespräch mit einem Kunden Informationen zum Anspruch eingeben und basierend auf Daten aus der Vergangenheit sofort die "erwarteten" Kosten des Anspruchs abrufen. Weitere Informationen finden Sie im Thema .

Feldanforderungen. Es müssen ein Ziel und mindestens eine Eingabe vorhanden sein. Standardmäßig werden Felder mit den vordefinierten Rollen "Beides" oder "Keine" nicht verwendet. Das Ziel muss stetig (Skala) sein. Bei Prädiktoren (Eingaben) gibt es keine Messniveaubeschränkungen. Kategoriale (Flag, nominal und ordinal) Felder werden als Faktoren im Modell verwendet, stetige Felder als Kovariaten.

Ziele

Was möchten Sie tun?

- **Neues Modell erstellen.** Ein vollständig neues Modell aufbauen. Dies ist die übliche Wirkungsweise des Knotens.
- **Training des bestehenden Modells fortsetzen.** Das Training wird mit dem letzten vom Knoten erfoglreich erstellten Modell fortgesetzt. Dies ermöglicht die Aktualisierung eines bestehenden Modells, ohne dass ein Zugriff auf die ursprünglichen Daten erforderlich ist, und kann zu einer wesentlich schnelleren Leistung führen, da ausschließlich die neuen bzw. aktualisierten Datensätze in den Stream eingespeist werden. Details zum vorherigen Modell werden zusammen mit dem Modellierungsknoten gespeichert, wodurch diese Option auch dann verwendet werden kann, wenn das vorherige Modellnugget nicht mehr im Stream oder in der Modellpalette verfügbar ist.

  Hinweis: Wenn diese Option aktiviert ist, werden alle anderen Steuerelemente auf den Registerkarten "Felder" und "Erstellungsoptionen" inaktiviert.

Was ist Ihr Hauptziel? Wählen Sie das geeignete Ziel aus.

- **Standardmodell erstellen.** Bei der Methode wird ein einziges Modell erstellt, um das Ziel unter Verwendung der Prädiktoren vorherzusagen. In der Regel gilt, dass Standardmodelle einfacher interpretiert und schneller gescort werden können, als verbesserte, verstärkte oder große Dataset-Ensembles.
- **Modellgenauigkeit verbessern (Boosting).** Bei der Methode wird mittels Boosting ein Ensemblemodell erstellt. Dabei wird eine Modellsequenz erzeugt, um genauere Vorhersagen zu erhalten. Möglicherweise nimmt das Erstellen und Scoren bei Ensembles mehr Zeit in Anspruch als bei Standardmodellen.

  Durch Boosting wird eine Reihe von "Komponentenmodellen" erstellt, von denen jede einzelne Komponente auf dem gesamten Dataset beruht. Vor dem Erstellen jedes aufeinander folgenden Komponentenmodells werden die Datensätze basierend auf den Residuen des vorangegangenen Komponentenmodells gewichtet. Größere Residuen erhalten eine höhere Analysegewichtung, sodass beim nächsten Komponentenmodell das Augenmerk auf einer hochwertigen Vorhersage dieser Datensätze liegt. Zusammen bilden diese Komponentenmodelle ein Ensemblemodell. Das Ensemblemodell scort basierend auf einer Kombinationssregel neue Datensätze. Die verfügbaren Regeln hängen vom Messniveau des Ziels ab.
- **Modellstabilität verbessern (Bagging).** Bei der Methode wird mittels Bagging (Bootstrap-Aggregation) ein Ensemblemodell erstellt. Dabei werden mehrere Modelle erzeugt, um zuverlässigere Vorhersagen zu erhalten. Möglicherweise nimmt das Erstellen und Scoren bei Ensembles mehr Zeit in Anspruch als bei Standardmodellen.

  Bei der Bootstrap-Aggregation (Bagging) werden Reproduktionen des Trainingsdataset erstellt, indem von der Ersetzung aus dem ursprünglichen Dataset Stichproben genommen werden. Dadurch werden Bootstrap-Stichproben mit der gleichen Größe wie beim ursprünglichen Dataset erstellt. Dann wird von jeder Reproduktion ein "Komponentenmodell" erstellt. Zusammen bilden diese Komponentenmodelle ein Ensemblemodell. Das Ensemblemodell scort basierend auf einer Kombinationssregel neue Datensätze. Die verfügbaren Regeln hängen vom Messniveau des Ziels ab.
- **Modell für sehr umfangreiche Datasets erstellen (erfordert IBM SPSS Modeler Server).** Bei dieser Methode wird ein Ensemblemodell durch Aufteilen des Datasets in separate Datenblöcke erstellt. Diese Option ist empfehlenswert, wenn Ihr Dataset zu groß für die Erstellung eines der oben erwähnten Modelle oder die inkrementelle Modellerstellung ist. Unter Umständen kann das Modell mit dieser Option schneller als ein Standardmodell erstellt werden, das Scoren dauert jedoch evtl. länger als bei einem Standardmodell. Für diese Option ist eine Verbindung zuIBM SPSS Modeler Server erforderlich.

Informationen zu Einstellungen für Boosting, Bagging und sehr umfangreiche Datasets finden Sie unter „Ensembles” auf Seite 165.

Grundeinstellungen

Daten automatisch vorbereiten. Mit dieser Option kann die Prozedur das Ziel und die Prädiktoren intern transformieren, um die Vorhersagekraft des Modells zu maximieren. Alle Transformationen werden im Modell gespeichert und auf neue Daten zum Scoring angewendet. Die Originalversionen der transformierten Felder werden vom Modell ausgeschlossen. Standardmäßig wird folgende automatische Datenaufbereitung durchgeführt.

- **Verarbeitung von Datum und Zeit.** Jeder Datenprädiktor wird in einen neuen stetigen Prädiktor transformiert, der die Zeit enthält, die seit einem Referenzdatum (1970-01-01) verstrichen ist. Jeder Zeitprädiktor wird in einen neuen stetigen Prädiktor transformiert, der die Zeit enthält, die seit einer Referenzzeit (00:00:00) vergangen ist.
- **Messniveau anpassen.** Stetige Prädiktoren mit weniger als 5 unterschiedlichen Werten werden in ordinale Prädiktoren umgewandelt. Ordinale Prädiktoren mit mehr als zehn eindeutigen Werten werden in stetige Prädiktoren umgewandelt.
- **Ausreißerbehandlung.** Werte von stetigen Prädiktoren, die über einem Trennwert liegen (3 Standardabweichungen vom Mittelwert) werden auf den Trennwert gesetzt.
- **Behandlung fehlender Werte.** Fehlende Werte nominaler Prädiktoren werden durch den Modus der Trainingspartition ersetzt. Fehlende Werte ordinaler Prädiktoren werden durch den Median der Trainingspartition ersetzt. Fehlende Werte stetiger Prädiktoren werden durch den Mittelwert der Trainingspartition ersetzt.
- **Überwachte Zusammenführung.** Mit dieser Option erstellen Sie ein sparsameres Modell, indem Sie die Anzahl der zu verarbeitenden Felder in Zusammenhang mit dem Ziel reduzieren. Ähnliche Kategorien werden anhand der Beziehung zwischen der Eingabe und dem Ziel identifiziert. Kategorien, die sich nicht signifikant unterscheiden (d. h. einen p-Wert aufweisen, der größer als 0,1 ist), werden zusammengeführt. Hinweis: Wenn alle Kategorien zu einer verschmolzen werden, werden die ursprünglichen und abgeleiteten Versionen des Felds aus dem Modell ausgeschlossen, da sie als Einflussgrößen keinen Wert haben.

Konfidenzniveau. Das Konfidenzniveau wird zur Berechnung der Intervallschätzungen der Modellkoeffizienten in der Ansicht Koeffizienten verwendet. Geben Sie einen Wert größer 0 und kleiner 100 ein. Der Standardwert ist 95.

Modellauswahl

Methode zur Modellauswahl. Wählen Sie eine der Methoden zur Modellauswahl (Details unten) oder **Alle Prädiktoren einschließen** aus, wodurch einfach alle verfügbaren Prädiktoren als Haupteffektmodellterme eingegeben werden. Standardmäßig wird **Schrittweise vorwärts** verwendet.

Auswahl schrittweise vorwärts. Diese Option beginnt ohne Effekte im Modell und nimmt jeweils einen Effekt auf bzw. schließt ihn aus, bis entsprechend den Kriterien bei "Schrittweise vorwärts" keine weiteren Vorgänge möglich sind.

- **Kriterien für Aufnahme bzw. Ausschluss.** Diese Statistik wird zur Bestimmung verwendet, ob ein Effekt im Modell aufgenommen oder aus diesem ausgeschlossen werden soll. Das **Informationskriterium (AICC)** basiert auf der Wahrscheinlichkeit des Trainingssets für das Modell und wird zur Penalisierung übermäßig komplexer Modelle angepasst. Die **F-Statistik** beruht auf einem statistischen Test der Verbesserung des Modellfehlers. **Korrigiertes R-Quadrat** beruht auf der Anpassungsgüte des Trainingssets und wird zur Penalisierung übermäßig komplexer Modelle angepasst. Das **Kriterium zur Verhinderung übermäßiger Anpassung (ASE)** basiert auf der Anpassungsgüte (mittlere quadratische Abweichung; Average Squared Error, ASE) des Sets zur Verhinderung übermäßiger Anpassung. Das Set zur Verhinderung übermäßiger Anpassung ist eine zufällige Teilprobe von etwa 30 % des Originaldatasets, die nicht zum Trainieren des Modells verwendet wird.

  Wenn ein anderes Kriterium als **F-Statistik** gewählt wird, wird bei jedem Schritt der Effekt im Modell aufgenommen, der dem größten positiven Zuwachs des Kriteriums entspricht. Alle Effekte, die einer Abnahme des Kriteriums entsprechen, werden aus dem Modell ausgeschlossen.

  Wenn **F-Statistik** als Kriterium gewählt wird, wird bei jedem Schritt der Effekt mit dem geringsten *p*-Wert kleiner als der festgelegte Schwellenwert, **Einschließen von Effekten mit p-Werten kleiner als**, in das Modell aufgenommen. Der Standardwert ist 0,05. Alle Effekte im Modell mit einem *p*-Wert größer als der festgelegte Schwellenwert, **Entfernen von Effekten mit p-Werten größer als** werden ausgeschlossen. Der Standardwert ist 0,10.

- **Maximale Anzahl von Effekten im endgültigen Modell anpassen.** Standardmäßig können alle verfügbaren Effekte in das Modell eingegeben werden. Wenn alternativ der schrittweise Algorithmus einen Schritt bei der festgelegten maximalen Anzahl an Effekten beendet, stoppt der Algorithmus beim aktuellen Effektset.
- **Maximale Schrittanzahl anpassen.** Der schrittweise Algorithmus stoppt nach einer bestimmten Anzahl von Schritten. Standardmäßig ist das dreimal die Anzahl an verfügbaren Effekten. Alternativ kann eine positive Ganzzahl als maximale Anzahl an Schritten angegeben werden.

Auswahl der besten Subsets. Diese Option überprüft "alle möglichen" Modelle oder zumindest eine größere Untergruppe der möglichen Modelle als "Schrittweise vorwärts", um die beste Möglichkeit entsprechend dem Kriterium "Beste Subsets" auszuwählen. Das **Informationskriterium (AICC)** basiert auf der Wahrscheinlichkeit des Trainingssets für das Modell und wird zur Penalisierung übermäßig komplexer Modelle angepasst. **Korrigiertes R-Quadrat** beruht auf der Anpassungsgüte des Trainingssets und wird zur Penalisierung übermäßig komplexer Modelle angepasst. Das **Kriterium zur Verhinderung übermäßiger Anpassung (ASE)** basiert auf der Anpassungsgüte (mittlere quadratische Abweichung; Average Squared Error, ASE) des Sets zur Verhinderung übermäßiger Anpassung. Das Set zur Verhinderung übermäßiger Anpassung ist eine zufällige Teilprobe von etwa 30 % des Originaldatasets, die nicht zum Trainieren des Modells verwendet wird.

Das Modell mit dem höchsten Wert für das Kriterium wird als das beste Modell ausgewählt.

Hinweis: Die Auswahl "Beste Subsets" ist rechenintensiver als die Auswahl "Schrittweise vorwärts". Wenn "Beste Subsets" zusammen mit "Boosting", "Bagging" oder "Sehr große Datasets" verwendet wird, kann das Erstellen deutlich länger dauern als das Erstellen eines Standardmodells mithilfe der Auswahl "Schrittweise vorwärts".

Ensembles

Diese Einstellungen legen das Verhalten der Ensemblebildung fest, die erfolgt, wenn auf der Registerkarte "Ziele" die Option "Boosting", "Bagging" oder "Sehr große Datasets" ausgewählt ist. Optionen, die für das ausgewählte Ziel nicht gelten, werden ignoriert.

Bagging und sehr umfangreiche Datasets. Beim Scoren eines Ensembles wird diese Regel angewendet, um die vorhergesagten Werte aus den Basismodellen für die Berechnung des Scorewerts für das Ensemble zu kombinieren.

- **Standardkombinationsregel für stetige Ziele.** Ensemblevorhersagewerte für stetige Ziele können unter Verwendung des Mittelwerts oder Medians der Vorhersagewerte aus den Basismodellen kombiniert werden.

Hinweis: Wenn als Ziel die Verbesserung der Modellgenauigkeit ausgewählt wurde, wird die Auswahl zum Kombinieren der Regeln ignoriert. Beim Boosting wird für das Scoring der kategorialen Ziele stets eine gewichtete Mehrheit verwendet und für das Scoring stetiger Ziele ein gewichteter Median.

Boosting und Bagging. Geben Sie die Anzahl der zu erstellenden Basismodelle an, wenn als Ziel die Verbesserung der Modellgenauigkeit oder -stabilität angegeben ist. Im Falle des Bagging ist das die Anzahl der Bootstrap-Stichproben. Muss eine positive ganze Zahl sein.

Erweitert

Ergebnisse replizieren. Durch Einstellen eines Startwerts für den Zufallsgenerator können Analysen reproduziert werden. Der Zufallszahlengenerator wird verwendet, um zu wählen, welche Datensätze sich im Set zur Verhinderung übermäßiger Anpassung befinden. Geben Sie eine ganze Zahl ein oder klicken Sie auf **Generieren**. Dadurch wird eine pseudozufällige Ganzzahl zwischen 1 und 2147483647 (einschließlich) erzeugt. Der Standardwert ist 54752075.

Modelloptionen

Modellname. Sie können den Modellnamen automatisch basierend auf den Zielfeldern generieren, oder einen benutzerdefinierten Namen angeben. Der automatisch generierte Name ist der Zielfeldname.

Hinweis: Der vorhergesagte Wert wird immer berechnet, wenn das Modell gescort wird. Der Name des neuen Felds ist der Name des Zielfelds mit einem vorangestellten *$L-*. Bei einem Zielfeld mit der Bezeichnung *Umsatz* beispielsweise erhält das neue Feld den Namen *$L-Umsatz*.

Modellübersicht

Die Ansicht "Modellübersicht" ist eine Momentaufnahme, eine Übersicht auf einen Blick über das Modell und seine Anpassungsgüte.

Tabelle. In der Tabelle werden einige allgemeine Modelleinstellungen dargestellt, darunter:

- Der Name des Ziels, das auf der Registerkarte Felder angegeben ist.
- Ob die automatische Datenvorbereitung durchgeführt wurde, wie in den Grundeinstellungen angegeben wurde.
- Die Modellauswahlmethode und das Auswahlkriterium, das in den Einstellungen für die Modellauswahl angegeben wurde. Der Wert für das Auswahlkriterium des finalen Modells wird ebenfalls angezeigt und in einem Format dargestellt, bei dem kleinere Werte vorteilhafter sind.

Diagramme. Das Diagramm zeigt die Genauigkeit des endgültigen Modells an, das nach dem Prinzip "größer ist besser" dargestellt wird. Der Wert ist 100 × dem angepassten R^2 für das endgültige Modell.

Automatische Datenaufbereitung

Diese Ansicht zeigt Informationen darüber an, welche Felder ausgeschlossen wurden und wie transformierte Felder im Schritt "automatische Datenaufbereitung" (ADP) abgeleitet wurden. Für jedes transformierte oder ausgeschlossene Feld listet die Tabelle den Feldnamen, die Rolle in der Analyse und die im ADP-Schritt vorgenommene Aktion auf. Die Felder werden in aufsteigender alphabetischer Reihenfolge der Feldnamen sortiert. Zu den möglichen Aktionen, die für die Felder durchgeführt wurden, zählen:

- Mit **Ableitung der Dauer: Monate** wird die verstrichene Zeit in Monaten aus den Werten in einem Feld mit Datumsangaben und dem aktuellen Datum des Systems berechnet.
- Mit **Ableitung der Dauer: Stunden** wird die verstrichene Zeit in Stunden aus den Werten in einem Feld mit Zeitangaben und der aktuellen Zeit des Systems berechnet.
- Mit **Messniveau von stetig auf ordinal ändern** werden stetige Felder mit weniger als fünf eindeutigen Werten in ordinale Felder umgewandelt.
- Mit **Messniveau von ordinal auf stetig ändern** werden ordinale Felder mit über zehn eindeutigen Werten in stetige Felder umgewandelt.
- **Ausreißer entfernen** setzt Werte von stetigen Prädiktoren, die über einem Trennwert liegen (3 Standardabweichungen vom Mittelwert) auf den Trennwert.
- **Fehlende Werte ersetzen** ersetzt fehlende Werte von nominalen Feldern durch den Modus, von ordinalen Feldern durch den Median und von stetigen Feldern durch den Mittelwert.
- **Kategorien zur Maximierung des Zielzusammenhangs zusammenführen** ermittelt "ähnliche" Prädiktorkategorien auf der Grundlage der Beziehung zwischen der Eingabe und dem Ziel. Kategorien, die sich nicht signifikant unterscheiden (d. h. einen *p*-Wert aufweisen, der größer als 0,05 ist), werden zusammengeführt.
- **Konstanten Prädiktor ausschließen/nach Ausreißer-Behandlung/nach der Zusammenführung von Kategorien** entfernt Prädiktoren, die einen einzelnen Wert aufweisen, möglicherweise nachdem andere ADP-Aktionen ausgeführt wurden.

Prädiktoreinfluss

In der Regel konzentriert man sich bei der Modellerstellung auf die Prädiktorfelder, die am wichtigsten sind, und vernachlässigt jene, die weniger wichtig sind. Dabei unterstützt Sie das Wichtigkeitsdiagramm für die Prädiktoren, da es die relative Wichtigkeit der einzelnen Prädiktoren für das Modell angibt. Da die Werte relativ sind, beträgt die Summe der Werte aller Prädiktoren im Diagramm 1,0. Die Wichtigkeit der Prädiktoren steht in keinem Bezug zur Genauigkeit des Modells. Sie bezieht sich lediglich auf die Wichtigkeit der einzelnen Prädiktoren für eine Vorhersage und nicht auf die Genauigkeit der Vorhersage.

Vorhergesagt/Beobachtet

Diese Ansicht zeigt ein klassiertes Streudiagramm der vorhergesagten Werte auf der vertikalen Achse durch die beobachteten Werte auf der horizontalen Achse. Idealerweise sollten die Werte entlang einer 45-Grad-Linie liegen. In dieser Ansicht können Sie erkennen, ob bestimmte Datensätze vom Modell besonders schlecht vorhergesagt werden.

Residuen

Diese Ansicht zeigt ein Diagnosediagramm von Modellresiduen.

Diagrammstile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Histogramm.** Hierbei handelt es sich um ein klassiertes Histogramm der studentisierten Residuen, das durch die Normalverteilung überlagert ist. Lineare Modelle gehen davon aus, dass Residuen eine normale Verteilung aufweisen. Das Histogramm sollte sich also idealerweise einer nahezu glatten Linie annähern.
- **P-P-Diagramm.** Hierbei handelt es sich um einen klassierten Wahrscheinlichkeit-Wahrscheinlichkeit-Plot, bei dem die studentisierten Residuen mit einer Normalverteilung verglichen werden. Wenn die Steigung der Diagrammpunkte weniger steil als die normale Linie ist, zeigen die Residuen eine größere Schwankung als eine normale Verteilung; ist die Steigung steiler, zeigen die Residuen weniger Schwankung als eine normale Verteilung. Wenn die Diagrammpunkte eine S-förmige Kurve aufweisen, ist die Verteilung der Residuen verzerrt.

Ausreißer

In dieser Tabelle sind Datensätze aufgelistet, die einen unverhältnismäßigen Einfluss auf das Modell ausüben. Außerdem werden die Datensatz-ID (sofern auf der Registerkarte "Felder" angegeben), der Zielwert und die Cook-Distanz angezeigt. Die Cook-Distanz ist ein Maß dafür, wie stark sich die Residuen aller Datensätze ändern würden, wenn ein spezieller Datensatz von der Berechnung der Modellkoeffizienten ausgeschlossen würde. Ein großer Wert der Cook-Distanz zeigt an, dass der Ausschluss eines Datensatzes von der Berechnung die Koeffizienten substanziell verändert, und sollte daher als einflussreich betrachtet werden.

Einflussreiche Datensätze sollten sorgfältig untersucht werden, um zu entscheiden, ob ihnen bei der Schätzung des Modells eine geringere Gewichtung gegeben werden kann, ob die extremen Werte auf einen akzeptablen Schwellenwert verringert werden können oder ob die einflussreichen Datensätze vollständig entfernt werden sollen.

Effekte

Diese Ansicht zeigt die Größe der einzelnen Effekte im Modell.

Stile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Diagramm.** In diesem Diagramm werden Effekte von oben nach unten nach absteigendem Prädiktoreinfluss sortiert. Verbindungslinien im Diagramm sind basierend auf der Effektsignifikanz gewichtet, wobei eine größere Linienbreite signifikanteren Effekten entspricht (kleinere *p*-Werte). Wenn Sie den Mauszeiger über eine Verbindungslinie bewegen, wird eine QuickInfo mit dem *p*-Wert und der Bedeutung des Effekts angezeigt. Dies ist die Standardeinstellung.
- **Tabelle.** Diese Ansicht zeigt eine ANOVA-Tabelle für das Gesamtmodell und die einzelnen Modelleffekte. Die einzelnen Effekte sind von oben nach unten nach absteigendem Prädiktoreinfluss sortiert. Beachten Sie, dass die Tabelle standardmäßig minimiert ist, sodass nur die Ergebnisse des Gesamtmodells angezeigt werden. Klicken Sie in der Tabelle auf die Zelle für das **korrigierte Modell**, um die Ergebnisse für die einzelnen Modelleffekte anzuzeigen.

Prädiktoreinfluss. Für den Prädiktoreinfluss gibt es einen Schieberegler, mit dem eingestellt wird, welche Prädiktoren in der Ansicht gezeigt werden. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Prädiktoren konzentrieren. Standardmäßig werden die zehn besten Effekte angezeigt.

Signifikanz. Mit dem Schieberegler "Signifikanz" wird gesteuert, welche Effekte in der Ansicht angezeigt werden. Diese Einstellungen gehen über die Eingaben, die auf dem Prädiktoreinfluss beruhen, hinaus. Effekte, deren Signifikanzwerte größer als der Wert des Schiebereglers sind, werden ausgeblendet. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Effekte konzentrieren. Standardmäßig ist der Wert 1,00 eingestellt, sodass keine Effekte basierend auf der Signifikanz herausgefiltert werden.

Koeffizienten

Diese Ansicht zeigt den Wert der einzelnen Koeffizienten im Modell. Hinweis: Faktoren (kategoriale Prädiktoren) sind innerhalb des Modells indikatorcodiert, sodass Faktoren, die **Effekte** enthalten, in der Regel mehrere zugehörige **Koeffizienten** aufweisen. Mit Ausnahme der Kategorie für den redundanten (Referenz-)Parameter erhält jede Kategorie einen solchen Koeffizienten.

Stile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Diagramm.** In diesem Diagramm wird zuerst der konstante Term angezeigt und dann die Effekte von oben nach unten nach abnehmendem Prädiktoreinfluss sortiert. In Faktoren, die Effekte enthalten, werden die Koeffizienten in aufsteigender Reihenfolge der Datenwerte sortiert. Verbindungslinien im Diagramm sind basierend auf dem Vorzeichen des Koeffizienten farbig dargestellt (siehe Diagrammschlüssel) und auf der Grundlage der Koeffizientensignifikanz gewichtet, wobei eine größere Linienbreite signifikanteren Koeffizienten entspricht (kleinere *p*-Werte). Wenn Sie den Mauszeiger über eine Verbindungslinie bewegen, wird eine QuickInfo mit dem Wert des Koeffizienten, seinem *p*-Wert und der Bedeutung des Effekts angezeigt, mit dem der Parameter verbunden ist. Dies ist der Standardstil.
- **Tabelle.** Diese Tabelle zeigt die Werte, Signifikanztests und Konfidenzintervalle für die einzelnen Modellkoeffizienten. Nach dem konstanten Term sind die einzelnen Effekte von oben nach unten nach absteigendem Prädiktoreinfluss sortiert. In Faktoren, die Effekte enthalten, werden die Koeffizienten in aufsteigender Reihenfolge der Datenwerte sortiert. Beachten Sie, dass die Tabelle standardmäßig minimiert ist, sodass nur der Koeffizient, die Signifikanz und die Bedeutung der einzelnen Modellparameter angezeigt werden. Klicken Sie zum Anzeigen des Standardfehlers, der *t*-Statistik und des Konfidenzintervalls in der Tabelle auf die Zelle **Koeffizient**. Wenn Sie den Mauszeiger in der Tabelle über den Namen eines Modellparameters bewegen, wird eine QuickInfo mit dem Namen des Parameters, dem Effekt, mit dem der Parameter verbunden ist, und (für kategoriale Prädiktoren) den Wertbeschriftungen angezeigt, die mit dem Modellparameter verbunden sind. Dies kann besonders hilfreich sein, um die neuen Kategorien anzuzeigen, die erstellt werden, wenn bei der automatischen Datenaufbereitung ähnliche Kategorien eines kategorialen Prädiktors zusammengeführt werden.

Prädiktoreinfluss. Für den Prädiktoreinfluss gibt es einen Schieberegler, mit dem eingestellt wird, welche Prädiktoren in der Ansicht gezeigt werden. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Prädiktoren konzentrieren. Standardmäßig werden die zehn besten Effekte angezeigt.

Signifikanz. Mit dem Schieberegler "Signifikanz" wird gesteuert, welche Koeffizienten in der Anzeige angezeigt werden. Diese Einstellungen gehen über die Eingaben hinaus, die auf dem Prädiktoreinfluss basieren. Koeffizienten, deren Signifikanzwerte größer als der Wert des Schiebereglers sind, werden ausgeblendet. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Koeffizienten konzentrieren. Standardmäßig ist der Wert 1,00 eingestellt, sodass keine Koeffizienten basierend auf der Signifikanz herausgefiltert werden.

Geschätzte Mittelwerte

Diese Diagramme werden für signifikante Prädiktoren angezeigt. Das Diagramm zeigt den vom Modell geschätzten Zielwert auf der vertikalen Achse für jeden Prädiktorwert auf der horizontalen Achse, wobei alle anderen Prädiktoren konstant gehalten werden. Es gewährt eine nützliche Visualisierung der Effekte der einzelnen Prädiktorkoeffizienten auf dem Ziel.

Hinweis: Wenn keine Prädiktoren signifikant sind, werden keine geschätzten Mittelwerte erzeugt.

Übersicht über Modellerstellung

Wenn ein anderer Modellauswahlalgorithmus als **Keiner** in den Einstellungen "Modellauswahl" gewählt wird, werden einige Details zum Modellerstellungsprozess angegeben.

Schrittweise vorwärts Wenn der Auswahlalgorithmus "Schrittweise vorwärts" ist, werden in der Tabelle die letzten 10 Schritte im schrittweisen Algorithmus angezeigt. Für jeden Schritt werden der Wert des Auswahlkriteriums und die Effekte im Modell an diesem Schritt angezeigt. Auf diese Weise bekommen Sie einen Eindruck davon, wie groß der Beitrag der einzelnen Schritte zum Modell ist. In jeder Spalte können Sie die Reihen so sortieren, dass Sie noch leichter erkennen können, welche Effekte sich bei einem bestimmten Schritt im Modell befinden.

Beste Subsets. Wenn der Auswahlalgorithmus "Beste Subsets" ist, werden in der Tabelle die 10 besten Modelle angezeigt. Für jedes Modell werden der Wert des Auswahlkriteriums und die Effekte im Modell angezeigt. So erhalten Sie einen Eindruck der Stabilität der besten Modelle; wenn sie zu vielen ähnlichen Effekten mit wenigen Unterschieden neigen, können Sie sich auf das "Top"-Modell verlassen; wenn sie dagegen sehr unterschiedliche Effekte aufweisen, sind eventuell einige Effekte zu ähnlich und sollten kombiniert (oder entfernt) werden. In jeder Spalte können Sie die Reihen so sortieren, dass Sie noch leichter erkennen können, welche Effekte sich bei einem bestimmten Schritt im Modell befinden.

Einstellungen

Hinweis: Der vorhergesagte Wert wird immer berechnet, wenn das Modell gescort wird. Der Name des neuen Felds ist der Name des Zielfelds mit einem vorangestellten *$L-*. Bei einem Zielfeld mit der Bezeichnung *Umsatz* beispielsweise erhält das neue Feld den Namen *$L-Umsatz*.

SQL für dieses Modell generieren. Bei Verwendung von Daten aus einer Datenbank kann SQL-Code per Pushback zur Ausführung an die Datenbank zurückübertragen werden. Dadurch kann bei vielen Operationen eine bessere Leistung erzielt werden.

Durch Konvertierung in natives SQL scoren. Wenn ausgewählt, wird SQL-Code zum Scoren des Modells innerhalb der Anwendung generiert.

Logistikknoten

Logistische Regression, auch als **nominale Regression** bekannt, ist ein statistisches Verfahren zur Klassifizierung von Datensätzen anhand der Werte der Eingabefelder. Sie ist analog zur linearen Regression, außer dass statt eines numerischen Zielfelds ein kategoriales verwendet wird. Es werden sowohl binomiale Modelle (für Ziele mit zwei diskreten Kategorien) als auch multinomiale Modelle (für Ziele mit mehr als zwei Kategorien) unterstützt.

Die logistische Regression funktioniert durch Erstellung einer Menge von Gleichungen, die die Werte der Eingabefelder mit den Wahrscheinlichkeiten in Relation setzen, die den einzelnen Ausgabefeldkategorien zugeordnet sind. Nach der Generierung des Modells kann es zur Schätzung der Wahrscheinlichkeiten für neue Daten verwendet werden. Für jeden Datensatz wird eine Wahrscheinlichkeit der Zugehörigkeit für jede mögliche Ausgabekategorie berechnet. Die Zielkategorie mit der höchsten Wahrscheinlichkeit wird als vorhergesagter Ausgabewert für den betreffenden Datensatz zugewiesen.

Beispiel für ein binomiales Modell. Ein Telekommunikationsanbieter ist besorgt über die Anzahl an Kunden, die er an Mitbewerber verliert. Mithilfe von Daten über die Servicenutzung können Sie ein bino-

miales Modell erstellen, mit dem Sie prognostizieren können, welche Kunden mit hoher Wahrscheinlichkeit zu einem anderen Anbieter wechseln, und Ihre Angebote emtsprechend anpassen, um so viele Kunden wie möglich zu halten. Ein binomiales Modell wird verwendet, da das Ziel zwei verschiedene Kategorien aufweist (hohe/geringe Wahrscheinlichkeit).

Hinweis: Nur bei binomialen Modellen müssen Zeichenfolgefelder auf acht Zeichen begrenzt sein. Längere Zeichenfolgen können, falls erforderlich, mithilfe eines Umcodierungsknotens umcodiert werden.

Beispiel für ein multinomiales Modell. Ein Telekommunikationsanbieter hat seinen Kundenstamm nach Serviceverwendungsmustern eingeteilt und die Kunden in vier Gruppen unterteilt. Mithilfe von demografischen Daten zum Vorhersagen der Gruppenzugehörigkeit können Sie ein multinomiales Modell erstellen, um potenzielle Kunden in Gruppen einzuteilen und anschließend Angebote für die einzelnen Kunden anzupassen.

Anforderungen. Es werden mindestens ein Eingabefeld und genau ein kategoriales Zielfeld mit mindestens zwei Kategorien benötigt. Bei einem binomialen Modell muss das Ziel über ein Messniveau des Typs *Flag* verfügen. Bei einem multinomialen Modell kann das Ziel ein Messniveau von *Flag* oder *Nominal* mit mindestens zwei Kategorien aufweisen. Felder, die auf *Beides* oder *Keine* gesetzt sind, werden ignoriert. Bei den im Modell verwendeten Feldern müssen die Typen vollständig instanziiert sein.

Stärken. Logistische Regressionsmodelle sind häufig ziemlich genau. Sie können symbolische und numerische Eingabefelder verarbeiten. Sie können die vorhergesagten Wahrscheinlichkeiten für alle Zielkategorien angeben, sodass der zweitbeste Kandidat problemlos ermittelt werden kann. Logistische Modelle sind am effektivsten, wenn es sich bei der Gruppenmitgliedschaft um ein echt kategoriales Feld handelt; wenn die Gruppenmitgliedschaft auf Werten eines stetigen Bereichsfelds (z. B. hoher IQ gegenüber niedrigem IQ) basiert, sollten Sie die Anwendung der linearen Regression in Erwägung ziehen, um die umfassenderen Informationen nutzen zu können, die der vollständige Wertebereich bietet. Logistische Modelle können auch eine automatische Feldauswahl durchführen, obwohl andere Ansätze, wie beispielsweise Baummodelle oder Merkmalauswahl, diese Aufgabe bei großen Datasets möglicherweise schneller durchführen. Und schließlich sind viele Analysten und Data-Mining-Experten gut mit logistischen Modellen vertraut, weshalb sie als Basis verwendet werden können, mit der andere Modellierungstechniken verglichen werden können.

Bei der Verarbeitung großer Datasets können Sie die Leistung deutlich verbessern, indem Sie den Likelihood-Quotienten-Test, eine erweiterte Ausgabeoption, inaktivieren. Weitere Informationen finden Sie im Thema „Logistische Regression - Erweiterte Ausgabe" auf Seite 175.

Logistikknoten - Modelloptionen

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Prozedur. Gibt an, ob ein binomiales oder ein multinomiales Modell erstellt wird. Welche Optionen im Dialogfeld verfügbar sind, hängt davon ab, welche Art von Modellierungsprozedur ausgewählt wurde.

- **Binomial.** Wird verwendet, wenn das Zielfeld ein Flagfeld oder ein nominales Feld mit zwei diskreten Werten (dichotom) ist, wie *Ja/Nein*, *Ein/Aus*, *männlich/weiblich*.
- **Multinomial.** Wird verwendet, wenn das Zielfeld ein nominales Feld mit mehr als zwei Werten ist. Sie können **Haupteffekte**, **Gesättigt** or **Benutzerdefiniert** auswählen.

Konstante in Gleichung einschließen. Mit dieser Option wird bestimmt, ob die entstehenden Gleichungen einen konstanten Term enthalten. In den meisten Fällen sollten Sie diese Option aktiviert lassen.

Binomiale Modelle

Für binomiale Modelle sind folgende Methoden und Optionen verfügbar:

Methode. Dient zur Angabe der bei der Erstellung des logistischen Regressionsmodells verwendeten Methode.

- **Einschluss.** Dies ist die Standardmethode, bei der alle Terme direkt in die Gleichung aufgenommen werden. Beim Erstellen des Modells wird keine Feldauswahl durchgeführt.
- **Vorwärts.** Die Feldauswahlmethode "Vorwärts" erstellt das Modell, indem schrittweise nach vorn gegangen wird. Bei dieser Methode ist das ursprüngliche Modell das einfachste und nur die Konstante und die Terme können zum Modell hinzugefügt werden. Bei jedem Schritt werden die Terme, die sich noch nicht im Modell befinden, darauf getestet, wie sehr sie das Modell verbessern würden und der beste davon wird zum Modell hinzugefügt. Wenn keine Terme mehr hinzugefügt werden können oder der beste der in Frage kommenden Terme nicht zu einer hinreichend großen Verbesserung des Modells führen würde, wird das endgültige Modell generiert.
- **Rückwärts.** Die Methode "Rückwärts" ist im Grunde das Gegenteil der Methode "Vorwärts". Bei dieser Methode enthält das ursprüngliche Modell alle Terme als Prädiktoren und es können nur Terme aus dem Modell entfernt werden. Modellterme, die kaum zum Modell beitragen, werden nach und nach entfernt, bis keine Terme mehr ohne eine signifikante Verschlechterung des Modells entfernt werden können. So entsteht das endgültige Modell.

Kategoriale Eingaben. Listet die Felder auf, die als kategorial gekennzeichnet sind, d. h. diejenigen mit einem Messniveau von "Flag", "Nominal" oder "Ordinal". Sie können den Kontrast und die Basiskategorie für jedes kategoriale Feld angeben.

- **Feldname.** Diese Spalte enthält die Feldnamen der kategorialen Eingaben und ist bereits automatisch mit allen Flagwerten und nominalen Werten in den Daten ausgefüllt. Um stetige oder numerische Eingaben in diese Spalte hinzuzufügen, klicken Sie auf das Symbol "Felder hinzufügen" auf der rechten Seite der Liste und wählen Sie die erforderlichen Eingaben aus.
- **Kontrast.** Die Interpretation der Regressionskoeffizienten für ein kategoriales Feld hängt von den verwendeten Kontrasten ab. Der Kontrast bestimmt, wie die Hypothesentests zum Vergleich der geschätzten Mittel eingerichtet werden. Beispiel: Wenn Sie wissen, dass ein kategoriales Feld eine implizite Reihenfolge aufweist, beispielsweise ein Muster oder eine Gruppierung, können Sie den Kontrast verwenden, um diese Reihenfolge zu modellieren. Folgende Kontraste sind verfügbar:

  Indikator. Die Kontraste kennzeichnen das Vorhandensein oder Nichtvorhandensein einer Kategoriezugehörigkeit. Dies ist die Standardmethode.

  Einfach. Jede Kategorie des Prädiktorfelds - mit Ausnahme der Referenzkategorie selbst - wird mit der Referenzkategorie verglichen.

  Differenz. Jede Kategorie des Prädiktorfelds - mit Ausnahme der ersten Kategorie - wird mit dem durchschnittlichen Effekt der vorherigen Kategorien verglichen. Dies ist auch als umgekehrte Helmert-Kontraste bekannt.

  Helmert. Jede Kategorie des Prädiktorfelds - mit Ausnahme der letzten Kategorie - wird mit dem durchschnittlichen Effekt der nachfolgenden Kategorien verglichen.

  Wiederholt. Jede Kategorie des Prädiktorfelds - mit Ausnahme der ersten Kategorie - wird mit der Kategorie verglichen, die ihr unmittelbar vorangeht.

  Polynomial. Orthogonale polynomiale Kontraste. Es wird angenommen, dass zwischen den Kategorien die gleichen Abstände vorliegen. Polynomiale Kontraste sind nur für numerische Felder verfügbar.

  Abweichung. Jede Kategorie des Prädiktorfelds - mit Ausnahme der Referenzkategorie - wird mit dem Gesamteffekt verglichen.

- **Basiskategorie.** Gibt an, wie die Referenzkategorie für den ausgewählten Kontrasttyp bestimmt wird. Wählen Sie **Erste** aus, um die erste Kategorie für das Eingabefeld zu verwenden (alphabetisch sortiert) oder wählen Sie **Letzte** aus, um die letzte Kategorie zu verwenden. Der Standardwert lautet "Erste".

  Hinweis: Dieses Feld ist bei den Kontrasteinstellungen "Differenz", "Helmert", "Wiederholt" oder "Polynomial" nicht verfügbar.

Die Schätzung des Effekts der einzelnen Felder auf die Gesamttrefferquote (Gesamtantwort) wird als Zunahme oder Abnahme der Wahrscheinlichkeit der einzelnen anderen Kategorien relativ zur Referenzkategorie berechnet. Dadurch können Sie besser die Felder und Werte ermitteln, die mit höherer Wahrscheinlichkeit zu einer bestimmten Antwort führen.

Die Basiskategorie wird in der Ausgabe als "0,0" angezeigt. Dies liegt daran, dass der Vergleich mit ihr selbst zu einem leeren Ergebnis führt. Alle anderen Kategorien werden als Gleichungen in Bezug auf die Basiskategorie angezeigt. Weitere Informationen finden Sie im Thema „Logistisches Modellnugget - Details” auf Seite 178.

Multinomiale Modelle

Für multinomiale Modelle sind folgende Methoden und Optionen verfügbar:

Methode. Dient zur Angabe der bei der Erstellung des logistischen Regressionsmodells verwendeten Methode.

- **Einschluss.** Dies ist die Standardmethode, bei der alle Terme direkt in die Gleichung aufgenommen werden. Beim Erstellen des Modells wird keine Feldauswahl durchgeführt.
- **Schrittweise.** Bei der Methode "Schrittweise" der Feldauswahl wird, wie der Name andeutet, die Gleichung in Schritten erstellt. Das anfängliche Modell ist das einfachste Modell, das möglich ist. Es enthält keine Modellterme (außer der Konstanten) in der Gleichung. Bei jedem Schritt werden die Terme, die noch nicht zum Modell hinzugefügt wurden, bewertet, und wenn der beste dieser Terme in signifikanter Weise zur Vorhersagekraft des Modells beiträgt, wird er hinzugefügt. Außerdem werden die derzeit im Modell enthaltenen Terme neu bewertet, um zu ermitteln, ob einige davon ohne signifikante Beeinträchtigung des Modells entfernt werden können. Wenn dies der Fall ist, werden sie entfernt. Der Vorgang wird wiederholt und andere Terme werden hinzugefügt und/oder entfernt. Wenn keine weiteren Terme zur Verbesserung des Modells hinzugefügt werden können und keine Terme mehr entfernt werden können, ohne das Modell zu beeinträchtigen, wird das endgültige Modell generiert.
- **Vorwärts.** Die Feldauswahlmethode "Vorwärts" ähnelt der Methode "Schrittweise" dahin gehend, dass das Modell in Schritten aufgebaut wird. Allerdings ist bei dieser Methode das ursprüngliche Modell das einfachste und die Konstante und die Terme können nur zum Modell hinzugefügt werden. Bei jedem Schritt werden die Terme, die sich noch nicht im Modell befinden, darauf getestet, wie sehr sie das Modell verbessern würden und der beste davon wird zum Modell hinzugefügt. Wenn keine Terme mehr hinzugefügt werden können oder der beste der in Frage kommenden Terme nicht zu einer hinreichend großen Verbesserung des Modells führen würde, wird das endgültige Modell generiert.
- **Rückwärts.** Die Methode "Rückwärts" ist im Grunde das Gegenteil der Methode "Vorwärts". Bei dieser Methode enthält das ursprüngliche Modell alle Terme als Prädiktoren und es können nur Terme aus dem Modell entfernt werden. Modellterme, die kaum zum Modell beitragen, werden nach und nach entfernt, bis keine Terme mehr ohne eine signifikante Verschlechterung des Modells entfernt werden können. So entsteht das endgültige Modell.
- **Schrittweise rückwärts.** Die Methode "Schrittweise rückwärts" ist im Grunde das Gegenteil der Methode "Schrittweise". Bei dieser Methode enthält das ursprüngliche Modell alle Terme als Prädiktoren. Bei jedem Schritt werden die Terme im Modell evaluiert und alle Terme, die ohne signifikante Beeinträchtigung des Modells entfernt werden können, werden entfernt. Außerdem werden die zuvor entfernten Terme erneut evaluiert, um zu ermitteln, ob der beste dieser Terme in signifikanter Weise zur Vorhersagekraft des Modells beiträgt. Ist dies der Fall, so wird er wieder in das Modell aufgenommen.

Wenn keine Terme mehr entfernt werden können, ohne das Modell wesentlich zu beeinträchtigen, und keine weiteren Terme zur Verbesserung des Modells hinzugefügt werden können, wird das endgültige Modell generiert.

Hinweis: Die automatischen Methoden ("Schrittweise", "Vorwärts" und "Rückwärts") sind sehr anpassungsfähige Lernmethoden und weisen eine starke Tendenz zur übermäßigen Anpassung an die Trainingsdaten auf. Bei der Verwendung dieser Methoden ist es ganz besonders wichtig, die Validität des entstehenden Modells zu überprüfen - entweder mit neuen Daten oder mithilfe einer zurückgehaltenen Teststichprobe, die mit dem Partitionsknoten erstellt wurde.

Basiskategorie für Ziel. Gibt an, wie die Referenzkategorie bestimmt wird. Diese wird als Basis verwendet, anhand deren die Regressionsgleichungen für alle anderen Kategorien im Ziel geschätzt werden. Wählen Sie **Erste** aus, um die erste Kategorie für das aktuelle Zielfeld zu verwenden (alphabetisch sortiert) oder wählen Sie **Letzte** aus, um die letzte Kategorie zu verwenden. Alternativ können Sie durch Klicken auf **Angeben** eine bestimmte Kategorie auswählen und dann den gewünschten Wert in der Liste auswählen. Die verfügbaren Werte lassen sich für jedes Feld in einem Typknoten definieren.

Häufig wird die Kategorie als Basiskategorie angegeben, an der das geringste Interesse besteht, beispielsweie ein Lockartikel. Die anderen Kategorien werden dann auf relative Weise in Bezug zur Basiskategorie gesetzt, um zu bestimmen, wodurch sie mit höherer Wahrscheinlichkeit in ihre eigene Kategorie fallen. Dadurch können Sie besser die Felder und Werte ermitteln, die mit höherer Wahrscheinlichkeit zu einer bestimmten Antwort führen.

Die Basiskategorie wird in der Ausgabe als "0,0" angezeigt. Dies liegt daran, dass der Vergleich mit ihr selbst zu einem leeren Ergebnis führt. Alle anderen Kategorien werden als Gleichungen in Bezug auf die Basiskategorie angezeigt. Weitere Informationen finden Sie im Thema „Logistisches Modellnugget - Details" auf Seite 178.

Modelltyp. Es gibt drei Optionen zur Definition der Terme im Modell. **Haupteffektmodelle** beinhalten nur die einzelnen Eingabefelder und testen nicht die Interaktionen (multiplikativen Effekte) zwischen den Eingabefeldern. **Gesättigte** Modelle enthalten alle Interaktionen sowie die Haupteffekte der Eingabefelder. Gesättigte Modelle sind besser zur Erfassung komplexer Beziehungen in der Lage, sind jedoch auch wesentlich schwieriger zu interpretieren und neigen wesentlich stärker zur übermäßigen Anpassung. Aufgrund der potenziell großen Anzahl möglicher Kombinationen sind die automatischen Methoden zur Feldauswahl (alle Methoden außer "Einschluss") für gesättigte Modelle inaktiviert. **Benutzerdefinierte Modelle** enthalten nur die von Ihnen angegebenen Terme (Haupteffekte und Interaktionen). Verwenden Sie bei der Auswahl dieser Option die Liste "Terme im Modell", um Terme zum Modell hinzuzufügen oder daraus zu entfernen.

Terme im Modell. Beim Erstellen eines benutzerdefinierten Modells müssen Sie die Terme im Modell explizit angeben. Die Liste zeigt die aktuelle Menge an Termen für das Modell. Mit den Schaltflächen auf der rechten Seite der Liste "Terme im Modell" können Sie Modellterme hinzufügen und entfernen.

- Um Terme zum Modell hinzuzufügen, klicken Sie auf die Schaltfläche *Neue Terme im Modell hinzufügen*.
- Zum Löschen von Termen wählen Sie die gewünschten Terme aus und klicken Sie auf die Schaltfläche *Ausgewählte Terme im Modell löschen*.

Hinzufügen von Termen zu einem logistischen Regressionsmodell

Beim Anfordern eines benutzerdefinierten logistischen Regressionsmodells können Sie Terme zum Modell hinzufügen, indem Sie auf der Registerkarte für das logistische Regressionsmodell auf die Schaltfläche *Neue Terme im Modell hinzufügen* klicken. Das Dialogfeld "Neue Terme" wird geöffnet, in dem Sie Terme angeben können.

Typ des hinzuzufügenden Terms. Es gibt mehrere Methoden zum Hinzufügen von Termen zum Modell, je nach der Auswahl der Eingabefelder in der Liste der verfügbaren Felder.

- **Einzelne Interaktion.** Fügt den Term ein, der für die Interaktion aller ausgewählten Felder steht.

- **Haupteffekte.** Fügt für jedes ausgewählte Eingabefeld einen Haupteffekt-Term (das Feld selbst) ein.
- **Alle zweifachen Interaktionen.** Fügt für jedes mögliche Paar ausgewählter Eingabefelder einen Zweifach-Interaktions-Term (das Produkt der Eingabefelder) ein. Beispiel: Bei Auswahl der Eingabefelder *A*, *B* und *C* in der Liste der verfügbaren Felder werden bei dieser Methode die Terme *A * B*, *A * C* und *B * C* eingefügt.
- **Alle dreifachen Interaktionen.** Fügt für jede mögliche Kombination aus jeweils drei ausgewählten Eingabefeldern einen Dreifach-Interaktions-Term (das Produkt der Eingabefelder) ein. Beispiel: Bei Auswahl der Eingabefelder *A*, *B*, *C* und *D* in der Liste der verfügbaren Felder werden bei dieser Methode die Terme *A * B * C*, *A * B * D*, *A * C * D* und *B * C * D* eingefügt.
- **Alle vierfachen Interaktionen.** Fügt für jede mögliche Kombination aus jeweils vier ausgewählten Eingabefeldern einen Vierfach-Interaktions-Term (das Produkt der Eingabefelder) ein. Beispiel: Bei Auswahl der Eingabefelder *A*, *B*, *C*, *D* und *E* in der Liste der verfügbaren Felder werden bei dieser Methode die Terme *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* und *B * C * D * E* eingefügt.

Verfügbare Felder. Listet die verfügbaren Eingabefelder auf, die bei der Konstruktion der Modellterme verwendet werden sollen.

Vorschau. Zeigt die Terme an, die beim Klicken auf **Einfügen** zum Modell hinzugefügt werden. Dabei werden die ausgewählten Felder und der Termtyp zugrunde gelegt.

Einfügen. Fügt (auf der Grundlage der aktuellen Auswahl von Feldern und des Termtyps) Terme in das Modell ein und schließt das Dialogfeld.

Expertenoptionen für Logistikknoten

Wenn Sie über detailliertes Wissen im Bereich logistische Regression verfügen, können Sie mithilfe der Expertenoptionen den Trainingsprozess optimieren. Für den Zugriff auf die Expertenoptionen setzen Sie den Modus auf der Registerkarte "Experten" auf **Experten**.

Skalieren (nur bei multinomialen Modellen). Hier können Sie den Skalierungswert für die Streuung angeben, mit dem die Schätzung der Parameter-Kovarianzmatrix korrigiert wird. Bei der Option **Pearson** wird der Skalierungswert unter Verwendung der Chi-Quadrat-Statistik nach Pearson geschätzt. Bei der Option **Devianz** wird der Skalierungswert unter Verwendung der Devianzfunktion (Likelihood-Quotienten-Chi-Quadrat) geschätzt. Außerdem können Sie auch einen eigenen, benutzerdefinierten Skalierungswert angeben. Hierbei muss es sich um einen positiven numerischen Wert handeln.

Alle Wahrscheinlichkeiten anhängen. Bei Auswahl dieser Option werden die Wahrscheinlichkeiten für die einzelnen Kategorien des Ausgabefelds zu den einzelnen Datensätzen hinzugefügt, die vom Knoten verarbeitet werden. Wenn diese Option nicht ausgewählt wurde, wird nur die Wahrscheinlichkeit der vorhergesagten Kategorie hinzugefügt.

Beispielsweise kommen zu einer Tabelle, die die Ergebnisse eines multinomialen Modells mit drei Kategorien enthält, fünf neue Spalten hinzu. In einer Spalte wird die Wahrscheinlichkeit dafür angegeben, dass das Ergebnis korrekt prognostiziert wird, in der nächsten Spalte wird die Wahrscheinlichkeit angegeben, dass diese Vorhersage ein Treffer oder ein Fehlschlag ist, und in drei weiteren Spalten wird die Wahrscheinlichkeit angegeben, dass die Vorhersage für die einzelnen Kategorien ein Treffer oder ein Fehlschlag ist. Weitere Informationen finden Sie im Thema „Logistisches Modellnugget" auf Seite 177.

Hinweis: Diese Option ist für binomiale Modelle immer ausgewählt.

Toleranz für Prüfung auf Singularität. Dient zur Angabe der Toleranz, die bei der Prüfung auf Singularitäten verwendet wird.

Konvergenz. Mit diesen Optionen können Sie die Parameter für Modellkonvergenz steuern. Bei der Ausführung des Modells steuern die Konvergenzkriterien, wie viele Male die verschiedenen Parameter wiederholt durchlaufen werden, um zu ermitteln, wie gut sie passen. Je häufiger die Parameter durchprobiert

werden, desto enger liegen die Ergebnisse beieinander (d. h. die Ergebnisse konvergieren). Weitere Informationen finden Sie im Thema „Logistische Regression - Konvergenzoptionen".

Ausgabe. Mit diesen Optionen können Sie zusätzliche Statistiken anfordern, die in der erweiterten Ausgabe des vom Knoten erstellten Modellnuggets angezeigt werden. Weitere Informationen finden Sie im Thema „Logistische Regression - Erweiterte Ausgabe".

Kriterien. Mit diesen Optionen können Sie die Kriterien zum Hinzufügen und Entfernen von Feldern mit den Schätzmethoden "Schrittweise", "Vorwärts", "Rückwärts" oder "Schrittweise rückwärts" festlegen. (Die Schaltfläche ist inaktiviert, wenn die Methode "Einschluss" ausgewählt ist.) Weitere Informationen finden Sie im Thema „Logistische Regression - Optionen für die Schrittkriterien" auf Seite 176.

Logistische Regression - Konvergenzoptionen

Sie können die Konvergenzparameter für die Schätzung des logistischen Regressionsmodells festlegen.

Maximale Iterationen. Dient zur Angabe der maximalen Anzahl der Iterationen, die für die Schätzung des Modells verwendet werden.

Maximale Schritthalbierung. Die Schritthalbierung ist ein Verfahren, das von der logistischen Regression verwendet wird, um Komplexitäten im Schätzvorgang zu verarbeiten. Unter normalen Umständen sollten Sie die Standardeinstellung verwenden.

Log-Likelihood-Konvergenz. Iterationen werden angehalten, wenn die relative Änderung der Log-Wahrscheinlichkeit (Log-Likelihood) kleiner als dieser Wert ist. Wenn der Wert gleich 0 ist, wird dieses Kriterium nicht angewendet.

Parameterkonvergenz. Die Iterationen werden angehalten, wenn die absolute oder relative Änderung in den Parameterschätzungen kleiner als dieser Wert ist. Wenn der Wert gleich 0 ist, wird dieses Kriterium nicht angewendet.

Delta (nur multinomiale Modelle). Sie können einen Wert zwischen 0 und 1 angeben, der zu jeder leeren Zelle hinzugefügt werden soll (Kombination aus Ein- und Ausgabefeldwerten). Dadurch kann der Schätzungsalgorithmus besser mit Daten umgehen, wenn viele mögliche Kombinationen von Feldwerten relativ zur Anzahl der Datensätze in den Daten vorhanden sind. Der Standardwert ist 0.

Logistische Regression - Erweiterte Ausgabe

Wählen Sie die optionalen Ausgaben aus, die in der erweiterten Ausgabe des Regressionsmodellnuggets angezeigt werden sollen. Zur Anzeige der erweiterten Ausgabe durchsuchen Sie das Modellnugget und klicken Sie auf die Registerkarte **Erweitert**. Weitere Informationen finden Sie im Thema „Logistisches Modellnugget - Erweiterte Ausgabe" auf Seite 179.

Binomialoptionen

Wählen Sie die für das Modell zu generierenden Ausgabetypen aus. Weitere Informationen finden Sie im Thema „Logistisches Modellnugget - Erweiterte Ausgabe" auf Seite 179.

Anzeigen. Hier können Sie auswählen, ob die Ergebnisse bei jedem Schritt angezeigt werden sollen oder ob gewartet werden soll, bis alle Schritte durchlaufen wurden.

Konfidenzintervall für exp(B). Dient zur Auswahl der Konfidenzintervalle für die einzelnen Koeffizienten (als Beta angezeigt) im Ausdruck. Geben Sie das Niveau des Konfidenzintervalls an (Standard: 95 %).

Restdiagnose. Fordert eine Tabelle mit den fallweisen Diagnosen der Residuen an.

- **Ausreißer außerhalb (Standardabweichung).** Listet nur Fälle mit Residuen auf, bei denen der absolute standardisierte Wert der aufgelisteten Variablen mindestens so groß ist wie der von Ihnen angegebene Wert. Der Standardwert ist 2.
- **Alle Fälle.** Schließt alle Fälle in die Tabelle mit den fallweisen Diagnosen der Residuen ein.

  Hinweis: Da diese Option alle Eingabedatensätze auflistet, kann dies zu einer außergewöhnlich großen Tabelle im Bericht führen, mit einer Zeile pro Datensatz.

Klassifizierungstrennwert. Hiermit können Sie den Trennwert für Klassifizierungsfälle festlegen. Fälle mit vorhergesagten Werten, die den Klassifikationsschwellenwert übersteigen, werden als positiv, vorhergesagte Werte, die unter dem Schwellenwert liegen, als negativ klassifiziert. Um die Standardeinstellung zu ändern, geben Sie einen Wert zwischen 0,01 und 0,99 ein.

Optionen für multinomiale Modelle

Wählen Sie die für das Modell zu generierenden Ausgabetypen aus. Weitere Informationen finden Sie im Thema „Logistisches Modellnugget - Erweiterte Ausgabe" auf Seite 179.

Hinweis: Durch Auswählen der Option **Tests für Likelihood-Quotienten** wird die Verarbeitungszeit, die zum Erstellen eines logistischen Regressionsmodells erforderlich ist, stark erhöht. Wenn die Erstellung des Modells zu lange dauert, sollten Sie diese Option inaktivieren oder stattdessen die Wald- und Scorestatistiken verwenden. Weitere Informationen finden Sie im Thema „Logistische Regression - Optionen für die Schrittkriterien".

Iterationsverlauf für alle. Dient zur Auswahl des Schrittintervalls für das Drucken des Iterationsstatus in der erweiterten Ausgabe.

Konfidenzintervall. Das Konfidenzintervall für Koeffizienten in den Gleichungen. Geben Sie das Niveau des Konfidenzintervalls an (Standard: 95 %).

Logistische Regression - Optionen für die Schrittkriterien

Mit diesen Optionen können Sie die Kriterien zum Hinzufügen und Entfernen von Feldern mit den Schätzmethoden "Schrittweise", "Vorwärts", "Rückwärts" oder "Schrittweise rückwärts" festlegen.

Anzahl der Terme im Modell (nur bei multinomialen Modellen). Für die Modelle vom Typ "Rückwärts" und "Schrittweise rückwärts" können Sie die Mindestzahl der Terme im Modell angeben, für "Vorwärts" und "Schrittweise vorwärts" die Höchstzahl der Terme. Wenn Sie einen Mindestwert über 0 angeben, enthält das Modell die angegebene Anzahl an Termen, selbst wenn einige davon auf der Grundlage statistischer Kriterien entfernt worden wären. Bei den Modellen "Vorwärts", "Schrittweise" und "Einschluss" wird die Mindesteinstellung ignoriert. Wenn Sie einen Maximalwert angeben, werden einige Terme möglicherweise aus dem Modell weggelassen, auch wenn sie aufgrund der statistischen Kriterien ausgewählt worden wären. Bei den Modellen "Rückwärts", "Schrittweise rückwärts" und "Einschluss" wird die Einstellung **Maximum angeben** ignoriert.

Kriterium für Eintragen (nur bei multinomialen Modellen). Wählen Sie **Wert**, um die Verarbeitungsgeschwindigkeit zu maximieren. Die Option **Likelihood-Quotient** kann zu robusteren Schätzungen führen, die Berechnung kann jedoch länger dauern. Standardmäßig wird die Scorestatistik verwendet.

Kriterium für Entfernen. Wählen Sie **Likelihood-Quotient** für ein robusteres Modell. Zur Verkürzung der für die Modellerstellung erforderlichen Zeit können Sie **Wald** auswählen. Wenn in den Daten jedoch eine vollständige oder quasi vollständige Trennung vorliegt (kann durch die Registerkarte "Erweitert" im Modellnugget bestimmt werden), wird die Wald-Statistik besonders unzuverlässig und sollte daher nicht verwendet werden. Standardmäßig wird der Statistiktyp "Likelihood-Quotient" verwendet. Bei binomialen Modellen gibt es die zusätzliche Option **Bedingt**. Diese bietet Ausschlusstests auf der Grundlage der Wahrscheinlichkeit der Likelihood-Quotienten-Statistik, die auf bedingten Parameterschätzungen beruht.

Signifikanzschwellen für LR-Kriterien. Mit dieser Option können Sie Auswahlkriterien basierend auf der statistischen Wahrscheinlichkeit (*p*-Wert) angeben, die den einzelnen Feldern zugeordnet ist. Felder werden nur zum Modell hinzugefügt, wenn der zugehörige *p*-Wert kleiner ist als der Wert für **Aufnahme**, und nur dann entfernt, wenn der *p*-Wert größer ist als der Wert für **Ausschluss**. Der Wert für **Aufnahme** muss unter dem Wert für **Ausschluss** liegen.

Anforderungen für Aufnahme oder Entfernung (nur multinomiale Modelle). Bei einigen Anwendungen hat es, mathematisch gesehen, keinen Sinn, Interaktionsterme zum Modell hinzuzufügen, es sei denn, das Modell enthält außerdem die Terme niedrigerer Ordnung für die zum Interaktionsterm gehörenden Felder. So ist es vielleicht nicht sinnvoll, *A * B* in das Modell aufzunehmen, es sei denn, *A* und *B* kommen ebenfalls im Modell vor. Mit diesen Optionen können Sie festlegen, wie während der schrittweisen Termauswahl mit solchen Abhängigkeiten umgegangen werden soll.

- **Hierarchie für einzelne Effekte.** Effekte höherer Ordnung (Interaktionen, an denen mehr Felder beteiligt sind) werden nur dann in das Modell aufgenommen, wenn alle Effekte niedrigerer Ordnung (Haupteffekte oder Interaktionen mit weniger Feldern) für die betreffenden Felder bereits im Modell enthalten sind, und Effekte niedrigerer Ordnung werden nicht entfernt, wenn Effekte höherer Ordnung, die dieselben Felder betreffen, im Modell vorhanden sind. Diese Option gilt nur für kategoriale Felder.
- **Hierarchie für alle Effekte.** Diese Option funktioniert genau wie die vorherige, außer dass sie auf alle Felder angewendet wird.
- **Alle Effekte einschließen.** Effekte können nur dann im Modell vorkommen, wenn alle in dem Effekt eingeschlossenen Effekte ebenfalls im Modell vorkommen. Diese Option ähnelt der Option **Hierarchie für alle Effekte**, mit der Ausnahme, dass die stetige Felder leicht abweichend behandelt werden. Damit ein Effekt einen anderen Effekt einschließt, muss der eingeschlossene Effekt (niedrigerer Ordnung) *alle* stetigen Felder enthalten, die im einschließenden Effekt (höherer Ordnung) enthalten sind, und bei den kategorialen Feldern des eingeschlossenen Effekts muss es sich um ein Subset der diskreten Felder im einschließenden Effekt handeln. Beispiel: Wenn *A* und *B* kategoriale Felder sind und *X* ein stetiges Feld ist, dann schließt der Term *A * B * X* die Terme *A * X* und *B * X* ein.
- **Keine.** Es werden keine Beziehungen erzwungen; die Terme werden unabhängig zum Modell hinzugefügt und daraus entfernt.

Logistisches Modellnugget

Ein Modellnugget vom Typ "Logistisch" steht für die Gleichung, die durch einen Logistikknoten geschätzt wurde. Diese enthält alle Informationen, die vom logistischen Regressionsmodell erfasst wurden, sowie Informationen über die Struktur und Leistung des Modells. Dieser Gleichungstyp kann auch von anderen Modellen, wie Oracle SVM, generiert werden.

Wenn Sie einen Stream ausführen, der ein Modellnugget vom Typ "Logistisch" enthält, fügt der Knoten zwei neue Felder hinzu, die die Vorhersage des Modells und die zugehörige Wahrscheinlichkeit enthalten. Die Namen der neuen Felder werden aus dem Namen des prognostizierten Ausgabefelds abgeleitet, dem *$L-* für die vorhergesagte Kategorie und *$LP-* für die zugehörige Wahrscheinlichkeit vorangestellt ist. Bei einem Ausgabefeld mit der Bezeichnung *Farbpräf* beispielsweise erhalten die neuen Felder die Namen *$L-Farbpräf* und *$LP-Farbpräf*. Wenn Sie außerdem im Logistikknoten die Option **Alle Wahrscheinlichkeiten ausgeben** ausgewählt haben, wird für jede Kategorie des Ausgabefelds ein zusätzliches Feld hinzugefügt, das die Wahrscheinlichkeit enthält, die zu der entsprechenden Kategorie für die einzelnen Datensätze gehört. Diese zusätzlichen Felder werden auf der Grundlage der Werte des Ausgabefelds benannt, denen *$LP-* vorangestellt wurde. Wenn die gültigen Werte von *Farbpäf* beispielsweise *Rot*, *Grün* und *Blau* sind, werden drei neue Felder hinzugefügt: *$LP-Rot*, *$LP-Grün* und *$LP-Blau*.

Generieren eines Filterknotens. Mit dem Menü "Generieren" können Sie einen neuen Filterknoten erstellen, um Eingabefelder basierend auf den Ergebnissen des Modells zu übergeben. Die Felder, die aufgrund von Multikollinearität aus dem Modell herausgenommen wurden, werden vom generierten Knoten gefiltert ebenso wie Felder, die nicht im Modell verwendet werden.

Logistisches Modellnugget - Details

Bei multinomialen Modellen weist die Registerkarte "Modell" in einem Modellnugget vom Typ "Logistisch" eine geteilte Anzeige auf. Dabei werden die Modellgleichungen im linken und der Prädiktoreinfluss im rechten Fensterbereich angezeigt. Bei binomialen Modellen wird auf der Registerkarte nur der Prädiktoreinfluss angezeigt. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Modellgleichungen

Bei multinomialen Modellen werden im linken Fensterbereich die tatsächlich für das logistische Regressionsmodell geschätzten Gleichungen angezeigt. Für jede Kategorie im Zielfeld (mit Ausnahme der Basiskategorie) gibt es jeweils genau eine Gleichung. Die Gleichungen werden in einem Baumformat angezeigt. Dieser Gleichungstyp kann auch von bestimmten anderen Modellen, wie beispielsweise Oracle SVM, generiert werden.

Gleichung für. Zeigt die Regressionsgleichungen, die bei einem Vorgegebenen Satz an Prädiktorwerten zur Ableitung der Wahrscheinlichkeiten für die Zielkategorie verwendet werden. Die letzte Kategorie des Zielfelds wird als **Basiskategorie** betrachtet; die angezeigten Gleichungen bieten für ein bestimmtes Set an Prädiktorwerten die Log-Odds für die anderen Zielkategorien relativ zur Basiskategorie. Die prognostizierte Wahrscheinlichkeit für die einzelnen Kategorien des jeweiligen Prädiktormusters wird aus diesen Log-Odds-Werten abgeleitet.

Wie werden die Wahrscheinlichkeiten berechnet?

Bei jeder Gleichung werden die Log-Odds für eine bestimmte Zielkategorie relativ zur Basiskategorie berechnet. Bei **Log-Odds**, auch als **Logit** bezeichnet, handelt es sich um den Quotienten aus der Wahrscheinlichkeit für eine angegebene Zielkategorie und der Wahrscheinlichkeit der Basiskategorie, wobei auf das Ergebnis der natürliche Logarithmus angewendet wird. Bei der Basiskategorie sind die Chancen für die Kategorie relativ zu sich selbst 1,0 und daher ist Log-Odds gleich 0. Dies kann als implizite Gleichung für die Basiskategorie betrachtet werden, bei der alle Koeffizienten gleich 0 sind.

Um die Wahrscheinlichkeit aus dem Log-Odds-Wert für eine bestimmte Zielkategorie abzuleiten, verwenden Sie den von der Gleichung für die betreffende Kategorie berechneten Logit-Wert und wenden Sie folgende Formel an:

$P(Gruppe_i) = \exp(g_i) / \Sigma_k \exp(g_k)$

Dabei ist g der berechnete Wert für Log-Odds, i der Kategorieindex und k liegt im Bereich von 1 bis zur Anzahl der Zielkategorien.

Bedeutung des Prädiktors

Optional kann auf der Registerkarte "Modell" auch ein Diagramm, das den relative Einfluss der einzelnen Prädiktoren für die Schätzung des Modells angibt, angezeigt werden. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und die Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass dieses Diagramm nur verfügbar ist, wenn vor dem Generieren des Modells **Prädiktoreinfluss berechnen** auf der Registerkarte "Analysieren" ausgewählt wurde. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Hinweis: Die Berechnung des Prädiktoreinflusses kann bei der logistischen Regression länger dauern als bei anderen Modelltypen und ist standardmäßig nicht auf der Registerkarte "Analysieren" ausgewählt. Die Auswahl dieser Option kann die Leistung verlangsamen, insbesondere bei großen Datasets.

Logistisches Modellnugget - Übersicht

In der Übersicht für ein logistisches Regressionsmodell werden die Felder und Einstellungen angezeigt, die zum Generieren des Modells verwendet wurden. Wenn Sie einen Analyseknoten ausgeführt haben, der an diesen Modellierungsknoten angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt. Allgemeine Informationen zum Verwenden des Modellbrowsers finden Sie unter „Durchsuchen von Modellnuggets" auf Seite 42.

Logistisches Modellnugget - Einstellungen

Auf der Registerkarte "Einstellungen" für ein Modellnugget vom Typ "Logistisch" werden während des Modellscorings Optionen für Konfidenzen, Wahrscheinlichkeiten, Propensity-Scores und SQL-Generierung angegeben. Diese Registerkarte ist erst verfügbar, nachdem das Modellnugget zu einem Stream hinzugefügt wurde, und zeigt je nach Modell- und Zieltyp verschiedene Optionen an.

Multinomiale Modelle

Für multinomiale Modelle sind folgende Optionen verfügbar:

Konfidenzen berechnen. Gibt an, ob während des Scorings die Konfidenzen berechnet werden sollen.

Raw-Propensity-Scores berechnen (nur bei Flagfeldern). Bei Modellen mit Flagzielen (und nur dort) können Sie Raw-Propensity-Scores anfordern, die die Likelihood des für das Zielfeld angegebenen Ergebnisses *true* (wahr) anzeigen. Diese Werte werden zusätzlich zu den standardmäßigen Vorhersage- und Konfidenzwerten ausgegeben. Adjusted-Propensity-Scores sind nicht verfügbar. Weitere Informationen finden Sie im Thema „Analyseoptionen bei Modellierungsknoten" auf Seite 35.

Alle Wahrscheinlichkeiten anhängen. Gibt an, ob die Wahrscheinlichkeiten für die einzelnen Kategorien des Ausgabefelds zu den einzelnen vom Knoten verarbeiteten Datensätzen hinzugefügt werden. Wenn diese Option nicht ausgewählt wurde, wird nur die Wahrscheinlichkeit der vorhergesagten Kategorie hinzugefügt. Bei einem nominalen Ziel mit drei Kategorien beispielsweise enthält das Scoring-Ergebnis eine Spalte für jede der drei Kategorien sowie eine vierte Spalte, die die Wahrscheinlichkeit der vorhergesagten Kategorie angibt. Beispiel: Wenn die Wahrscheinlichkeiten für die Kategorien *Rot*, *Grün* und *Blau* 0,6; 0,3 bzw. 0,1 betragen, ist die vorhergesagte Kategorie *Rot*, mit einer Wahrscheinlichkeit von 0,6.

Durch Konvertierung in natives SQL scoren. Wenn ausgewählt, wird SQL-Code zum Scoren des Modells innerhalb der Anwendung generiert.

Hinweis: Bei multinomialen Modellen ist die SQL-Generierung nicht verfügbar, wenn die Option **Alle Wahrscheinlichkeiten anhängen** ausgewählt wurde. Bei Modellen mit nominalen Zielen ist sie nicht verfügbar, wenn die Option **Konfidenzen berechnen** ausgewählt wurde. SQL-Generierung mit Konfidenzberechnungen wird nur für multinomiale Modelle mit Flagzielen unterstützt. Für binomiale Modelle ist die SQL-Generierung nicht verfügbar.

Binomiale Modelle

Bei binomialen Modellen sind Konfidenzen und Wahrscheinlichkeiten immer aktiviert und die Einstellungen, mit denen Sie diese Optionen inaktivieren könnten, sind nicht verfügbar. Für binomiale Modelle ist die SQL-Generierung nicht verfügbar. Die einzige Einstellung, die bei binomialen Modellen geändert werden kann, ist die Möglichkeit zur Berechnung der Raw-Propensity-Scores. Wie bereits für multinomiale Modelle angegeben, gilt dies nur für Modelle mit Flagzielen. Weitere Informationen finden Sie im Thema „Analyseoptionen bei Modellierungsknoten" auf Seite 35.

Logistisches Modellnugget - Erweiterte Ausgabe

Die erweiterte Ausgabe für die logistische Regression (auch als **nominale Regression** bekannt) bietet detaillierte Informationen zum geschätzten Modell und seiner Leistung. Die meisten der in der erweiterten

Ausgabe enthaltenen Informationen weisen einen hohen Fachlichkeitsgrad auf und zur richtigen Interpretation dieser Ausgaben sind umfassende Kenntnisse der logistischen Regressionsanalyse erforderlich.

Warnungen. Zeigt etwaige Warnungen oder potenzielle Probleme mit den Ergebnissen an.

Zusammenfassung der Fallverarbeitung. Listet die Anzahl der verarbeiteten Datensätze auf, nach den einzelnen symbolischen Feldern im Modell aufgeschlüsselt.

Zusammenfassung der Schritte (optional). Listet die Effekte auf, die bei Verwendung der automatischen Feldauswahl bei jedem Schritt der Modellerstellung hinzugefügt bzw. entfernt werden.

Hinweis: Wird nur für die folgenden Methoden angezeigt: "Schrittweise", "Vorwärts", "Rückwärts" oder "Schrittweise rückwärts".

Iterationsverlauf (optional). Zeigt den Iterationsverlauf von Parameterschätzungen für jede *n*-te Iteration, ausgehend von den ursprünglichen Schätzungen. Dabei ist *n* der Wert des Druckintervalls. Bei der Standardvorgabe wird jede Iteration gedruckt (n=1).

Informationen zur Modellanpassung (multinomiale Modelle). Zeigt den Likelihood-Quotienten-Test für Ihr Modell (endgültig) im Vergleich zu einem, bei dem alle Parameterkoeffizienten 0 sind (nur konstanter Term).

Klassifizierung (optional). Zeigt die Matrix der vorhergesagten und tatsächlichen Ausgabefeldwerte mit den zugehörigen Prozentsätzen an.

Chi-Quadrat-Anpassungsgütestatistiken (optional). Zeigt die Chi-Quadrat-Statistiken nach Pearson sowie die Likelihood-Quotienten-Chi-Quadrat-Statistiken an. Diese Statistiken testen die Gesamtanpassung des Modells für die Trainingsdaten.

Hosmer-Lemeshow-Anpassungsgüte (optional). Zeigt die Ergebnisse der Gruppierung von Fällen in Risikodezile und des Vergleichs der beobachteten Wahrscheinlichkeit mit der erwarteten Wahrscheinlichkeit innerhalb jedes Dezils. Diese Statistik für die Anpassungsgüte ist robuster als die herkömmliche Statistik für die Anpassungsgüte, die in multinomialen Modellen verwendet wird, insbesondere bei Modellen mit stetigen Kovariaten und bei Studien mit kleinen Stichprobenumfängen.

Pseudo-R-Quadrat (optional). Zeigt die *R*-Quadrat-Maße für die Anpassungsgüte nach Cox und Snell, Nagelkerke und McFadden an. Diese Statistiken sind in gewisser Weise analog zu der *R*-Quadrat-Statistik in der linearen Regression.

Monotonizitätsmaße (optional). Zeigt die Anzahl konkordanter Paare, diskordanter Paare und gebundener Paare in den Daten an sowie den Prozentsatz der Gesamtzahl der Paare, den die einzelnen Gruppen darstellen. In dieser Tabelle werden außerdem die Werte Somers-D, Goodman-und-Kruskal-Gamma, Kendall-Tau-a und Konkordanzindex C angezeigt.

Informationskriterien (optional). Zeigt das Akaike-Informationskriterium (AIC) und das Schwarz-Bayes-Informationskriterium (BIC).

Tests für Likelihood-Quotienten (optional). Zeigt Statistiken, die testen, ob die Koeffizienten der Modelleffekte statistisch von 0 abweichen. Signifikante Eingabefelder sind Felder mit sehr niedrigen Signifikanzniveaus in der Ausgabe (mit *Sig.* beschriftet).

Parameterschätzungen (optional). Zeigt die Schätzungen der Gleichungskoeffizienten, Tests für diese Koeffizienten, aus den Koeffizienten abgeleitete Quotenverhältnisse (beschriftet mit *Exp(B)*) sowie Konfidenzintervalle für die Quotenverhältnisse an.

Asymptotische Kovarianz-/Korrelationsmatrix (optional). Zeigt die asymptotischen Kovarianzen und/oder Korrelationen der Koeffizientenschätzungen an.

Beobachtete und vorhergesagte Häufigkeiten (optional). Zeigt für jede Kovariaten-Struktur die beobachteten und vorhergesagten Häufigkeiten für die einzelnen Ausgabefeldwerte an. Diese Tabelle kann ziemlich groß sein, insbesondere bei Modellen mit numerischen Eingabefeldern. Wenn die resultierende Tabelle so groß werden würde, dass sie völlig unhandlich wird, wird sie weggelassen und eine Warnung ausgegeben.

Faktor/PCA-Knoten

Der Faktor/PCA-Knoten bietet leistungsstarke Datenreduktionsverfahren zur Verringerung der Komplexität der Daten. Es sind zwei ähnliche, aber doch völlig getrennte Ansätze verfügbar:

- Die **Hauptkomponentenanalyse (PCA)** findet lineare Kombinationen der Eingabefelder, die die Varianz im gesamten Set der Felder am besten erfassen, wenn die Komponenten orthogonal (lotrecht) zueinander sind. PCA gilt für sämtliche Varianz, darunter sowohl gemeinsame als auch nur für bestimmte Felder geltende Varianz.
- Mit der **Faktorenanalyse** wird versucht, die zugrunde liegenden Konzepte oder **Faktoren** zu bestimmen, die die Korrelationsmuster innerhalb eines Sets beobachteter Felder erklären. Die Faktorenanalyse zielt nur auf die gemeinsame Varianz ab. Varianz, die nur für bestimmte Felder gilt, wird bei der Modellschätzung nicht berücksichtigt. Der Faktor/PCA-Knoten bietet mehrere Methoden der Faktorenanalyse.

Bei beiden Ansätzen besteht das Ziel darin, eine kleinere Zahl abgeleiteter Felder zu finden, mit denen die Informationen in der ursprünglichen Menge der Felder effektiv zusammengefasst werden können.

Anforderungen. In PCA-Faktormodellen können nur numerische Felder verwendet werden. Zum Schätzen einer Faktoranalyse oder PCA ist mindestens ein Feld mit der Rolle *Eingabe* erforderlich. Felder, deren Rolle auf *Ziel*, *Beides* oder *Keine* festgelegt ist, wie nicht-numerische Felder.

Stärken. Die Faktorenanalyse und die Hauptkomponentenanalyse (PCA) können die Komplexität der Daten effektiv reduzieren, ohne den Informationsgehalt wesentlich zu beeinträchtigen. Mit diesen Verfahren können Sie robustere Modelle erstellen, die schneller ausgeführt werden können, als dies mit den rohen Eingabefeldern der Fall wäre.

Faktor/PCA-Knoten - Modelloptionen

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Extraktionsmethode. Dient zur Angabe der für die Datenreduktion verwendeten Methode.

- **Hauptkomponenten.** Dies ist die Standardmethode. Dabei wird die Hauptkomponentenanalyse (PCA) verwendet, um Komponenten zu finden, die die Eingabefelder zusammenfassen.
- **Ungewichtete kleinste Quadrate.** Diese Faktorenanalysemethode beruht auf der Suche nach dem Faktorenset, das am besten das Muster der Beziehungen (Korrelationen) zwischen den Eingabefeldern reproduzieren kann.
- **Verallgemeinerte kleinste Quadrate.** Diese Faktorenanalysemethode ist ähnlich der Methode der ungewichteten kleinsten Quadrate, mit dem Unterschied, dass Gewichtung verwendet wird, um die verstärkte Berücksichtigung von Feldern mit einer großen Menge an spezieller (nicht gemeinsamer) Varianz aufzuheben.

- **Maximum Likelihood.** Bei dieser Faktorenanalysemethode werden Faktorengleichungen erstellt, die höchstwahrscheinlich zu dem beobachteten Muster von Beziehungen (Korrelationen) in den Eingabefeldern geführt haben. Hierbei werden Annahmen über die Form dieser Beziehungen zugrunde gelegt. Die Methode geht insbesondere davon aus, dass für die Trainingsdaten eine multivariate Normalverteilung gilt.
- **Hauptachsen-Faktorenanalyse.** Diese Faktorenanalysemethode ähnelt stark der Hauptkomponentenmethode, mit der Ausnahme, dass sie sich nur auf die gemeinsame Varianz konzentriert.
- **Alpha-Faktorisierung.** Diese Faktorenanalysemethode betrachtet die Felder in der Analyse als Beispiel für die Grundgesamtheit potenzieller Eingabefelder. Dadurch wird die statistische Reliabilität der Faktoren maximiert.
- **Bildfaktorisierung.** Diese Faktorenanalysemethode verwendet die Datenschätzung zur Isolation der gemeinsamen Varianz und zum Ermitteln der Faktoren, die sie beschreiben.

Faktor/PCA-Knoten - Expertenoptionen

Wenn Sie über detailliertes Wissen in den Bereichen Faktorenanalyse und PCA verfügen, können Sie mithilfe der Expertenoptionen den Trainingsprozess optimieren. Für den Zugriff auf die Expertenoptionen setzen Sie den Modus auf der Registerkarte "Experten" auf **Experten**.

Fehlende Werte. Standardmäßig verwendet IBM SPSS Modeler nur Datensätze mit gültigen Werten für alle im Modell verwendeten Felder. (Dies wird zuweilen als **listenweiser Ausschluss** fehlender Werte bezeichnet.) Wenn sehr viele fehlende Daten vorliegen, werden mit diesem Ansatz möglicherweise zu viele Datensätze entfernt, sodass nicht mehr genügend Daten zu Erstellung eines guten Modells vorhanden sind. In solchen Fällen können Sie die Option **Nur vollständige Datensätze verwenden** inaktivieren. IBM SPSS Modeler versucht anschließend, so viele Informationen wie möglich zu verwenden, um das Modell zu schätzen. Hierzu zählen auch Datensätze, bei denen bei einigen Feldern Werte fehlen. (Dies wird zuweilen als **paarweiser Ausschluss** fehlender Werte bezeichnet.) In einigen Situationen jedoch kann eine derartige Verwendung unvollständiger Datensätze zu Berechnungsproblemen bei der Schätzung des Modells führen.

Felder. Dient zur Angabe, ob die Korrelationsmatrix (Standard) oder die Kovarianzmatrix der Eingabefelder für die Schätzung des Modells verwendet werden soll.

Maximale Anzahl der Iterationen für Konvergenz. Dient zur Angabe der maximalen Anzahl der Iterationen, die für die Schätzung des Modells verwendet werden.

Faktoren extrahieren. Es gibt zwei Methoden zur Auswahl der Anzahl der Faktoren, die aus den Eingabefeldern extrahiert werden sollen.

- **Eigenwerte über.** Bei dieser Option werden alle Faktoren oder Komponenten beibehalten, die Eigenwerte aufweisen, die größer sind als das angegebene Kriterium. **Eigenwerte** messen die Fähigkeit der einzelnen Faktoren oder Komponenten zur Zusammenfassung der Varianz in der Menge der Eingabefelder. Das Modell führt bei Verwendung der Korrelationsmatrix zur Beibehaltung aller Faktoren oder Komponenten mit Eigenwerten, die größer sind als der angegebene Wert. Bei Verwendung der Kovarianzmatrix wird das Kriterium als Wert mal mittlerer Eigenwert festgelegt. Bei dieser Skalierung hat diese Option eine ähnliche Bedeutung für beide Matrixtypen.
- **Maximale Anzahl.** Bei dieser Option wird die angegebene Anzahl von Faktoren bzw. Komponenten in absteigender Reihenfolge der Eigenwerte beibehalten. Die Faktoren bzw. Komponenten, die den *n* höchsten Eigenwerten entsprechen, werden also beibehalten. Dabei ist *n* das angegebene Kriterium. Das Standardextraktionskriterium liegt bei fünf Faktoren/Komponenten.

Komponenten-/Faktorladungsmatrix. Mit diesen Optionen wird das Format der Faktorladungsmatrix (bzw. der Komponentenladungsmatrix bei PCA-Modellen) festgelegt.

- **Werte sortieren.** Bei Auswahl dieser Option werden die Faktorladungen in der Modellausgabe numerisch sortiert.

- **Werte ausblenden unter.** Bei Auswahl dieser Option werden die Scores unterhalb des angegebenen Schwellenwerts in der Matrix ausgeblendet, damit das Muster in der Matrix besser erkannt werden kann.

Rotation. Mit diesen Optionen können Sie die Rotationsmethode für das Modell steuern. Weitere Informationen finden Sie im Thema „Faktor/PCA-Knoten - Rotationsoptionen".

Faktor/PCA-Knoten - Rotationsoptionen

In vielen Fällen kann die mathematische Rotation des Sets der beibehaltenen Faktoren ihre Nützlichkeit und insbesondere ihre Interpretierbarkeit erhöhen. Wählen Sie eine Rotationsmethode aus:

- **Keine Rotation.** Standardoption. Es wird keine Rotation verwendet.
- **Varimax.** Eine orthogonale Rotation, bei der die Anzahl der Felder mit hohen Ladungen für die einzelnen Faktoren minimiert wird. Dadurch wird die Interpretation der Faktoren vereinfacht.
- **Direkte Oblimin-Rotation.** Eine Methode für schiefe (nicht orthogonale) Rotation. Wenn **Delta** gleich null 0 ist (Standard), sind die Lösungen schief. Mit zunehmendem negativem Wert von Delta werden die Faktoren weniger schiefwinklig. Um den Standardwert von 0 zu überschreiben, geben Sie eine Zahl kleiner gleich 0,8 ein.
- **Quartimax.** Eine orthogonale Methode, bei der die Anzahl der Faktoren, die für die Erklärung der einzelnen Felder erforderlich sind, minimiert wird. Dadurch wird die Interpretation der beobachteten Felder vereinfacht.
- **Equamax.** Eine Rotationsmethode, bei der es sich um eine Kombination der Varimax-Methode, die die Faktoren vereinfacht, und der Quartimax-Methode, die die Felder vereinfacht, handelt. Die Anzahl der Felder mit hoher Ladung bei einem Faktor und die Anzahl der Faktoren, die für die Erklärung eines Felds erforderlich sind, werden minimiert.
- **Promax.** Eine schiefe Rotation, bei der Faktoren korreliert sein dürfen. Sie lässt sich schneller berechnen als eine direkte Oblimin-Rotation, sodass sie auch für große Datasets verwendet werden kann. **Kappa** steuert die Schiefe der Lösung (den Grad, in dem die Faktoren korreliert werden können).

Modellnugget vom Typ "Faktor/PCA"

Ein Modellnugget vom Typ "Faktor/PCA" stellt das Modell der Faktorenanalyse und der Hauptkomponentenanalyse (PCA) dar, das durch einen Faktor/PCA-Knoten erstellt wurde. Sie enthalten alle Informationen, die vom trainierten Modell erfasst wurden, sowie Informationen über Leistung und Merkmale des Modells.

Wenn Sie einen Stream ausführen, der ein Faktorgleichungsmodell enthält, fügt der Knoten ein neues Feld für jeden Faktor bzw. jede Komponente im Modell hinzu. Die neuen Feldnamen werden vom Modellnamen abgeleitet, mit *$F-* präfigiert und mit *-n* suffigiert. Dabei ist *n* die Nummer des Faktors bzw. der Komponente. Wenn Ihr Modell beispielsweise den Namen *Faktor* aufweist und drei Faktoren enthält, werden die neuen Felder wie folgt benannt: *$F-Faktor-1*, *$F-Faktor-2* und *$F-Faktor-3*.

Um besser zu verstehen, was das Faktormodell codiert hat, können Sie weiter unten im Stream weitere Analysen durchführen. Eine günstige Methode zur Anzeige des Ergebnisses des Faktormodells besteht in der Anzeige der Korrelationen zwischen den Faktoren und den Eingabefeldern mithilfe eines Statistikknotens. Dadurch wird aufgezeigt, welche Eingabefelder welche Faktoren stark belasten, und hilft bei der Ermittlung, ob den Faktoren eine Bedeutung oder Interpretation zugrunde liegt.

Außerdem können Sie das Faktormodell mit den in der erweiterten Ausgabe verfügbaren Informationen bewerten. Zur Anzeige der erweiterten Ausgabe klicken Sie im Browser für Modellnuggets auf die Registerkarte **Erweitert**. Die erweiterte Ausgabe enthält zahlreiche detaillierte Informationen und ist für Benutzer mit umfassenden Kenntnissen im Bereich der Faktoranalyse bzw. der Hauptkomponentenanalyse (PCA) gedacht. Weitere Informationen finden Sie im Thema „Modellnuggets vom Typ "Faktor/PCA" - Erweiterte Ausgabe" auf Seite 184.

Modellnugget vom Typ "Faktor/PCA" - Gleichungen

Auf der Registerkarte "Modell" für ein Modellnugget vom Typ "Faktor/PCA" wird die Faktorscoregleichung für die einzelnen Faktoren angezeigt. Faktor- bzw. Komponentenscores werden berechnet, indem jeder Eingabefeldwert mit seinem Koeffizienten multipliziert und die Summe der Ergebnisse gebildet wird.

Modellnugget vom Typ "Faktor/PCA" - Übersicht

Auf der Registerkarte "Übersicht" für ein Faktormodell werden die Anzahl der im Faktor/PCA-Modell beibehaltenen Faktoren sowie zusätzliche Informationen zu den für die Generierung des Modells verwendeten Feldern und Einstellungen angezeigt. Weitere Informationen finden Sie im Thema „Durchsuchen von Modellnuggets" auf Seite 42.

Modellnuggets vom Typ "Faktor/PCA" - Erweiterte Ausgabe

Die erweiterte Ausgabe für die Faktorenanalyse bietet detaillierte Informationen zum geschätzten Modell und dessen Leistung. Die meisten der in der erweiterten Ausgabe enthaltenen Informationen weisen einen hohen Fachlichkeitsgrad auf und zur richtigen Interpretation dieser Ausgaben sind umfassende Kenntnisse der Faktorenanalyse erforderlich.

Warnungen. Zeigt etwaige Warnungen oder potenzielle Probleme mit den Ergebnissen an.

Kommunalitäten. Zeigt an, welcher Anteil der Varianz der einzelnen Felder durch die Faktoren oder Komponenten erklärt wird. *Initial* gibt die Anfangskommunalitäten mit dem vollständigen Faktorenset aus (das Modell wird mit so vielen Faktoren gestartet, wie Eingabefelder vorhanden sind) und *Extraction* gibt die Kommunalitäten auf der Grundlage des beibehaltenen Faktorensets aus.

Erklärte Gesamtvarianz. Zeigt die von den Faktoren im Modell erklärte Gesamtvarianz an. *Anfängliche Eigenwerte* zeigt die Varianz an, die vom vollständigen Set der Anfangsfaktoren erklärt wird. *Extrahierte Summen von quadrierten Faktorladungen* zeigt die Varianz an, die von den im Modell beibehaltenen Faktoren erklärt wird. *Rotierte Summen von quadrierten Ladungen* zeigt die Varianz an, die von den rotierten Faktoren erklärt wird. Beachten Sie, dass bei schiefen Rotationen *Rotierte Summen von quadrierten Ladungen* nur die Summen der quadrierten Ladungen und keine Varianzprozentsätze angezeigt werden.

Faktormatrix (bzw. Komponentenmatrix). Zeigt die Korrelationen zwischen Eingabefeldern und nicht rotierten Faktoren.

Rotierte Faktormatrix (bzw. Komponentenmatrix). Zeigt die Korrelationen zwischen Eingabefeldern und rotierten Faktoren für orthogonale Rotationen an.

Mustermatrix. Zeigt die partiellen Korrelationen zwischen Eingabefeldern und rotierten Faktoren für schiefe Rotationen an.

Strukturmatrix. Zeigt die einfachen Korrelationen zwischen Eingabefeldern und rotierten Faktoren für schiefe Rotationen an.

Faktorkorrelationsmatrix. Zeigt die Korrelationen zwischen den Faktoren für schiefe Rotationen an.

Diskriminanzknoten

Die Diskriminanzanalyse dient zur Erstellung eines Vorhersagemodells der Gruppenzugehörigkeit. Das Modell besteht aus einer Diskriminanzfunktion (oder, bei mehr als zwei Gruppen, einem Set von Diskriminanzfunktionen) auf der Grundlage derjenigen linearen Kombinationen der Prädiktorvariablen, die die beste Diskriminanz zwischen den Gruppen ergeben. Die Funktionen werden aus einer Stichprobe der Fäl-

le erzeugt, bei denen die Gruppenzugehörigkeit bekannt ist. Diese Funktionen können dann auf neue Fälle mit Messungen für die Prädiktorvariablen, aber unbekannter Gruppenzugehörigkeit angewandt werden.

Beispiel. Ein Telekommunikationsunternehmen kann mithilfe der Diskriminanzanalyse Kunden anhand der Nutzungsdaten in Gruppen einteilen. Hierdurch kann das Unternehmen potenzielle Kunden scoren und sich gezielt denjenigen zuwenden, die mit der größten Wahrscheinlichkeit zu den einträglichsten Gruppen gehören.

Anforderungen. Es werden mindestens ein Eingabefeld und genau ein Zielfeld benötigt. Bei dem Ziel muss es sich um ein kategoriales Feld (mit dem Messniveau *Flag* oder *Nominal*) mit dem Speichertyp "Zeichenfolge" oder "Ganze Zahl" handeln. (Der Speichertyp kann, falls erforderlich, mithilfe eines Füller- oder Ableitungsknotens konvertiert werden.) Felder, die auf *Beides* oder *Keine* gesetzt sind, werden ignoriert. Bei den im Modell verwendeten Feldern müssen die Typen vollständig instanziiert sein.

Stärken. Sowohl die Diskriminanzanalyse als auch die logistische Regression eignen sich jeweils als Klassifizierungsmodell. Die Diskriminanzanalyse geht jedoch von mehr Annahmen zu den Eingabefeldern aus, beispielsweise davon, dass sie normal verteilt sind und stetig sein sollten und dass sie zu besseren Ergebnissen führen, wenn diese Anforderungen erfüllt sind, besonders, wenn der Stichprobenumfang klein ist.

Diskriminanzknoten - Modelloptionen

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen” auf Seite 28.

Methode. Zur Eingabe von Prädiktoren in das Modell sind folgende Optionen verfügbar:

- **Einschluss.** Dies ist die Standardmethode, bei der alle Terme direkt in die Gleichung aufgenommen werden. Terme, die nicht in signifikanter Weise zur Vorhersagekraft des Modells beitragen, werden nicht hinzugefügt.
- **Schrittweise.** Das anfängliche Modell ist das einfachste Modell, das möglich ist. Es enthält keine Modellterme (außer der Konstanten) in der Gleichung. Bei jedem Schritt werden die Terme, die noch nicht zum Modell hinzugefügt wurden, bewertet, und wenn der beste dieser Terme in signifikanter Weise zur Vorhersagekraft des Modells beiträgt, wird er hinzugefügt.

Hinweis: Die Methode "Schrittweise" weist eine starke Tendenz zur übermäßigen Anpassung an die Trainingsdaten auf. Bei der Verwendung dieser Methoden ist es ganz besonders wichtig, die Validität des entstehenden Modells mithilfe einer zurückgehaltenen Teststichprobe oder mit neuen Daten zu überprüfen.

Diskriminanzknoten - Expertenoptionen

Wenn Sie über umfassende Kenntnisse im Bereich der Diskriminanzanalyse verfügen, können Sie mithilfe der Expertenoptionen eine Feinabstimmung des Trainingsvorgangs vornehmen. Für den Zugriff auf die Expertenoptionen stellen Sie auf der Registerkarte "Experten" **Modus** auf **Experten** ein.

A-priori-Wahrscheinlichkeit. Diese Option bestimmt, ob die Klassifikationskoeffizienten für A-priori-Kenntnis über Gruppenzugehörigkeiten angepasst werden.

- **Alle Gruppen gleich.** Es wird von gleichen A-priori-Wahrscheinlichkeiten für alle Gruppen ausgegangen; dies hat keine Auswirkungen auf die Koeffizienten.
- **Aus Gruppengrößen berechnen.** Die beobachteten Gruppengrößen in Ihrem Beispiel bestimmen die A-priori-Wahrscheinlichkeiten der Gruppenzugehörigkeit. Wenn zum Beispiel 50 % der Beobachtungen der Analyse in die erste, 25 % in die zweite und 25 % in die dritte Gruppe fallen, werden die Klassifikationskoeffizienten angepasst, um die Wahrscheinlichkeit der Zugehörigkeit zur ersten Gruppe relativ zu den beiden anderen zu erhöhen.

Kovarianzmatrix verwenden. Sie können wählen, ob zur Klassifikation der Fälle die Kovarianzmatrix innerhalb der Gruppen oder die gruppenspezifische Kovarianzmatrix verwendet werden soll.

- *Innerhalb der Gruppen*. Zur Klassifizierung von Fällen wird die in Pools zusammengefasste Kovarianzmatrix innerhalb der Gruppen verwendet.
- *Gruppenspezifisch.*. Für die Klassifizierung werden gruppenspezifische Kovarianzmatrizen verwendet. Da die Klassifizierung auf Diskriminanzfunktionen und nicht auf ursprünglichen Variablen basiert, entspricht diese Option nicht immer der Verwendung einer quadratischen Diskriminanzfunktion.

Ausgabe. Mit diesen Optionen können Sie zusätzliche Statistiken anfordern, die in der erweiterten Ausgabe des vom Knoten erstellten Modellnugget erscheinen. Weitere Informationen finden Sie im Thema „Diskriminanzknoten - Ausgabeoptionen".

Kriterien. Mit diesen Optionen können Sie die Kriterien zum Hinzufügen und Entfernen von Feldern mit der Schätzmethode "Schrittweise" festlegen. (Die Schaltfläche ist inaktiviert, wenn die Methode "Einschluss" ausgewählt ist.) Weitere Informationen finden Sie im Thema „Diskriminanzknoten - Schrittoptionen" auf Seite 187.

Diskriminanzknoten - Ausgabeoptionen

Wählen Sie die optionalen Ausgaben aus, die in der erweiterten Ausgabe des Modellnuggets vom Typ "Logistische Regression" angezeigt werden sollen. Zur Anzeige der erweiterten Ausgabe durchsuchen Sie das Modellnugget und klicken Sie auf die Registerkarte **Erweitert**. Weitere Informationen finden Sie im Thema „Diskriminanzmodellnugget - Erweiterte Ausgabe" auf Seite 188.

Deskriptive Statistiken. Verfügbare Optionen sind Mittelwerte (einschließlich Standardabweichungen), univariate ANOVA und der Box-*M*-Test.

- *Mittelwerte*. Zeigt Gesamt- und Gruppenmittelwerte sowie Standardabweichungen für die unabhängigen Variablen an.
- *Univariate ANOVA*. Führt für jede unabhängige Variable eine einfaktorielle Varianzanalyse durch, d. h. einen Test auf Gleichheit der Gruppenmittelwerte.
- *Box-M*. Ein Test auf Gleichheit der Kovarianzmatrizen der Gruppen. Bei hinreichend großen Stichproben bedeutet ein nicht signifikanter p-Wert, dass die Anhaltspunkte für unterschiedliche Matrizen nicht ausreichend sind. Der Test ist empfindlich gegenüber Abweichungen von der multivariaten Normalverteilung.

Funktionskoeffizienten. Verfügbare Optionen sind Klassifikationskoeffizienten nach Fisher und nicht standardisierte Koeffizienten.

- *Fisher*. Zeigt die Koeffizienten der Klassifizierungsfunktion nach Fisher an, die direkt für die Klassifizierung verwendet werden können. Es wird ein eigenes Set von Koeffizienten der Klassifizierungsfunktion für jede Gruppe ermittelt. Ein Fall wird der Gruppe zugewiesen, für die er den größten Diskriminanzscore (Klassifizierungsfunktionswert) aufweist.
- *Nicht standardisiert*. Zeigt die nicht standardisierten Koeffizienten der Diskriminanzfunktion an.

Matrizen. Als Koeffizientenmatrizen für unabhängige Variablen sind die Korrelationsmatrix innerhalb der Gruppen, die Kovarianzmatrix innerhalb der Gruppen, die gruppenspezifische Kovarianzmatrix und die Kovarianzmatrix für alle Fälle verfügbar.

- *Korrelationsmatrix innerhalb der Gruppen*. Zeigt eine in Pools zusammengefasste Korrelationsmatrix innerhalb der Gruppen an, die als Durchschnitt der separaten Kovarianzmatrizen für alle Gruppen vor der Berechnung der Korrelationen bestimmt wird.
- *Kovarianz innerhalb der Gruppen*. Zeigt eine Pools zusammengefasste Kovarianzmatrix innerhalb der Gruppen an, die sich von der Gesamtkovarianzmatrix unterscheiden kann. Die Matrix wird als Mittel der einzelnen Kovarianzmatrizen für alle Gruppen berechnet.
- *Kovarianz der einzelnen Gruppen*. Zeigt separate Kovarianzmatrizen für jede Gruppe an.
- *Gesamte Kovarianz*. Zeigt die Kovarianzmatrix für alle Fälle an, so als wären sie aus einer einzigen Stichprobe.

Klassifizierung. Folgende Ausgaben gehören zu den Klassifikationsergebnissen.

- *Fallweise Ergebnisse*. Für jeden Fall werden Codes für die tatsächliche Gruppe, die vorhergesagte Gruppe, A-posteriori-Wahrscheinlichkeiten und Diskriminanzscores angezeigt.
- *Zusammenfassungstabelle*. Die Anzahl der Fälle, die auf Grundlage der Diskriminanzanalyse jeder der Gruppen richtig oder falsch zugeordnet werden. Zuweilen auch als Konfusionsmatrix bezeichnet.
- *Klassifikation mit Fallauslassung*. Jeder Fall der Analyse wird durch Funktionen aus allen anderen Fällen unter Auslassung dieses Falls klassifiziert. Diese Klassifikation wird auch als "U-Methode" bezeichnet.
- *Territorien*. Ein Diagramm der Grenzen, mit denen Fälle auf der Grundlage von Funktionswerten in Gruppen klassifiziert werden. Die Zahlen entsprechen den Gruppen, in die die Fälle klassifiziert wurden. Der Mittelwert jeder Gruppe wird durch einen darin liegenden Stern (*) angezeigt. Dieses Diagramm wird nicht angezeigt, wenn nur eine einzige Diskriminanzfunktion vorliegt.
- *Kombinierte Gruppen*. Erzeugt ein alle Gruppen umfassendes Streudiagramm der Werte für die ersten beiden Diskriminanzfunktionen. Wenn nur eine Funktion vorliegt, wird stattdessen ein Histogramm angezeigt.
- *Gruppenspezifisch.*. Erzeugt gruppenspezifische Streudiagramme der Werte für die ersten beiden Diskriminanzfunktionen. Wenn nur eine Funktion vorliegt, werden stattdessen Histogramme angezeigt.

Schrittweise. Zusammenfassung der Schritte zeigt nach jedem Schritt Statistiken für alle Variablen an. **F für paarweise Distanzen** zeigt eine Matrix mit paarweisen *F*-Quotienten für jedes Gruppenpaar an. Die *F*-Quotienten können für Signifikanztests der Mahalanobis-Abstände zwischen Gruppen verwendet werden.

Diskriminanzknoten - Schrittoptionen

Methode. Wählen Sie die Statistiken aus, die für die Aufnahme oder den Ausschluss neuer Variablen dienen sollen. Die Optionen Wilks-Lambda, nicht erklärte Varianz, Mahalanobis-Abstand, kleinster *F*-Quotient und Rao-*V* stehen zur Verfügung. Mit Rao-*V* können Sie den Mindestanstieg von *V* für eine einzugebende Variable angeben.

- *Wilks-Lambda*. Eine Auswahlmethode für Variablen bei der schrittweisen Diskriminanzanalyse. Die Aufnahme von Variablen in die Gleichung erfolgt anhand der jeweiligen Verringerung von Wilks-Lambda. Bei jedem Schritt wird diejenige Variable aufgenommen, die den Gesamtwert von Wilks-Lambda am meisten vermindert.
- *Nicht erklärte Varianz*. Bei jedem Schritt wird die Variable aufgenommen, welche die Summe der nicht erklärten Streuung zwischen den Gruppen minimiert.
- *Mahalanobis-Distanz*. Dieses Maß gibt an, wie weit die Werte der unabhängigen Variablen eines Falls vom Mittelwert aller Fälle abweichen. Eine große Mahalanobis-Distanz charakterisiert einen Fall, der bei einer oder mehreren unabhängigen Variablen Extremwerte besitzt.
- *Kleinster F-Quotient*. Eine Methode für die Variablenauswahl in einer schrittweisen Analyse. Sie beruht auf der Maximierung eines F-Quotienten, der aus der Mahalanobis-Distanz zwischen den Gruppen errechnet wird.

- *Rao-V*. Ein Maß für die Unterschiede zwischen Gruppenmittelwerten. Auch Lawley-Hotelling-Spur genannt. Bei jedem Schritt wird die Variable aufgenommen, die den Anstieg des Rao-V maximiert. Wenn Sie diese Option ausgewählt haben, geben Sie den Minimalwert ein, den eine Variable für die Aufnahme in die Analyse aufweisen muss.

Kriterien. Verfügbare Alternativen sind **F-Wert verwenden** und **F-Wahrscheinlichkeit verwenden**. Geben Sie Werte für die Eingabe und das Entfernen von Variablen ein.

- *F-Wert verwenden*. Eine Variable wird in ein Modell aufgenommen, wenn ihr F-Wert größer als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn der F-Wert kleiner als der Ausschlusswert ist. Der Aufnahmewert muss größer sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen in das Modell aufzunehmen, senken Sie den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, erhöhen Sie den Ausschlusswert.
- *F-Wahrscheinlichkeit verwenden*. Eine Variable wird in das Modell aufgenommen, wenn das Signifikanzniveau ihres F-Werts kleiner als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn das Signifikanzniveau größer als der Ausschlusswert ist. Der Aufnahmewert muss kleiner sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen in das Modell aufzunehmen, erhöhen Sie den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, senken Sie den Ausschlusswert.

Diskriminanzmodellnugget

Modellnuggets vom Typ "Diskriminanz" stehen für die durch Diskriminanzknoten geschätzten Gleichungen. Sie enthalten alle Informationen, die vom Diskriminanzmodell erfasst wurden, sowie Informationen über die Struktur und Leistung des Modells.

Wenn Sie einen Stream ausführen, der ein Modellnugget vom Typ "Diskriminanz" enthält, fügt der Knoten zwei neue Felder hinzu, die die Vorhersage des Modells und die zugehörige Wahrscheinlichkeit enthalten. Die Namen der neuen Felder werden aus dem Namen des prognostizierten Ausgabefelds abgeleitet, dem *$D-* für die vorhergesagte Kategorie und *$DP-* für die zugehörige Wahrscheinlichkeit vorangestellt ist. Bei einem Ausgabefeld mit der Bezeichnung *Farbpräf* beispielsweise erhalten die neuen Felder die Namen *$D-Farbpräf* und *$DP-Farbpräf*.

Generieren eines Filterknotens. Im Menü "Generieren" können Sie einen neuen Filterknoten zur Übergabe der Eingabefelder auf der Grundlage der Ergebnisse erstellen.

Bedeutung des Prädiktors

Optional kann auf der Registerkarte "Modell" auch ein Diagramm, das den relative Einfluss der einzelnen Prädiktoren für die Schätzung des Modells angibt, angezeigt werden. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und die Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass dieses Diagramm nur verfügbar ist, wenn vor dem Generieren des Modells **Prädiktoreinfluss berechnen** auf der Registerkarte "Analysieren" ausgewählt wurde. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors” auf Seite 43.

Diskriminanzmodellnugget - Erweiterte Ausgabe

Die erweiterte Ausgabe für die Diskriminanzanalyse bietet detaillierte Informationen zum geschätzten Modell und dessen Leistung. Die meisten der in der erweiterten Ausgabe enthaltenen Informationen weisen einen hohen Fachlichkeitsgrad auf und zur richtigen Interpretation dieser Ausgaben sind umfassende Kenntnisse der Diskriminanzanalyse erforderlich. Weitere Informationen finden Sie im Thema „Diskriminanzknoten - Ausgabeoptionen” auf Seite 186.

Diskriminanzmodellnugget - Einstellungen

Über die Registerkarte "Einstellungen" für ein Modellnugget vom Typ "Diskriminanz" können Sie beim Scoren des Modells Propensity-Scores ermitteln. Diese Registerkarte ist nur für Modelle mit Flagzielen verfügbar und erst nachdem das Modellnugget einem Stream hinzugefügt wurde.

Raw-Propensity-Scores berechnen. Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Diese Werte werden zusätzlich zu den anderen Vorhersage- und Konfidenzwerten angegeben, die gegebenenfalls während des Scorings erstellt werden.

Adjusted-Propensity-Scores berechnen. Raw-Propensity-Scores beruhen ausschließlich auf den Trainingsdaten und sind aufgrund der Neigung vieler Modelle zur übermäßigen Anpassung an die Trainingsdaten möglicherweise zu optimistisch. Bei Adjusted Propensitys wird versucht, dies durch Evaluation der Modellleistung anhand einer Test- bzw. Validierungspartition zu kompensieren. Bei dieser Option muss im Stream ein Partitionsfeld definiert sein und die Adjusted-Propensity-Scores müssen im Modellierungsknoten aktiviert werden, bevor das Modell generiert wird.

Diskriminanzmodellnugget - Übersicht

Auf der Registerkarte "Übersicht" für ein Modellnugget vom Typ "Diskriminanz" werden die Felder und Einstellungen angezeigt, die zum Generieren des Modells verwendet wurden. Wenn Sie einen Analyseknoten ausgeführt haben, der an diesen Modellierungsknoten angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt. Allgemeine Informationen zum Verwenden des Modellbrowsers finden Sie unter „Durchsuchen von Modellnuggets” auf Seite 42.

GenLin-Knoten

Das verallgemeinerte lineare Modell erweitert das allgemeine lineare Modell so, dass die abhängige Variable über eine angegebene Verknüpfungsfunktion in linearem Zusammenhang zu den Faktoren und Kovariaten steht. Außerdem ist es mit diesem Modell möglich, dass die abhängige Variable eine von der Normalverteilung abweichende Verteilung aufweist. Es deckt durch seine sehr allgemein gehaltene Modellformulierung häufig verwendete statistische Modelle ab, wie beispielsweise die lineare Regression für normalverteilte Antworten, logistische Modelle für binäre Daten und loglineare Modelle für Häufigkeitsdaten, Modelle vom Typ "Log-Log komplementär" für intervallzensierte Überlebensdaten sowie viele andere statistische Modelle.

Beispiele. Eine Reederei kann verallgemeinerte lineare Modelle verwenden, um eine Poisson-Regression an die Anzahl von Beschädigungen für mehrere Schifftypen anzupassen, die in verschiedenen Zeiträumen gebaut wurden. Anhand des daraus hervorgehenden Modells kann bestimmt werden, welche Schiffstypen besonders anfällig auf Schäden sind.

Ein KFZ-Versicherungsunternehmen kann verallgemeinerte lineare Modelle verwenden, um eine Gammaregression an die Schadensersatzansprüche für Autos anzupassen. Anhand des daraus hervorgehenden Modells können die Faktoren bestimmt werden, die am meisten zur Anspruchshöhe beitragen.

Mediziner können mithilfe von verallgemeinerten linearen Modellen eine komplementäre Log-Log-Regression an intervallzensierte Überlebensdaten anpassen, um die Dauer bis zum Wiederauftreten einer Erkrankung vorherzusagen.

Bei verallgemeinerten linearen Modellen wird eine Gleichung erstellt, die die Werte der Eingabefelder mit den Werten der Ausgabefelder in Bezug setzt. Nach der Generierung des Modells kann es zur Schätzung der Werte für neue Daten verwendet werden. Für jeden Datensatz wird eine Wahrscheinlichkeit der Zugehörigkeit für jede mögliche Ausgabekategorie berechnet. Die Zielkategorie mit der höchsten Wahrscheinlichkeit wird als vorhergesagter Ausgabewert für den betreffenden Datensatz zugewiesen.

Anforderungen. Es werden mindestens ein Eingabefeld und genau ein Zielfeld (mit dem Messniveau *Stetig* oder *Flag*) mit mindestens zwei Kategorien benötigt. Bei den im Modell verwendeten Feldern müssen die Typen vollständig instanziiert sein.

Stärken. Das verallgemeinerte lineare Modell ist extrem flexibel, jedoch ist der Prozess für die Auswahl der Modellstruktur nicht automatisiert, weshalb ein Grad an Vertrautheit mit den Daten nötig ist, der bei einem "Black Box"-Algorithmus nicht erforderlich ist.

Feldoptionen für den GenLin-Knoten

Zusätzlich zu den benutzerdefinierten Optionen für Ziel, Eingabe und Partition, die normalerweise auf der Registerkarte "Felder" des Modellierungsknotens bereitgestellt werden (siehe „Feldoptionen der Modellierungsknoten” auf Seite 31), stellt der GenLin-Knoten die folgenden Zusatzfunktionen bereit.

Gewichtungsfeld verwenden. Der Skalenparameter ist ein geschätzter Modellparameter, der mit der Varianz der Antwort zusammenhängt. Die Skalengewichtungen sind "bekannte" Werte, die sich zwischen den einzelnen Beobachtungen unterscheiden können. Wenn die Skalengewichtungsvariable angegeben ist, wird der Skalenparameter, der mit der Varianz der Antwort zusammenhängt, für jede Beobachtung durch diese Variable geteilt. Datensätze, bei denen die Werte für die Skalengewichtung kleiner oder gleich 0 sind oder fehlen, werden nicht in der Analyse verwendet.

Zielfeld enthält Anzahl der Ereignisse, die in einem Set von Versuchen eintreten. Wenn es sich bei der Antwort um eine Reihe von Ereignissen handelt, die in einem Set von Versuchen eintreten, enthält das Zielfeld die Anzahl der Ereignisse und Sie können eine zusätzliche Variable auswählen, die die Anzahl der Versuche enthält. Wenn die Anzahl der Versuche über alle Subjekte gleich ist, können die Versuche alternativ auch über einen festen Wert angegeben werden. Die Anzahl der Versuche sollte größer oder gleich der Anzahl der Ereignisse für die einzelnen Datensätze sein. Bei den Ereignissen sollte es sich um nicht negative Ganzzahlen und bei den Versuchen um positive Ganzzahlen handeln.

Modelloptionen für den GenLin-Knoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen” auf Seite 28.

Modelltyp. Es gibt zwei Optionen für den zu erstellenden Modelltyp. **Haupteffekte** sorgt dafür, dass das Modell nur die einzelnen Eingabefelder enthält und nicht die Interaktionen (multiplikativen Effekte) zwischen den Eingabefeldern testet. **Haupteffekte und alle Zweifach-Interaktionen** umfasst alle Zwei-Wege-Interktionen sowie die Haupteffekte der Eingabefelder.

Offset. Der Term "Offset" ist einer "struktureller" Prädiktor. Ihr Koeffizient wird nicht vom Modell geschätzt, sondern es wird davon ausgegangen, dass er den Wert 1 aufweist. Daher werden die Werte des Offsets einfach zur linearen Prädiktoren des Ziels addiert. Dies ist besonders nützlich bei Poisson-Regressionsmodellen, bei denen die verschiedenen Fälle dem relevanten Ereignis unterschiedlich stark ausgesetzt sein können

Beispielsweise gibt es bei der Modellierung der Unfallraten für einzelne Fahrer einen wichtigen Unterschied zwischen einem Fahrer, der in 3 Jahren Fahrpraxis einen Unfall verursacht hat und einem Fahrer, der in 25 Jahren einen Unfall verursacht hat. Die Anzahl der Unfälle kann als Poisson-Antwort oder als negative binomiale Antwort mit einer Log-Verknüpfung angezeigt werden, wenn der natürliche Logarithmus der Fahrpraxis des Fahrers als Offset-Term eingeschlossen wird.

Bei anderen Kombinationen von Verteilung und Verknüpfungstypen wären andere Transformationen der Offset-Variablen erforderlich.

Hinweis: Bei Verwendung eines variablen Offsetfelds sollte das angegebene Feld nicht als Eingabe verwendet werden. Setzen Sie, falls erforderlich, die Rolle des Offsetfelds in einem Quellen- oder Typknoten weiter oben im Stream auf **Keine**.

Basiskategorie für Flagziel.

Bei binären Antworten können Sie die Referenzkategorie für die abhängige Variable auswählen. Dies kann sich auf bestimmte Ausgaben, wie beispielsweise Parameterschätzungen und gespeicherte Werte, auswirken, sollte jedoch nicht die Anpassungsgüte des Modells verändern. Beispiel: Angenommen, Ihre binäre Antwort nimmt die Werte 0 und 1 an:

- Standardmäßig verwendet die Prozedur die letzte Kategorie (die mit dem höchsten Wert), also 1, als Referenzkategorie. In dieser Situation wird anhand der vom Modell gespeicherten Wahrscheinlichkeiten die Wahrscheinlichkeit geschätzt, mit der ein bestimmter Fall den Wert 0 annimmt, und die Parameterschätzungen sollten als in Beziehung zur Likelihood der Kategorie 0 stehend interpretiert werden.
- Wenn Sie die erste Kategorie (die mit dem niedrigsten Wert), also 0, als Referenzkategorie angeben, wird anhand der im Modell gespeicherten Wahrscheinlichkeitswerte die Wahrscheinlichkeit geschätzt, dass ein bestimmter Fall den Wert 1 annimmt.
- Wenn Sie die benutzerdefinierte Kategorie angeben und für Ihre Variable Beschriftungen definiert sind, können Sie die Referenzkategorie durch Auswahl eines Werts aus der Liste festlegen. Dies kann nützlich sein, wenn Sie bei der Festlegung eines Modells nicht mehr wissen, wie genau eine bestimmte Variable codiert war.

Konstanten Term in Modell einschließen. Der konstante Term wird für gewöhnlich in das Modell aufgenommen. Wenn anzunehmen ist, dass die Daten durch den Koordinatenursprung verlaufen, können Sie den konstanten Term ausschließen.

Expertenoptionen für den GenLin-Knoten

Wenn Sie über umfassende Kenntnisse im Bereich der verallgemeinerten linearen Modelle verfügen, können Sie mithilfe der Expertenoptionen eine Feinabstimmung des Trainingsvorgangs vornehmen. Für den Zugriff auf die Expertenoptionen stellen Sie auf der Registerkarte "Experten" **Modus** auf **Experten** ein.

Verteilung im Zielfeld und Linkfunktion

Verteilung.

Diese Auswahl gibt die Verteilung der abhängigen Variablen an. Die Möglichkeit einer anderen Verteilung als "Normal" und einer anderen Verknüpfungsfunktion als "Identität" ist die wichtigste Verbesserung des verallgemeinerten linearen Modells gegenüber dem allgemeinen linearen Modell. Es gibt zahlreiche mögliche Kombinationen aus Verteilung und Verknüpfungsfunktion und es können mehrere davon für das jeweils vorliegende Dataset geeignet sein. Daher können Sie sich in Ihrer Wahl durch theoretische Vorüberlegungen leiten lassen oder davon, welche Kombination am besten zu passen scheint.

- **Binomial.** Diese Verteilung ist nur für Variablen geeignet, die eine binäre Antwort oder eine Anzahl von Ereignissen repräsentieren.
- **Gamma.** Diese Verteilung eignet sich für Variablen mit positiven Skalenwerten, die in Richtung größerer positiver Werte verzerrt sind. Wenn ein Datenwert kleiner oder gleich 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet.
- **Invers normal.** Diese Verteilung eignet sich für Variablen mit positiven Skalenwerten, die in Richtung größerer positiver Werte verzerrt sind. Wenn ein Datenwert kleiner oder gleich 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet.
- **Negativ binomial.** Diese Verteilung lässt sich als Anzahl der Versuche betrachten, die erforderlich sind, um *k* Erfolge zu beobachten, und eignet sich für Variablen mit nicht negativen ganzzahligen Werten. Wenn ein Datenwert keine ganze Zahl oder kleiner als 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet. Der feste Wert des Hilfsparameters der negativen Binomialverteilung kann jede beliebige Zahl größer oder gleich 0 sein. Wenn der Hilfsparameter auf 0 gesetzt wird, entspricht die Verwendung dieser Verteilung der Verwendung der Poisson-Verteilung.
- **Normal.** Diese Option eignet sich für metrische Variablen, deren Werte eine symmetrische, glockenförmige Verteilung um einen Mittelwert aufweisen. Die abhängige Variable muss numerisch sein.

- **Poisson.** Diese Verteilung lässt sich als Anzahl der Vorkommen eines untersuchten Ereignisses in einem festen Zeitraum betrachten und eignet sich für Variablen mit nicht negativen ganzzahligen Werten. Wenn ein Datenwert keine ganze Zahl oder kleiner als 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet.
- **Tweedie.** Diese Verteilung eignet sich für Variablen, die durch Poisson-Mischungen von Gamma-Verteilungen repräsentiert werden können. Die Verteilung ist dahin gehend "gemischt", dass sie sowohl Eigenschaften von stetigen Verteilungen (nimmt nicht negative reelle Werte an) als auch von diskreten Verteilungen (positive Wahrscheinlichkeitsmasse an einem Einzelwert, 0) aufweist. Die abhängige Variable muss numerisch sein, mit Datenwerten größer oder gleich 0. Wenn ein Datenwert kleiner als 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet. Der feste Wert des Parameters der Tweedie-Verteilung kann jede beliebige Zahl zwischen 1 und 2 sein.
- **Multinomial.** Diese Verteilung eignet sich für Variablen, die eine ordinale Antwort repräsentieren. Bei der abhängigen Variablen kann es sich um eine numerische Variable oder eine Zeichenfolgevariable handeln. Sie muss mindestens zwei verschiedene gültige Datenwerte aufweisen.

Linkfunktionen.

Die Verknüpfungsfunktion ist eine Transformation der abhängigen Variable, die eine Schätzung des Modells ermöglicht. Die folgenden Funktionen sind verfügbar:

- **Identität.** $f(x)=x$. Die abhängige Variable wird nicht transformiert. Diese Verknüpfung kann mit jeder beliebigen Verteilung verwendet werden.
- **Log-Log komplementär.** $f(x)=\log(-\log(1-x))$. Nur für die Binomialverteilung geeignet.
- **Cauchit (kumulativ).** $f(x) = \tan(\pi\ (x - 0.5))$, auf die kumulative Wahrscheinlichkeit der einzelnen Kategorien der Antwort angewendet. Nur für die Multinomialverteilung geeignet.
- **Log-Log komplementär (kumulativ).** $f(x)=\ln(-\ln(1-x))$, auf die kumulative Wahrscheinlichkeit der einzelnen Kategorien der Antwort angewendet. Nur für die Multinomialverteilung geeignet.
- **Logit (kumulativ).** $f(x)=\ln(x\ /\ (1-x))$, auf die kumulative Wahrscheinlichkeit der einzelnen Kategorien der Antwort angewendet. Nur für die Multinomialverteilung geeignet.
- **Log-Log negativ (kumulativ).** $f(x)=-\ln(-\ln(x))$, auf die kumulative Wahrscheinlichkeit der einzelnen Kategoriender Antwort angewendet. Nur für die Multinomialverteilung geeignet.
- **Probit (kumulativ).** $f(x)=\Phi^{-1}(x)$, auf die kumulative Wahrscheinlichkeit der einzelnen Kategorien der Antwort angewendet. Dabei steht Φ^{-1} für die umgekehrte kumulative Standardnormalverteilungsfunktion steht. Nur für die Multinomialverteilung geeignet.
- **Log.** $f(x)=\log(x)$. Diese Verknüpfung kann mit jeder beliebigen Verteilung verwendet werden.
- **Log-Komplement.** $f(x)=\log(1-x)$. Nur für die Binomialverteilung geeignet.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Nur für die Binomialverteilung geeignet.
- **Negativ binomial.** $f(x)=\log(x\ /\ (x+k^{-1}))$. Dabei steht k für den Hilfsparameter der negativen Binomialverteilung. Nur für die negative Binomialverteilung geeignet.
- **Log-Log negativ.** $f(x)=-\log(-\log(x))$. Nur für die Binomialverteilung geeignet.
- **Odds-Potenz.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, wenn $\alpha \neq 0$. $f(x)=\log(x)$, wenn $\alpha=0$. α ist die erforderliche Zahlenangabe. Es muss sich dabei um eine reelle Zahl handeln. Nur für die Binomialverteilung geeignet.
- **Probit.** $f(x)=\Phi^{-1}(x)$, wobei Φ^{-1} die kumulative Standardnormalverteilungsfunktion ist. Nur für die Binomialverteilung geeignet.
- **Exponentiell.** $f(x)=x^{\alpha}$, wenn $\alpha \neq 0$. $f(x)=\log(x)$, wenn $\alpha=0$. α ist die erforderliche Zahlenangabe. Es muss sich dabei um eine reelle Zahl handeln. Diese Verknüpfung kann mit jeder beliebigen Verteilung verwendet werden.

Parameter. Mit den Steuerelementen in dieser Gruppe können Sie Parameterwerte festlegen, wenn bestimmte Verteilungsoptionen gewählt werden.

- **Parameter für negativ binomial.** Für negative binomiale Verteilung geben Sie entweder einen Wert an oder Sie gestatten dem System, einen geschätzten Wert bereitzustellen.

- **Parameter für Tweedie.** Geben Sie als festen Wert des Parameters der Tweedie-Verteilung eine Zahl zwischen 1,0 und 2,0 an.

  Parameterschätzung. Mit den Steuerelementen in dieser Gruppe können Sie Schätzmethoden festlegen und Anfangswerte für die Parameterschätzungen angeben.

  – **Methode.** Sie können eine Methode für die Parameterschätzung auswählen. Sie haben die Wahl zwischen "Newton-Raphson", "Fisher-Bewertung" und einer Hybridmethode, bei der zuerst Iterationen des Fisher-Scorings durchgeführt werden und dann zur Methode "Newton-Raphson" gewechselt wird. Wenn während der Phase "Fisher-Bewertung" der Hybridmethode Konvergenz erreicht wird, bevor die maximale Anzahl an Fisher-Iterationen erreicht wurde, fährt der Algorithmus mit der Newton-Raphson-Methode fort.

  – **Skalenparametermethode.** Sie können eine Schätzmethode für den Skalenparameter auswählen. Bei der Maximum-Likelihood-Methode wird der Skalenparameter zusammen mit den Modelleffekten geschätzt. Beachten Sie, dass diese Option nicht gültig ist, wenn die Antwort eine negative Binomialverteilung, eine Poisson-Verteilung oder eine Binomialverteilung aufweist. Die Optionen für die Abweichung und das Pearson-Chi-Quadrat schätzen den Skalenparameter aus dem Wert der jeweiligen Statistik. Alternativ können Sie einen festen Wert für den Skalenparameter angeben.

- **Kovarianzmatrix.** Der modellbasierte Schätzer ist das Negative der verallgemeinerten Inversen der Hesse-Matrix. Der robuste Schätzer (auch Huber-/White-/Sandwich-Schätzer genannt) ist ein "korrigierter" modellbasierter Schätzer, der eine konsistente Schätzung der Kovarianz bietet, selbst wenn Varianz und Verknüpfungsfunktionen falsch angegeben wurden.

Iterationen. Mit diesen Optionen können Sie die Parameter für die Modellkonvergenz festlegen. Weitere Informationen finden Sie im Thema „Verallgemeinerte lineare Modelle - Iterationen".

Ausgabe. Mit diesen Optionen können Sie zusätzliche Statistiken anfordern, die in der erweiterten Ausgabe des vom Knoten erstellten Modellnugget erscheinen. Weitere Informationen finden Sie im Thema „Verallgemeinerte lineare Modelle - Erweiterte Ausgabe" auf Seite 194.

Toleranz für Prüfung auf Singularität. Singuläre (bzw. nicht invertierbare) Matrizen weisen linear abhängige Spalten auf, die zu ernsten Problemen für den Schätzalgorithmus führen können. Auch annähernd singuläre Matrizen können zu schlechten Ergebnissen führen, daher behandelt die Prozedur eine Matrix, deren Determinante unter dem Toleranzwert liegt, als singulär. Geben Sie einen positiven Wert ein.

Verallgemeinerte lineare Modelle - Iterationen

Sie können die Konvergenzparameter für die Schätzung des verallgemeinerten linearen Modells festlegen.

Iterationen. Die folgenden Optionen sind verfügbar:

- **Maximale Iterationen.** Dies ist die maximale Anzahl der Iterationen, die im Algorithmus vorgenommen werden. Geben Sie eine nicht negative Ganzzahl an.
- **Maximale Schritthalbierung.** Bei jeder Iteration wird die Schrittgröße um den Faktor 0,5 reduziert, bis die Log-Likelihood ansteigt oder die Maximalzahl für die Schritthalbierung erreicht ist. Geben Sie eine positive Ganzzahl ein.
- **Auf Trennung der Datenpunkte prüfen.** Mit dieser Option lassen Sie Tests durch den Algorithmus durchführen, mit denen sichergestellt wird, dass die Parameterschätzungen eindeutige Werte aufweisen. Eine Trennung wird vorgenommen, sobald ein Modell erzeugt werden kann, in dem alle Fälle fehlerfrei klassifiziert werden. Diese Option ist für binomiale Antworten mit Binärformat verfügbar.

Konvergenzkriterien. Die folgenden Optionen sind verfügbar:

- **Parameterkonvergenz.** Mit dieser Option wird der Algorithmus nach einer Iteration angehalten, bei der die absolute oder relative Änderung bei den Parameterschätzungen unter dem angegebenen (positiven) Wert liegt.

- **Log-Likelihood-Konvergenz.** Mit dieser Option wird der Algorithmus nach einer Iteration angehalten, bei der die absolute oder relative Änderung bei der Log-Likelihood-Funktion unter dem angegebenen (positiven) Wert liegt.
- **Konvergenz der Hesse-Matrix.** Für die Spezifikation "Absolut" wird angenommen, dass eine Konvergenz vorliegt, wenn eine Statistik auf der Basis der Konvergenz der Hesse-Matrix kleiner als der angegebene positive Wert ist. Für die Spezifikation "Relativ" wird angenommen, dass eine Konvergenz vorliegt, wenn die Statistik kleiner als das Produkt aus dem angegebenen positiven Wert und dem absoluten Wert der Log-Likelihood ist.

Verallgemeinerte lineare Modelle - Erweiterte Ausgabe

Wählen Sie die optionalen Ausgaben aus, die in der erweiterten Ausgabe des Modellnuggets für das verallgemeinerte lineare Modell angezeigt werden sollen. Zur Anzeige der erweiterten Ausgabe durchsuchen Sie das Modellnugget und klicken Sie auf die Registerkarte **Erweitert**. Weitere Informationen finden Sie im Thema „GenLin-Modellnugget - Erweiterte Ausgabe" auf Seite 195.

Die folgenden Ausgaben sind verfügbar:

- **Zusammenfassung der Fallverarbeitung.** Zeigt die Anzahl und den Prozentsatz der Fälle an, die in die Analyse und in die Tabelle "Korrelierte Datenzusammenfassung" aufgenommen bzw. daraus ausgeschlossen werden.
- **Deskriptive Statistiken.** Zeigt eine deskriptive Statistik und Zusammenfassungsinformationen über die abhängige Variable, die Kovarianten und die Faktoren an.
- **Modellinformationen.** Zeigt den Namen des Datasets, die abhängige Variable bzw. die Ereignis- und Versuchsvariablen, die Offset-Variable, die Skalengewichtungsvariable, die Wahrscheinlichkeitsverteilung und die Verknüpfungsfunktion an.
- **Statistik für Anpassungsgüte.** Zeigt an: Abweichung und skalierte Abweichung, Pearson-Chi-Quadrat und skaliertes Pearson-Chi-Quadrat, Log-Likelihood, Akaike-Informationskriterium (AIC), AIC mit Korrektur für endliche Stichproben (AICC), Bayes-Informationskriterium (BIC) und konsistentes AIC (CAIC).
- **Modellübersichtsstatistik.** Zeigt Tests für die Anpassungsgüte des Modells an, darunter Likelihood-Quotienten-Statistiken für den Omnibus-Test für die Anpassungsgüte, sowie Statistiken für Kontraste des Typs I bzw. III für jeden Effekt.
- **Parameterschätzungen.** Zeigt Parameterschätzungen und entsprechende Teststatistiken und Konfidenzintervalle an. Wahlweise können Sie zusätzlich zu den rohen, unbearbeiteten Parameterschätzungen auch potenzierte Parameterschätzungen anzeigen.
- **Kovarianzmatrix für Parameterschätzungen.** Zeigt die Kovarianzmatrix für die geschätzten Parameter an.
- **Korrelationsmatrix für Parameterschätzungen.** Zeigt die Korrelationsmatrix für die geschätzten Parameter an.
- **Kontrastkoeffizientenmatrizen (L-Matrizen).** Zeigt die Kontrastkoeffizienten für die Standardeffekte und für die geschätzten Randmittel an, sofern auf der Registerkarte "Geschätzte Randmittel" angefordert.
- **Allgemeine schätzbare Funktionen.** Zeigt die Matrizen für die Generierung der Kontrastkoeffizientenmatrizen (L-Matrizen) an.
- **Iterationsverlauf.** Zeigt den Iterationsverlauf für Parameterschätzungen und Log-Likelihood an und druckt die letzte Auswertung des Gradientenvektors und der Hesse-Matrix. Die Tabelle des Iterationsverlaufs zeigt Parameterschätzungen für jede n^{te} Iteration an, beginnend mit der 0^{ten} Iteration (Anfangsschätzungen), wobei *n* für den Wert des Druckintervalls steht. Wenn der Iterationsverlauf angefordert wird, wird die letzte Iteration unabhängig von *n* stets angezeigt.
- **Lagrange-Multiplikator-Test.** Zeigt die Statistiken für den Lagrange-Multiplikator-Test an, die zur Bewertung der Gültigkeit eines Skalenparameters dienen, der mithilfe des Pearson-Chi-Quadrats berechnet wurde oder für den bei der Normal-, Gamma- und inversen Normalverteilung ein fester Wert festgelegt wurde. Bei der negativen Binomialverteilung wird hiermit der feste Hilfsparameter getestet.

Modelleffekte. Die folgenden Optionen sind verfügbar:

- **Analysetyp.** Geben Sie den Typ der zu erstellenden Analyse an. Eine Analyse des Typs I ist im Allgemeinen dann angebracht, wenn Sie von vorneherein Gründe dafür haben, die Prädiktoren im Modell zu ordnen. Typ III dagegen ist allgemeiner anwendbar. Wald- oder Likelihood-Quotienten-Statistiken werden auf der Basis der Auswahl in der Chi-Square-Statistikgruppe berechnet.
- **Konfidenzintervalle.** Geben Sie für das Konfidenzniveau einen Wert an, der über 50 und unter 100 liegt. Wald-Intervalle beruhen auf der Annahme, dass die Parameter eine asymptotische Normalverteilung aufweisen. Profil-Likelihood-Intervalle sind präziser, können aber rechnerisch aufwendig sein. Die Toleranzstufe für Profil-Likelihood-Intervalle ist das Kriterium, anhand dessen der iterative Algorithmus zur Intervallberechnung gestoppt wird.
- **Log-Likelihood-Funktion.** Legt das Anzeigeformat der Log-Likelihood-Funktion fest. Die vollständige Funktion enthält einen zusätzlichen Term, der hinsichtlich der Parameterschätzungen konstant ist. Er hat keine Auswirkungen auf die Parameterschätzung und wird bei einigen Softwareprodukten nicht angezeigt.

GenLin-Modellnugget

Ein GenLin-Modellnugget steht für die Gleichungen, die durch einen GenLin-Knoten geschätzt wurden. Sie enthalten alle Informationen, die vom Modell erfasst wurden, sowie Informationen über die Struktur und Leistung des Modells.

Bei der Ausführung eines Streams, der ein GenLin-Modellnugget enthält, fügt der Knoten neue Felder hinzu, deren Inhalt von der Art des Zielfelds abhängt:

- **Flagziel.** Fügt Felder hinzu, die die vorhergesagte Kategorie und die zugehörige Wahrscheinlichkeit enthalten sowie die Wahrscheinlichkeiten für die einzelnen Kategorien. Die Namen der ersten beiden neuen Felder werden aus dem Namen des prognostizierten Ausgabefelds abgeleitet, dem *$G-* für die vorhergesagte Kategorie und *$GP-* für die zugehörige Wahrscheinlichkeit vorangestellt ist. Bei einem Ausgabefeld mit der Bezeichnung *Standard* beispielsweise erhalten die neuen Felder die Namen *$G-Standard* und *$GP-Standard*. Diese letzteren beiden zusätzlichen Felder werden auf der Grundlage der Werte des Ausgabefelds benannt, denen *$GP-* vorangestellt ist. Wenn für *Standard* die Werte *Ja* und *Nein* zulässig sind, lauten die Namen der neuen Felder *$GP-Ja* und *$GP-Nein*.
- **Stetiges Ziel.** Fügt Felder hinzu, die den vorhergesagten Mittelwert und Standardfehler enthalten.
- **Stetiges Ziel, enthält die Anzahl von Ereignissen in einer Reihe von Versuchen.** Fügt Felder hinzu, die den vorhergesagten Mittelwert und Standardfehler enthalten.
- **Ordinales Ziel.** Fügt Felder hinzu, die die vorhergesagte Kategorie und die zugehörige Wahrscheinlichkeit für die einzelnen Werte des sortierten Sets enthalten. Die Namen der Felder werden von dem Wert des vorhergesagten sortierten Sets abgeleitet, dem *$G-* für die vorhergesagte Kategorie und *$GP-* für die zugehörige Wahrscheinlich vorangestellt ist.

Generieren eines Filterknotens. Im Menü "Generieren" können Sie einen neuen Filterknoten zur Übergabe der Eingabefelder auf der Grundlage der Ergebnisse erstellen.

Bedeutung des Prädiktors

Optional kann auf der Registerkarte "Modell" auch ein Diagramm, das den relative Einfluss der einzelnen Prädiktoren für die Schätzung des Modells angibt, angezeigt werden. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und die Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass dieses Diagramm nur verfügbar ist, wenn vor dem Generieren des Modells **Prädiktoreinfluss berechnen** auf der Registerkarte "Analysieren" ausgewählt wurde. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors” auf Seite 43.

GenLin-Modellnugget - Erweiterte Ausgabe

Die erweiterte Ausgabe für verallgemeinerte lineare Modelle bietet detaillierte Informationen zum geschätzten Modell und dessen Leistung. Die meisten der in der erweiterten Ausgabe enthaltenen Informa-

tionen weisen einen hohen Fachlichkeitsgrad auf und zur richtigen Interpretation dieser Ausgaben sind umfassende Kenntnisse dieser Art von Analysen erforderlich. Weitere Informationen finden Sie im Thema „Verallgemeinerte lineare Modelle - Erweiterte Ausgabe” auf Seite 194.

GenLin-Modellnugget - Einstellungen

Auf der Registerkarte "Einstellungen" für ein GenLin-Modellnugget können Sie beim Scoren des Modells Propensity-Scores ermitteln. Diese Registerkarte ist nur für Modelle mit Flagzielen verfügbar und erst nachdem das Modellnugget einem Stream hinzugefügt wurde.

Raw-Propensity-Scores berechnen. Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Diese Werte werden zusätzlich zu den anderen Vorhersage- und Konfidenzwerten angegeben, die gegebenenfalls während des Scorings erstellt werden.

Adjusted-Propensity-Scores berechnen. Raw-Propensity-Scores beruhen ausschließlich auf den Trainingsdaten und sind aufgrund der Neigung vieler Modelle zur übermäßigen Anpassung an die Trainingsdaten möglicherweise zu optimistisch. Bei Adjusted Propensitys wird versucht, dies durch Evaluation der Modellleistung anhand einer Test- bzw. Validierungspartition zu kompensieren. Bei dieser Option muss im Stream ein Partitionsfeld definiert sein und die Adjusted-Propensity-Scores müssen im Modellierungsknoten aktiviert werden, bevor das Modell generiert wird.

GenLin-Modellnugget - Übersicht

Auf der Registerkarte "Übersicht" für ein GenLin-Modellnugget werden die Felder und Einstellungen angezeigt, die zum Generieren des Modells verwendet wurden. Wenn Sie einen Analyseknoten ausgeführt haben, der an diesen Modellierungsknoten angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt. Allgemeine Informationen zum Verwenden des Modellbrowsers finden Sie unter „Durchsuchen von Modellnuggets” auf Seite 42.

Verallgemeinerte lineare gemischte Modelle

GLMM-Knoten

Verwenden Sie diesen Knoten, um ein verallgemeinertes lineares gemischtes Modell (Generalized Linear Mixed Model, GLMM) zu erstellen.

Verallgemeinerte lineare gemischte Modelle

Verallgemeinerte lineare gemischte Modelle erweitern das lineare Modell wie folgt:

- Das Ziel steht über eine angegebene Verknüpfungsfunktion in einer linearen Beziehung zu den Faktoren und Kovariaten.
- Das Ziel kann eine von der Normalverteilung abweichende Verteilung aufweisen.
- Es kann eine Korrelation zwischen den Beobachtungen bestehen.

Verallgemeinerte lineare gemischte Modelle decken eine breite Palette verschiedener Modelle ab, von einfacher linearer Regression bis hin zu komplexen Mehrebenenmodellen für nicht normalverteilte Longitudinaldaten.

Beispiele. Eine Schulbehörde kann ein verallgemeinertes lineares gemischtes Modell verwenden, um zu ermitteln, ob eine experimentelle Lehrmethode die Mathematikleistungen effektiv verbessert. Schüler aus derselben Klasse sollten korreliert werden, da sie von demselben Lehrer unterrichtet werden. Klassen in der derselben Schule können ebenfalls korreliert werden, sodass wir Zufallseffekte auf Schul- und Klassenebene einschließen können, um die verschiedenen Quellen für Variabilität zu berücksichtigen. Weitere Informationen finden Sie im Thema .

Wissenschaftler aus der Medizinforschung können ein verallgemeinertes lineares gemischtes Modell verwenden, um zu ermitteln, ob ein neues Antikonvulsivum die Häufigkeit epileptischer Anfälle bei einem Patienten verringern kann. Messwiederholungen bei ein und demselben Patienten sind typischerweise positiv korreliert. Daher sollte ein gemischtes Modell mit einigen Zufallseffekten angemessen sein. Das Zielfeld, die Anzahl der Anfälle, nimmt positive ganzzahlige Werte an. Daher kann ein verallgemeinertes lineares gemischtes Modell mit einer Poisson-Verteilung und einer Log-Verknüpfung geeignet sein. Weitere Informationen finden Sie im Thema .

Die Geschäftsführung eines Kabelanbieters für Fernseh-, Telefon- und Internetdienstleistungen kann ein verallgemeinertes lineares gemischtes Modell verwenden, um mehr über potenzielle Kunden zu erfahren. Da mögliche Antworten nominale Messniveaus aufweisen, verwendet der Unternehmensanalyst ein verallgemeinertes gemischtes Logit-Modell mit einer Zufallskontante, um die Korrelation zwischen den Antworten auf Fragen zur Servicenutzung für die verschiedenen Servicetypen (Fernsehen, Telefon, Internet) innerhalb der Antworten eines bestimmten Umfrageteilnehmers zu erfassen. Weitere Informationen finden Sie im Thema .

Über die Registerkarte "Datenstruktur" können Sie die strukturellen Beziehungen zwischen Datensätzen in Ihrem Dataset festlegen, wenn Beobachtungen miteinander korrelieren. Wenn die Datensätze im Dataset unabhängige Beobachtungen darstellen, müssen Sie auf dieser Registerkarte nichts festlegen.

Subjekte. Die Wertekombination der angegebenen kategorialen Felder sollte die Subjekte innerhalb des Datasets eindeutig definieren. Beispiel: Ein einzelnes Feld *Patienten-ID* sollte ausreichen, um die Subjekte in einem einzelnen Krankenhaus zu definieren, doch die Kombination aus *Krankenhaus-ID* und *Patienten-ID* kann erforderlich sein, wenn die Patienten-IDs nicht krankenhausübergreifend eindeutig sind. Bei einer Einstellung mit wiederholten Messungen werden für jedes Subjekt mehrere Beobachtungen aufgezeichnet, sodass jedes Subjekt mehrere Datensätze im Dataset belegen kann.

Ein **Subjekt** ist eine Beobachtungseinheit, die als unabhängig von anderen Subjekten betrachtet werden kann. Die Blutdruckmessungen eines Patienten in einer medizinischen Studie können beispielsweise als unabhängig von den Messungen anderer Patienten angesehen werden. Die Definition von Subjekten ist vor allem dann wichtig, wenn für jedes Subjekt Messwiederholungen durchgeführt werden und Sie die Korrelation zwischen diesen Beobachtungen analysieren möchten. So ist beispielsweise zu erwarten, dass Blutdruckmessungen bei einem bestimmten Patienten bei aufeinander folgenden Arztbesuchen miteinander korrelieren.

Alle Felder, die auf der Registerkarte "Datenstruktur" als Subjekte angegeben sind, werden dazu verwendet, Subjekte für die Kovarianzstruktur der Residuen zu definieren, und stellen die Liste der möglichen Felder zum Definieren der Subjekte für die Kovarianzstrukturen der Zufallseffekte im Block für zufällige Effekte bereit.

Messwiederholung. Die hier angegebenen Felder werden verwendet, um Beobachtungswiederholungen zu kennzeichnen. So können beispielsweise mit einer einzigen Variablen für *Woche* alle 10 Wochen der Beobachtungen in einer medizinischen Studie bezeichnet werden oder die Variablen *Monat* und *Tag* können gemeinsam verwendet werden, um tägliche Beobachtungen im Verlauf eines Jahres zu bezeichnen.

Kovarianzgruppe definieren nach. Die hier angegebenen Felder definieren unabhängige Sets von Kovarianzparametern wiederholter Effekte, einen für jede Kategorie, die durch die Kreuzklassifikation der Gruppierungsfelder definiert werden. Alle Subjekte weisen denselben Kovarianztyp auf. Subjekte innerhalb derselben Kovarianzgruppierung weisen dieselben Werte für die Parameter auf.

Kovarianztyp bei Messwiederholung. Hiermit wird die Kovarianzstruktur für die Residuen angegeben. Die folgenden Strukturen sind verfügbar:

- Autoregressiv der ersten Ordnung (AR1)
- Autoregressiv mit gleitendem Durchschnitt (1,1) (ARMA11)
- Zusammengesetzt symmetrisch (ZS)

- Diagonal
- Skalierte Identität
- Toeplitz
- Unstrukturiert (UN)
- Varianzkomponenten

Ziel: Mit diesen Einstellungen werden das Ziel, seine Verteilung und seine Beziehung zu den Prädiktoren über die Verknüpfungsfunktion definiert.

Ziel. Das Ziel muss angegeben werden. Es kann jedes beliebige Messniveau aufweisen. Durch das Messniveau des Ziels wird die Menge der jeweils geeigneten Verteilungen und Verknüpfungsfunktionen eingegrenzt.

- **Anzahl der Versuche als Nenner verwenden.** Wenn die Zielantwort eine Anzahl von Ereignissen ist, die in einer Menge von Versuchen eintreten, enthält das Zielfeld die Anzahl der Ereignisse und Sie können ein zusätzliches Feld auswählen, das die Anzahl der Versuche enthält. Beim Testen eines neuen Pestizids können Sie beispielsweise Stichproben von Ameisen verschiedenen Konzentrationen des Schädlingsbekämpfungsmittels aussetzen. Zeichnen Sie dabei die Anzahl der vernichteten Ameisen und die Anzahl der Ameisen in den einzelnen Stichproben auf. In diesem Fall sollte das Feld, in dem die Zahl der vernichteten Ameisen aufgezeichnet wird, als Zielfeld (Ereignisfeld) und das Feld, in dem die Anzahl der Ameisen in den einzelnen Stichproben aufgezeichnet wird, als Feld für die Versuche festgelegt werden. Wenn die Zahl der Ameisen in den einzelnen Stichproben gleich ist, kann die Anzahl der Versuche mit einem festen Wert angegeben werden.

  Die Anzahl der Versuche sollte größer oder gleich der Anzahl der Ereignisse für die einzelnen Datensätze sein. Bei den Ereignissen sollte es sich um nicht negative Ganzzahlen und bei den Versuchen um positive Ganzzahlen handeln.
- **Referenzkategorie anpassen.** Bei einem kategorialen Ziel können Sie die Referenzkategorie auswählen. Dies kann sich auf bestimmte Ausgaben, wie beispielsweise Parameterschätzungen, auswirken, sollte jedoch nicht die Anpassungsgüte des Modells verändern. Beispiel: Angenommen, Ihr Ziel kann die Werte 0, 1 und 2 annehmen. In diesem Fall verwendet die Prozedur standardmäßig die letzte Kategorie (die mit dem höchsten Wert), also 2, als Referenzkategorie. In diesem Fall sollten Parameterschätzungen als Bezug auf die Wahrscheinlichkeit von Kategorie 0 oder 1 *im Verhältnis* zur Wahrscheinlichkeit von Kategorie 2 interpretiert werden. Wenn Sie eine benutzerdefinierte Kategorie festlegen und Ihr Ziel über definierte Beschriftungen verfügt, können Sie die Referenzkategorie festlegen, indem Sie einen Wert aus der Liste auswählen. Dies kann nützlich sein, wenn Sie bei der Festlegung eines Modells nicht mehr wissen, wie genau ein bestimmtes Feld codiert war.

Zielverteilung und Beziehung (Verknüpfung) mit dem linearen Modell. Angesichts der Werte der Prädiktoren geht das Modell davon aus, dass die Verteilung der Werte des Ziels der angegebenen Form folgt und dass die Zielwerte über die angegebene Verknüpfungsfunktion in einer linearen Beziehung zu den Prädiktoren stehen. Für mehrere allgemeine Modelle werden Verknüpfungen bereitgestellt. Sie können aber auch eine **benutzerdefinierte** Einstellung auswählen, wenn es eine bestimmte Kombination einer Verteilung und einer Verknüpfungsfunktion gibt, die Sie anpassen möchten und die nicht in der Liste der Verknüpfungen enthalten ist.

- **Lineares Modell.** Gibt eine Normalverteilung mit einer Identitätsverknüpfung an, was nützlich ist, wenn sich das Ziel mit einer linearen Regression oder einem ANOVA-Modell vorhersagen lässt.
- **Gammaregression.** Gibt eine Gammaverteilung mit einer Log-Verknüpfung an, die eingesetzt werden sollte, wenn das Ziel ausschließlich positive Werte enthält und eine Verzerrung hin zu größeren Werten aufweist.
- **Loglinear.** Gibt eine Poisson-Verteilung mit einer Log-Verknüpfung an, die eingesetzt werden sollte, wenn das Ziel eine Anzahl an Vorkommen in einem festen Zeitraum darstellt.
- **Negative binomiale Regression.** Gibt eine negative Binomialverteilung mit einer Log-Verknüpfung an, die eingesetzt werden sollte, wenn Ziel und Nenner die Anzahl der Versuche darstellen, die erforderlich sind, um *k* Erfolge zu beobachten.

- **Multinomiale logistische Regression.** Gibt eine Multinomialverteilung an, die eingesetzt werden sollte, wenn es sich bei dem Ziel um eine Antwort mit mehreren Kategorien handelt. Verwendet entweder eine Verknüpfung vom Typ "Logit (kumulativ)" (ordinale Ergebnisse) oder eine Verknüpfung vom Typ "Logit (verallgemeinert)" (nominale Antwort mit mehreren Kategorien).
- **Binäre logistische Regression.** Gibt eine Binomialverteilung mit einer Logit-Verknüpfung an, die eingesetzt werden sollte, wenn es sich bei dem Ziel um eine Binärantwort handelt, die durch ein logistisches Regressionsmodell vorhergesagt wird.
- **Binär Probit.** Gibt eine Binomialverteilung mit einer Probit-Verknüpfung an, die eingesetzt werden sollte, wenn es sich bei dem Ziel um eine Binärantwort handelt, der eine Normalverteilung zugrunde liegt.
- **Intervallzensiertes Überleben.** Gibt eine Binomialverteilung mit einer Verknüpfung vom Typ "Log-Log komplementär" an, die sinnvoll für Überlebensanalysen ist, bei denen einige Beobachtungen kein Beendigungsereignis aufweisen.

Verteilung

Diese Auswahl gibt die Verteilung des Ziels an. Die Möglichkeit einer anderen Verteilung als "Normal" und einer anderen Verknüpfungsfunktion als "Identität" ist die wichtigste Verbesserung des verallgemeinerten linearen gemischten Modells gegenüber dem linearen gemischten Modell. Es gibt zahlreiche mögliche Kombinationen aus Verteilung und Verknüpfungsfunktion und es können mehrere davon für das jeweils vorliegende Dataset geeignet sein. Daher können Sie sich in Ihrer Wahl durch theoretische Vorüberlegungen leiten lassen oder davon, welche Kombination am besten zu passen scheint.

- **Binomial.** Diese Verteilung ist nur für Ziele geeignet, die eine binäre Antwort oder eine Anzahl von Ereignissen repräsentieren.
- **Gamma.** Diese Verteilung eignet sich für Ziele mit positiven Skalenwerten, die in Richtung größerer positiver Werte verzerrt sind. Wenn ein Datenwert kleiner oder gleich 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet.
- **Invers normal.** Diese Verteilung eignet sich für Ziele mit positiven Skalenwerten, die in Richtung größerer positiver Werte verzerrt sind. Wenn ein Datenwert kleiner oder gleich 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet.
- **Multinomial.** Diese Verteilung eignet sich für ein Ziel, das eine Antwort mit mehreren Kategorien darstellt. Die Form des Modells hängt vom Messniveau des Ziels ab.

  Ein **nominales** Ziel führt zu einem nominalen multinomialen Modell, in dem für jede Kategorie des Ziels (mit Ausnahme der Referenzkategorie) ein separates Set an Modellparametern geschätzt wird. Die Parameterschätzungen für einen Prädiktor zeigen jeweils die Beziehung zwischen dem betreffenden Prädiktor und der Wahrscheinlichkeit der einzelnen Kategorien des Ziels, relativ zur Referenzkategorie.

  Ein **ordinales** Ziel führt zu einem ordinalen multinomialen Modell, in dem der herkömmliche konstante Term durch eine Menge an **Schwellen**-Parametern ersetzt ist, die sich auf die kumulative Wahrscheinlichkeit der Zielkategorien beziehen.
- **Negativ binomial.** Für die negative binomiale Regression wird eine negative Binomialverteilung mit einer Log-Verknüpfung genutzt, die verwendet werden sollte, wenn das Ziel eine Anzahl an Vorkommen mit hoher Varianz darstellt.
- **Normal.** Diese Option eignet sich für stetige Ziele, deren Werte eine symmetrische, glockenförmige Verteilung um einen Mittelwert aufweisen.
- **Poisson.** Diese Verteilung lässt sich als Anzahl der Vorkommen eines untersuchten Ereignisses in einem festen Zeitraum betrachten und eignet sich für Variablen mit nicht negativen ganzzahligen Werten. Wenn ein Datenwert keine ganze Zahl oder kleiner als 0 ist oder fehlt, wird der entsprechende Fall nicht in der Analyse verwendet.

Verknüpfungsfunktionen

Die Verknüpfungsfunktion ist eine Transformation des Ziels, die eine Schätzung des Modells ermöglicht. Die folgenden Funktionen sind verfügbar:

- **Identität.** $f(x)=x$. Das Ziel wird nicht transformiert. Diese Verknüpfung kann abgesehen von der Multinomialverteilung mit jeder beliebigen Verteilung verwendet werden.
- **Log-Log komplementär.** $f(x)=\log(-\log(1-x))$. Nur für die Binomialverteilung oder Multinomialverteilung geeignet.
- **Cauchit.** $f(x) = \tan(\pi\ (x - 0{,}5))$. Nur für die Binomialverteilung oder Multinomialverteilung geeignet.
- **Log.** $f(x)=\log(x)$. Diese Verknüpfung kann abgesehen von der Multinomialverteilung mit jeder beliebigen Verteilung verwendet werden.
- **Log-Komplement.** $f(x)=\log(1-x)$. Nur für die Binomialverteilung geeignet.
- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Nur für die Binomialverteilung oder Multinomialverteilung geeignet.
- **Log-Log negativ.** $f(x)=-\log(-\log(x))$. Nur für die Binomialverteilung oder Multinomialverteilung geeignet.
- **Probit.** $f(x)=\Phi^{-1}(x)$, wobei Φ^{-1} die kumulative Standardnormalverteilungsfunktion ist. Nur für die Binomialverteilung oder Multinomialverteilung geeignet.
- **Exponentiell.** $f(x)=x^{\alpha}$, wenn $\alpha \neq 0$. $f(x)=\log(x)$, wenn $\alpha=0$. α ist die erforderliche Zahlenangabe. Es muss sich dabei um eine reelle Zahl handeln. Diese Verknüpfung kann abgesehen von der Multinomialverteilung mit jeder beliebigen Verteilung verwendet werden.

Feste Effekte: Faktoren mit festen Effekten werden im Allgemeinen als Felder betrachtet, deren relevante Werte alle im Dataset dargestellt werden und zum Scoren verwendet werden können. Standardmäßig werden Felder mit der vordefinierten Eingaberolle, die nicht an anderer Stelle des Dialogs angegeben sind, in den Bereich des Modells eingegeben, der feste Effekte aufweist. Kategoriale Felder (Flag, nominal und ordinal) werden als Faktoren im Modell verwendet und stetige Felder werden als Kovariaten verwendet.

Geben Sie Effekte in das Modell ein, indem Sie ein oder mehrere Felder in der Quellenliste auswählen und sie in die Liste der Effekte ziehen. Welche Art von Effekt erstellt wird, hängt davon ab, auf welchem Hotspot Sie die Auswahl ablegen.

- **Haupt.** Die abgelegten Felder werden unten in der Liste der Effekte als separate Haupteffekte angezeigt.
- **2-fach.** Alle möglichen Paare der abgelegten Felder werden unten in der Liste der Effekte als Zweiwegeinteraktionen angezeigt.
- **3-fach.** Alle möglichen Dreiergruppen der abgelegten Felder werden unten in der Liste der Effekte als Dreiwegeinteraktionen angezeigt.
- ***.** Die Kombination aller abgelegten Felder wird unten in der Liste der Effekte als Einzelinteraktion angezeigt.

Schaltflächen rechts neben dem Effektgenerator ermöglichen die Ausführung verschiedener Aktionen:

Tabelle 10. Beschreibungen der Schaltflächen des Effektgenerators.

Symbol	Beschreibung
	Löschen von Termen aus dem Modell mit festen Effekten durch Auswahl der Terme, die Sie löschen möchten, und durch Klicken auf die Schaltfläche zum Löschen.
	Umordnen der Terme innerhalb des Modells mit festen Effekten, durch Auswahl der Terme, die Sie umordnen möchten, und durch Klicken auf den Aufwärts- oder Abwärtspfeil.
	Hinzufügen von verschachtelten Termen zum Modell mithilfe des Dialogfelds „Hinzufügen eines benutzerdefinierten Terms" auf Seite 201 durch Klicken auf die Schaltfläche "Benutzerdefinierten Term hinzufügen"

Konstanten Term einschließen. Der konstante Term wird für gewöhnlich in das Modell eingeschlossen. Wenn anzunehmen ist, dass die Daten durch den Koordinatenursprung verlaufen, können Sie den konstanten Term ausschließen.

Hinzufügen eines benutzerdefinierten Terms: In dieser Prozedur können Sie verschachtelte Terme für Ihr Modell erstellen. Verschachtelte Terme sind nützlich, um den Effekt eines Faktors oder einer Kovariate zu modellieren, deren Werte nicht mit den Stufen eines anderen Faktors interagieren. Eine Lebensmittelkette kann beispielsweise das Kaufverhalten ihrer Kunden in mehreren Filialen untersuchen. Da jeder Kunde nur eine dieser Filialen besucht, kann der Effekt *Kunde* als **verschachtelt innerhalb** des Effekts *Filiale* beschrieben werden.

Darüber hinaus können Sie Interaktionseffekte, wie polynomiale Terme mit derselben Kovariaten, einschließen oder dem verschachtelten Term mehrere Verschachtelungsebenen hinzufügen.

Einschränkungen. Für verschachtelte Terme gelten die folgenden Einschränkungen:

- Alle Faktoren innerhalb einer Interaktion müssen eindeutig sein. Dementsprechend ist die Angabe von *A*A* unzulässig, wenn *A* ein Faktor ist.
- Alle Faktoren innerhalb eines verschachtelten Effekts müssen eindeutig sein. Dementsprechend ist die Angabe von *A(A)* unzulässig, wenn *A* ein Faktor ist.
- Effekte dürfen nicht in einer Kovariaten verschachtelt werden. Dementsprechend ist die Angabe von *A(X)* unzulässig, wenn *A* ein Faktor und *X* eine Kovariate ist.

Erstellen eines verschachtelten Terms

1. Wählen Sie einen Faktor oder eine Kovariate aus, der bzw. die in einem anderen Faktor verschachtelt ist, und klicken Sie auf die Pfeilschaltfläche.
2. Klicken Sie auf **(Innerhalb)**.
3. Wählen Sie den Faktor aus, in dem der vorherige Faktor bzw. die vorherige Kovariate verschachtelt ist, und klicken Sie dann auf die Pfeilschaltfläche.
4. Klicken Sie auf **Term hinzufügen**.

Optional können Sie Interaktionseffekte einschließen oder dem verschachtelten Term mehrere Verschachtelungsebenen hinzufügen.

Zufällige Effekte: Zufallseffektfaktoren sind Felder, deren Wert in der Datendatei als zufällige Stichprobe aus einer größeren Gesamtheit von Werten betrachtet werden kann. Sie helfen bei der Erklärung von übermäßiger Variabilität beim Ziel. Wenn Sie auf der Registerkarte "Datenstruktur" mehrere Objekte ausgewählt haben, wird standardmäßig für jedes Subjekt, das über das innerste Subjekt hinausgeht, ein Block mit zufälligen Effekten erstellt. Wenn Sie beispielsweise auf der Registerkarte "Datenstruktur" als Subjekte "Schule", "Klasse" und "Schüler" ausgewählt haben, werden automatisch die folgenden Blöcke für zufällige Effekte erstellt:

- Zufälliger Effekt 1: Subjekt ist Schule (ohne Effekte, nur konstanter Term)
- Zufälliger Effekt 2: Subjekt ist Schule * Klasse (keine Effekte, nur konstanter Term)

Sie können wie folgt mit Blöcken mit zufälligen Effekten arbeiten:

1. Klicken Sie zum Hinzufügen eines neuen Blocks auf **Block hinzufügen...** Hierdurch wird das Dialogfeld „Block für zufällige Effekte” auf Seite 202 geöffnet.
2. Wählen Sie zum Bearbeiten eines vorhandenen Blocks den zu bearbeitenden Block aus und klicken Sie auf **Block bearbeiten...** Hierdurch wird das Dialogfeld „Block für zufällige Effekte” auf Seite 202 geöffnet.
3. Um einen oder mehrere Blöcke zu löschen, wählen Sie die betreffenden Blöcke aus und klicken Sie auf die Löschschaltfläche.

Block für zufällige Effekte: Geben Sie Effekte in das Modell ein, indem Sie ein oder mehrere Felder in der Quellenliste auswählen und sie in die Liste der Effekte ziehen. Welche Art von Effekt erstellt wird, hängt davon ab, auf welchem Hotspot Sie die Auswahl ablegen. Kategoriale Felder (Flag, nominal und ordinal) werden als Faktoren im Modell verwendet und stetige Felder werden als Kovariaten verwendet.

- **Haupt.** Die abgelegten Felder werden unten in der Liste der Effekte als separate Haupteffekte angezeigt.
- **2-fach.** Alle möglichen Paare der abgelegten Felder werden unten in der Liste der Effekte als Zweiwegeinteraktionen angezeigt.
- **3-fach.** Alle möglichen Dreiergruppen der abgelegten Felder werden unten in der Liste der Effekte als Dreiwegeinteraktionen angezeigt.
- ***.** Die Kombination aller abgelegten Felder wird unten in der Liste der Effekte als Einzelinteraktion angezeigt.

Schaltflächen rechts neben dem Effektgenerator ermöglichen die Ausführung verschiedener Aktionen:

Tabelle 11. Beschreibungen der Schaltflächen des Effektgenerators.

Symbol	Beschreibung
	Löschen Sie Terme aus dem Modell, indem Sie die Terme, die Sie löschen möchten, auswählen und auf die Schaltfläche zum Löschen klicken.
	Ordnen Sie die Terme innerhalb des Modells neu, indem Sie die Terme auswählen, die Sie neu ordnen möchten, und auf den Aufwärts- oder Abwärtspfeil klicken.
	Hinzufügen von verschachtelten Termen zum Modell mithilfe des Dialogfelds „Hinzufügen eines benutzerdefinierten Terms" auf Seite 201 durch Klicken auf die Schaltfläche "Benutzerdefinierten Term hinzufügen"

Konstanten Term einschließen. Der konstante Term ist nicht standardmäßig im Modell mit zufälligen Effekten enthalten. Wenn anzunehmen ist, dass die Daten durch den Koordinatenursprung verlaufen, können Sie den konstanten Term ausschließen.

Kovarianzgruppe definieren nach. Die hier angegebenen Felder definieren unabhängige Sets von Kovarianzparametern zufälliger Effekte, einen für jede Kategorie, die durch die Kreuzklassifikation der Gruppierungsfelder definiert wird. Für die einzelnen Blöcke für zufällige Effekte können unterschiedliche Mengen von Gruppierungsfeldern festgelegt werden. Alle Subjekte weisen denselben Kovarianztyp auf. Subjekte innerhalb derselben Kovarianzgruppierung weisen dieselben Werte für die Parameter auf.

Subjektkombination. Hiermit können Sie eine Auswahl aus vorab festgelegten Kombinationen von Subjekten mit zufälligen Effekten aus der Registerkarte "Datenstruktur" treffen. Wenn beispielsweise auf der Registerkarte "Datenstruktur" die Subjekte *Schule*, *Klasse* und *Schüler* (in dieser Reihenfolge) definiert sind, enthält die Dropdown-Liste für die Subjektkombinationen folgende Optionen: **Keines**, **Schule**, **Schule * Klasse** und **Schule * Klasse * Schüler**.

Kovarianztyp für Zufallseffekte. Hiermit wird die Kovarianzstruktur für die Residuen angegeben. Die folgenden Strukturen sind verfügbar:

- Autoregressiv der ersten Ordnung (AR1)
- Autoregressiv mit gleitendem Durchschnitt (1,1) (ARMA11)
- Zusammengesetzt symmetrisch (ZS)
- Diagonal
- Skalierte Identität

- Toeplitz
- Unstrukturiert (UN)
- Varianzkomponenten

Gewichtung und Offset: Analysegewichtung. Der Skalenparameter ist ein geschätzter Modellparameter, der mit der Varianz der Antwort zusammenhängt. Die Analysegewichtungen sind "bekannte" Werte, die sich zwischen den einzelnen Beobachtungen unterscheiden können. Wenn das Feld "Analysegewichtung" angegeben ist, wird der Skalenparameter, der mit der Varianz der Antwort zusammenhängt, für jede Beobachtung durch die Werte für die Analysegewichtung geteilt. Datensätze, deren Analysegewichtungswerte kleiner oder gleich 0 sind oder fehlen, werden in der Analyse nicht verwendet.

Offset. Der Term "Offset" ist einer "struktureller" Prädiktor. Ihr Koeffizient wird nicht vom Modell geschätzt, sondern es wird davon ausgegangen, dass er den Wert 1 aufweist. Daher werden die Werte des Offsets einfach zur linearen Prädiktoren des Ziels addiert. Dies ist besonders nützlich bei Poisson-Regressionsmodellen, bei denen die verschiedenen Fälle dem relevanten Ereignis unterschiedlich stark ausgesetzt sein können

Beispielsweise gibt es bei der Modellierung der Unfallraten für einzelne Fahrer einen wichtigen Unterschied zwischen einem Fahrer, der in 3 Jahren Fahrpraxis einen Unfall verursacht hat und einem Fahrer, der in 25 Jahren einen Unfall verursacht hat. Die Anzahl der Unfälle kann als Poisson-Antwort oder als negative binomiale Antwort mit einer Log-Verknüpfung angezeigt werden, wenn der natürliche Logarithmus der Fahrpraxis des Fahrers als Offset-Term eingeschlossen wird.

Bei anderen Kombinationen von Verteilung und Verknüpfungstypen wären andere Transformationen der Offset-Variablen erforderlich.

Erstellungsoptionen: Mit dieser Auswahl werden einige erweiterte Kriterien angegeben, die für die Erstellung des Modells verwendet werden.

Sortierreihenfolge. Mit diesen Steuerelementen wird die Reihenfolge der Kategorien für die Ziele und Faktoren (kategoriale Eingaben) festgelegt, um die "letzte" Kategorie zu ermitteln. Die Einstellung für die Sortierreihenfolge des Ziels wird ignoriert, wenn das Ziel nicht kategorial ist oder wenn in den Einstellungen „Ziel” auf Seite 198 eine benutzerdefinierte Referenzkategorie angegeben wird.

Stoppregeln. Sie können die maximale Anzahl an Iterationen angeben, die im Algorithmus vorgenommen werden. Geben Sie eine nicht negative Ganzzahl an. Der Standardwert ist 100.

Einstellungen nach der Schätzung. Mit diesen Einstellungen wird festgelegt, wie ein Teil der Modellausgabe für die Anzeige berechnet wird.

- **Konfidenzniveau.** Das Konfidenzniveau wird zur Berechnung der Intervallschätzungen der Modellkoeffizienten verwendet. Geben Sie einen Wert größer 0 und kleiner 100 ein. Der Standardwert ist 95.
- **Freiheitsgrade.** Damit wird angegeben, wie Freiheitsgrade für Signifikanztests berechnet werden. Wählen Sie die Option **Fest für alle Tests (Residualmethode)**, wenn Ihre Stichprobe groß genug ist oder die Daten ausgewogen sind oder das Modell einen einfacheren Kovarianztyp verwendet, z. B. "Skalierte Identität" oder "Diagonal"). Dies ist die Standardeinstellung. Wählen Sie die Option **Zwischen den einzelnen Tests unterschiedlich (Satterthwaite-Approximation)**, wenn Ihre Stichprobe klein ist oder die Daten unausgewogen sind oder das Modell einen komplizierten Kovarianztyp verwendet, z. B. "Unstrukturiert").
- **Test für feste Effekte und Koeffizienten.** Dies ist die Methode zur Berechnung der Kovarianzmatrix der Parameterschätzungen. Wählen Sie die robuste Schätzung, wenn Sie befürchten, dass die Modellannahmen verletzt sein könnten.

Allgemein: Modellname. Sie können den Modellnamen automatisch basierend auf den Zielfeldern generieren, oder einen benutzerdefinierten Namen angeben. Der automatisch generierte Name ist der Zielfeld-

name. Bei mehreren Zielen besteht der Modellname aus den Feldnamen, die der Reihe nach durch Und-Zeichen verbunden werden. Wenn beispielsweise*Feld1 Feld2 Feld3* Ziele sind, lautet der Modellname *Feld1 & Feld2 & Feld3*.

Für Scoring bereitstellen. Beim Scoren des Modells sollten die ausgewählten Elemente in dieser Gruppe erzeugt werden. Wenn das Modell gescort wird, werden stets der vorhergesagte Wert (für alle Ziele) und die Konfidenz (für stetige Ziele) berechnet. Die berechnete Konfidenz kann auf der Wahrscheinlichkeit des vorhergesagten Werts (die höchste vorhergesagte Wahrscheinlichkeit) oder der Differenz zwischen der höchsten vorhergesagten Wahrscheinlichkeit und der zweithöchsten vorhergesagten Wahrscheinlichkeit basieren.

- **Vorhergesagte Wahrscheinlichkeit für kategoriale Ziele.** Mit dieser Option werden die vorhergesagten Wahrscheinlichkeiten für kategoriale Ziele produziert. Für jede Kategorie wird ein Feld erstellt.
- **Propensity-Scores für Flagziele.** Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Das Modell erzeugt Raw-Propensity-Scores. Bei aktiven Partitionen erzeugt das Modell außerdem Adjusted-Propensitys-Scores anhand der Testpartition.

Geschätzte Mittelwerte: Auf dieser Registerkarte können Sie die geschätzten Randmittel für die Ebenen der Faktoren und die Faktorinteraktionen anzeigen. Geschätzte Randmittel sind für multinomiale Modelle nicht verfügbar.

Terme. Die Modellterme in den festen Effekten, die gänzlich aus kategorialen Feldern bestehen, werden hier aufgeführt. Markieren Sie jeden Term, für den das Modell geschätzte Randmittel erstellen soll.

- **Kontrasttyp.** Damit wird der Kontrasttyp angegeben, der für die Stufen des Kontrastfelds verwendet wird. Wenn **Keine** ausgewählt ist, werden keine Kontraste erstellt. **Paarweise** führt zu paarweisen Vergleichen für Kombinationen über alle Stufen der angegebenen Faktoren. Dies ist der einzige verfügbare Kontrast für Interaktionen zwischen Faktoren. Bei Kontrasten vom Typ **Abweichung** wird jede Faktorstufe mit dem Gesamtmittelwert verglichen. Kontraste des Typs **Einfach** vergleichen jede Stufe des Faktors, mit Ausnahme der letzten, mit der letzten Stufe. Die "letzte" Stufe wird durch die Sortierreihenfolge für Faktoren bestimmt, die in den Erstellungsoptionen festgelegt ist. Beachten Sie, dass keiner dieser Kontrasttypen orthogonal ist.
- **Kontrastfeld.** Hier wird ein Faktor angegeben, dessen Stufen mithilfe des ausgewählten Kontrasttyps verglichen werden. Wenn **Keine** als Kontrasttyp ausgewählt ist, kann (bzw. muss) kein Kontrastfeld ausgewählt werden.

Stetige Felder. Die aufgeführten stetigen Felder werden aus den Termen in den festen Effekten extrahiert, bei denen stetige Felder verwendet werden. Bei der Berechnung der geschätzten Randmittel sind die Kovariaten auf die angegebenen Werte festgelegt. Wählen Sie den Mittelwert aus oder geben Sie einen benutzerdefinierten Wert an.

Geschätzte Mittelwerte anzeigen als. Damit wird angegeben, ob geschätzte Randmittel anhand der ursprünglichen Skala des Ziels oder anhand der Transformation der Verknüpfungsfunktion berechnet werden sollen. **Ursprüngliche Zielskala** berechnet geschätzte Randmittel für das Ziel. Beachten Sie: Wenn das Ziel mithilfe der Option für Ereignisse/Versuche angegeben wird, werden hiermit die geschätzten Randmittel für das Verhältnis Ereignisse/Versuche ausgegeben und nicht für die Anzahl der Ereignisse. **Transformation einer Verknüpfungsfunktion** berechnet geschätzte Randmittel für den linearen Prädiktor.

Anpassen für Mehrfachvergleiche mithilfe von. Bei der Durchführung von Hypothesentests mit mehreren Kontrasten kann das Gesamtsignifikanzniveau mithilfe der Signifikanzniveaus der eingeschlossenen Kontraste angepasst werden. Damit können Sie die Anpassungs-/Korrekturmethode auswählen.

- **Geringste signifikante Differenz.** Bei dieser Methode wird nicht die Gesamtwahrscheinlichkeit für das Verwerfen der Hypothesen reguliert, dass einige lineare Kontraste von den Werten der Nullhypothese abweichen.

- *Bonferroni sequenziell*. Hierbei handelt es sich um ein sequentielles schrittweises Bonferroni-Verfahren, das deutlich weniger konservativ ist was die Ablehnung einzelner Hypothesen anbelangt, aber dennoch dasselbe allgemeine Signifikanzniveau beibehält.
- *Sidak (sequenziell)*. Hierbei handelt es sich um ein sequentielles schrittweises Sidak-Verfahren, das deutlich weniger konservativ ist was die Ablehnung einzelner Hypothesen anbelangt, aber dennoch dasselbe allgemeine Signifikanzniveau beibehält.

Die Methode der geringsten signifikanten Differenz ist weniger konservativ als die sequenzielle Sidak-Methode, die wiederum weniger konservativ ist als die sequenzielle Bonferroni-Methode. Bei der Methode der geringsten signifikanten Differenz werden also mindestens so viele einzelne Hypothesen verworfen werden wie bei der sequenziellen Sidak-Methode, bei der wiederum mindestens so viele einzelne Hypothesen verworfen werden wie bei der sequenziellen Bonferroni-Methode.

Modellansicht: Standardmäßig wird die Ansicht "Modellübersicht" angezeigt. Um eine andere Modellansicht anzuzeigen, wählen Sie sie aus den Piktogrammen aus.

Modellübersicht: Diese Ansicht bietet eine übersichtliche und aktuelle Zusammenfassung des Modells und seiner Anpassungsgüte.

Tabelle. In der Tabelle sind das Ziel, die Wahrscheinlichkeitsverteilung und die Verknüpfungsfunktion zu sehen, die in den Zieleinstellungen angegeben wurden. Wenn das Ziel durch Ereignisse und Versuche definiert ist, wird die Zelle aufgeteilt und zeigt das Ereignisfeld und das Feld für die Versuche oder eine feste Anzahl von Versuchen an. Außerdem werden das Akaike-Informationskriterium mit Korrektur für endliche Stichproben (AICC) und das Bayes-Informationskriterium (BIC) angezeigt.

- *Akaike (korrigiert)*. Ein Maß für die Auswahl und den Vergleich von gemischten Modellen, das auf -2 (Restricted) Log-Likelihood beruht. Kleinere Werte stehen für bessere Modelle. Das AICC "korrigiert" das AIC für kleine Stichprobenumfänge. Wenn die Stichprobengröße zunimmt, konvergiert das AICC zu dem AIC.
- *Bayes*. Ein Maß für die Auswahl und den Vergleich von Modellen, das auf -2 Log-Likelihood beruht. Kleinere Werte stehen für bessere Modelle. Das BIC bestraft ebenfalls überparametrisierte Modelle, und zwar stärker als das AIC.

Diagramme. Bei kategorialen Zielen zeigt ein Diagramm die Genauigkeit des endgültigen Modells an, also den Prozentsatz der korrekten Klassifizierungen.

Datenstruktur: Diese Ansicht bietet einen Überblick über die von Ihnen angegebene Datenstruktur und hilft Ihnen bei der Überprüfung, ob die Subjekte und Messwiederholungen richtig angegeben wurden. Die beobachteten Informationen für das erste Subjekt werden für jedes Subjektfeld und Messwiederholungsfeld sowie für das Ziel angegeben. Außerdem wird die Anzahl der Stufen für die einzelnen Subjektfelder und Messwiederholungsfelder angezeigt.

Vorhergesagt/Beobachtet: Für stetige Ziele, einschließlich Zielen, die als Ereignisse/Versuche angegeben sind, zeigt diese Ansicht ein klassiertes Streudiagramm, das die vorhergesagten Werte auf der vertikalen Achse in Abhängigkeit von den beobachteten Werten auf der horizontalen Achse darstellt. Idealerweise sollten die Werte entlang einer 45-Grad-Linie liegen. In dieser Ansicht können Sie erkennen, ob bestimmte Datensätze vom Modell besonders schlecht vorhergesagt werden.

Klassifizierung: Bei kategorialen Zielen wird hiermit die Kreuzklassifikation der beobachteten Werte in Abhängigkeit von den vorhergesagten Werten in einer Heat-Map angezeigt, zuzüglich des Gesamtprozentsatzes der korrekten Werte.

Tabellenstile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Zeilenprozentwerte.** Hiermit werden die Zeilenprozentsätze (die Zellenhäufigkeiten ausgedrückt als Prozentsatz der Gesamtzeilenzahl) in den Zellen angezeigt. Dies ist die Standardeinstellung.

- **Zellenhäufigkeit.** Hiermit werden die Zellenhäufigkeiten in den Zellen angezeigt. Die Schattierung für die Heat-Map beruht weiterhin auf den Zeilenprozentsätzen.
- **Heat-Map.** Hiermit werden in den Zellen keine Werte, sondern nur die Schattierung angezeigt.
- **Komprimiert.** Hiermit werden keine Zeilen- oder Spaltenüberschriften oder Werte in den Zellen angezeigt. Diese Option kann nützlich sein, wenn das Ziel sehr viele Kategorien aufweist.

Fehlende Werte. Wenn Datensätze fehlende Werte im Ziel aufweisen, werden diese unter allen gültigen Werten in der Zeile **(Fehlend)** angezeigt. Datensätze mit fehlenden Werten tragen nicht zum Wert von "Gesamtprozent korrekt" bei.

Mehrere Ziele. Bei mehreren kategorialen Zielen wird jedes Ziel in einer separaten Tabelle dargestellt. Dazu gibt es eine Dropdown-Liste **Ziel**, aus der das anzuzeigende Ziel ausgewählt werden kann.

Große Tabellen. Wenn das angezeigte Ziel mehr als 100 Kategorien enthält, wird keine Tabelle angezeigt.

Feste Effekte: Diese Ansicht zeigt die Größe der einzelnen festen Effekte im Modell.

Stile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Diagramm.** In diesem Diagramm werden Effekte von oben nach unten in der Reihenfolge sortiert, in der sie in den Einstellungen "Feste Effekte" angegeben wurden. Verbindungslinien im Diagramm sind basierend auf der Effektsignifikanz gewichtet, wobei eine größere Linienbreite signifikanteren Effekten entspricht (kleinere *p*-Werte). Dies ist die Standardeinstellung.
- **Tabelle.** Diese Ansicht zeigt eine ANOVA-Tabelle für das Gesamtmodell und die einzelnen Modelleffekte. Die einzelnen Effekte sind von oben nach unten in der Reihenfolge sortiert, in der sie in den Einstellungen unter "Feste Effekte" angegeben wurden.

Signifikanz. Mit dem Schieberegler "Signifikanz" wird gesteuert, welche Effekte in der Ansicht angezeigt werden. Effekte, deren Signifikanzwerte größer als der Wert des Schiebereglers sind, werden ausgeblendet. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Effekte konzentrieren. Standardmäßig ist der Wert 1,00 eingestellt, sodass keine Effekte basierend auf der Signifikanz herausgefiltert werden.

Feste Koeffizienten: Diese Ansicht zeigt den Wert der einzelnen festen Koeffizienten im Modell. Hinweis: Faktoren (kategoriale Prädiktoren) sind innerhalb des Modells indikatorcodiert, sodass Faktoren, die **Effekte** enthalten, in der Regel mehrere zugehörige **Koeffizienten** aufweisen. Mit Ausnahme der Kategorie für den redundanten Koeffizienten erhält jede Kategorie einen solchen Koeffizienten.

Stile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Diagramm.** In diesem Diagramm wird zuerst der konstante Term angezeigt. Anschließend werden die Effekte von oben nach unten in der Reihenfolge sortiert, in der sie in den Einstellungen "Feste Effekte" angegeben wurden. In Faktoren, die Effekte enthalten, werden die Koeffizienten in aufsteigender Reihenfolge der Datenwerte sortiert. Verbindungslinien im Diagramm sind farbig dargestellt und basierend auf der Koeffizientensignifikanz gewichtet, wobei eine größere Linienbreite signifikanteren Koeffizienten entspricht (kleinere *p*-Werte). Dies ist der Standardstil.
- **Tabelle.** Diese Tabelle zeigt die Werte, Signifikanztests und Konfidenzintervalle für die einzelnen Modellkoeffizienten. Nach dem konstanten Term sind die einzelnen Effekte von oben nach unten in der Reihenfolge sortiert, in der sie in den Einstellungen unter "Feste Effekte" angegeben wurden. In Faktoren, die Effekte enthalten, werden die Koeffizienten in aufsteigender Reihenfolge der Datenwerte sortiert.

Multinomial. Wenn die Multinomialverteilung in Kraft ist, steuert die Dropdown-Liste "Multinomial", welche Zielkategorie angezeigt werden soll. Die Sortierreihenfolge der Werte in der Liste richtet sich nach den Angaben unter "Erstellungsoptionen".

Exponentialverteilung. Hiermit werden exponentielle Koeffizientenschätzungen und Konfidenzintervalle für bestimmte Modelltypen angegeben, einschließlich "Binäre logistische Regression" (Binomialverteilung und Logit-Verknüpfung), "Nominale logistische Regression" (multinomiale Verteilung und Logit-Verknüpfung), "Negative binomiale Regression" (negative Binomialverteilung und Log-Verknüpfung) und Log-Linear-Modell (Poisson-Verteilung und Log-Verknüpfung).

Signifikanz. Mit dem Schieberegler "Signifikanz" wird gesteuert, welche Koeffizienten in der Ansicht angezeigt werden. Koeffizienten, deren Signifikanzwerte größer als der Wert des Schiebereglers sind, werden ausgeblendet. Dadurch wird das Modell nicht verändert, doch Sie können sich ganz problemlos auf die wichtigsten Koeffizienten konzentrieren. Standardmäßig ist der Wert 1,00 eingestellt, sodass keine Koeffizienten basierend auf der Signifikanz herausgefiltert werden.

Kovarianzen der zufälligen Effekte: Damit wird die Kovarianzmatrix der Zufallseffekte (**G**) angezeigt.

Stile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Kovarianzwerte.** Dies ist eine Heat-Map der Kovarianzmatrix, in der Effekte von oben nach unten in der Reihenfolge sortiert sind, in der sie in den Einstellungen "Feste Effekte" angegeben wurden. Die Farben im Korrelogramm entsprechen den im Schlüssel angezeigten Zellenwerten. Dies ist die Standardeinstellung.
- **Korrelogramm.** Eine Heat-Map der Kovarianzmatrix.
- **Komprimiert.** Eine Heat-Map der Kovarianzmatrix ohne Zeilen- und Spaltenüberschriften.

Blöcke. Wenn mehrere Blöcke für zufällige Effekte vorhanden sind, gibt es eine Dropdown-Liste mit dem Titel "Block", die zur Auswahl des anzuzeigenden Blocks dient.

Gruppen. Wenn ein Block für zufällige Effekte eine Gruppenspezifikation aufweist, gibt es eine Dropdown-Liste mit dem Titel "Gruppe", in der die anzuzeigende Stufe ausgewählt werden kann.

Multinomial. Wenn die Multinomialverteilung in Kraft ist, steuert die Dropdown-Liste "Multinomial", welche Zielkategorie angezeigt werden soll. Die Sortierreihenfolge der Werte in der Liste richtet sich nach den Angaben unter "Erstellungsoptionen".

Kovarianzparameter: Diese Ansicht zeigt die Kovarianzparameterschätzungen und verwandte Statistiken für Residuum und Zufallseffekte. Dies sind erweiterte, aber dennoch grundlegende Ergebnisse, die Informationen darüber bieten, ob die Kovarianzstruktur geeignet ist.

Zusammenfassende Tabelle. Dies ist eine Kurzübersicht für die Anzahl der Parameter in den Kovarianzmatrizen für das Residuum (**R**) und den Zufallseffekt (**G**), den Rang (Anzahl der Spalten) in den Designmatrizen für den festen Effekt (**X**) und den Zufallseffekt (**Z**) und die Anzahl der Subjekte, die durch die Subjektfelder zur Definition der Datenstruktur festgelegt sind.

Tabelle für Kovarianzparameter. Für den ausgewählten Effekt werden die Schätzung, der Standardfehler und das Konfidenzintervall für jeden einzelnen Kovarianzparameter angezeigt. Die Anzahl der angezeigten Parameter hängt von der Kovarianzstruktur für den Effekt und, bei Blöcken für zufällige Effekte, von der Anzahl der Effekte im Block ab. Wenn Sie sehen, dass die außerhalb der Diagonalen liegenden Parameter nicht signifikant sind, können Sie eine einfachere Kovarianzstruktur verwenden.

Effekte. Wenn Blöcke für zufällige Effekte vorhanden sind, gibt es eine Dropdown-Liste mit dem Titel "Effekt", die zur Auswahl des anzuzeigenden Residuums oder des Blocks für zufällige Effekte dient. Der Residualeffekt ist immer verfügbar.

Gruppen. Wenn ein Residuum oder ein Block für zufällige Effekte eine Gruppenspezifikation aufweist, gibt es eine Dropdown-Liste mit dem Titel "Gruppe", in der die anzuzeigende Gruppenstufe ausgewählt werden kann.

Multinomial. Wenn die Multinomialverteilung in Kraft ist, steuert die Dropdown-Liste "Multinomial", welche Zielkategorie angezeigt werden soll. Die Sortierreihenfolge der Werte in der Liste richtet sich nach den Angaben unter "Erstellungsoptionen".

Geschätzte Mittelwerte: Signifikante Effekte: Hierbei handelt es sich um Diagramme, die für die 10 "signifikantesten" festen Effekte für alle Faktoren angezeigt werden, zuerst kommen die Dreifach-Interaktionen, dann die Zweifach-Interaktionen und schließlich die Haupteffekte. Das Diagramm zeigt den vom Modell geschätzten Zielwert auf der vertikalen Achse für jeden Wert des Haupteffekts (bzw. des ersten aufgeführten Effekts in einer Interaktion) auf der horizontalen Achse an. Für jeden Wert des zweiten in einer Interaktion aufgelisteten Effekts wird eine separate Linie erstellt. Für jeden Wert des dritten aufgelisteten Effekts in einer Dreifach-Interaktion wird ein separates Diagramm erstellt. Alle anderen Prädiktoren werden konstant gehalten. Es gewährt eine nützliche Visualisierung der Effekte der einzelnen Prädiktorkoeffizienten auf dem Ziel. Beachten Sie: Wenn keine Prädiktoren signifikant sind, werden keine geschätzten Mittel produziert.

Konfidenz. Hierdurch werden die oberen und unteren Konfidenzgrenzen für die Randmittel angezeigt, wobei das im Rahmen der Erstellungsoptionen angegebene Konfidenzniveau verwendet wird.

Geschätzte Mittelwerte: Benutzerdefinierte Effekte: Hierbei handelt es sich um Tabellen und Diagramme für vom Benutzer angeforderte feste Effekte für alle Faktoren.

Stile. Es sind verschiedene Anzeigestile verfügbar, auf die über die Dropdown-Liste **Stil** zugegriffen werden kann.

- **Diagramm.** Dieser Still zeigt ein Liniendiagramm des vom Modell geschätzten Zielwerts auf der vertikalen Achse für jeden Wert des Haupteffekts (oder des ersten in einer Interaktion aufgelisteten Effekts) auf der horizontalen Achse an. Für jeden Wert des zweiten aufgelisteten Effekts in einer Interaktion wird eine separate Linie erzeugt. Für jeden Wert des dritten aufgelisteten Effekts in einer Dreiwegeinteraktion wird ein separates Diagramm erzeugt. Alle anderen Prädiktoren werden konstant gehalten.

  Wenn Kontraste angefordert wurden, wird ein weiteres Diagramm erstellt, um die Stufen für das Kontrastfeld zu vergleichen. Bei Interaktionen wird ein Diagramm für jede Stufenkombination der Effekte (abgesehen vom Kontrastfeld) angezeigt. Bei **paarweisen** Kontrasten handelt es sich um ein Abstandsnetzdiagramm, also um eine grafische Darstellung der Vergleichstabelle, in der die Abstände zwischen Knoten im Netz den Unterschieden zwischen Stichproben entsprechen. Gelbe Linien entsprechen statistisch signifikanten Unterschieden, schwarze Linien nicht signifikanten Unterschieden. Wenn Sie die Maus über eine Linie im Netz bewegen, wird eine QuickInfo mit der angepassten Signifikanz des Unterschieds zwischen den durch die Linie verbundenen Knoten angezeigt.

  Bei **Abweichungs**-Kontrasten wird ein Balkendiagramm mit dem vom Modell geschätzten Zielwert der vertikalen Achse und den Werten des Kontrastfelds auf der horizontalen Achse angezeigt. Bei Interaktionen wird ein Diagramm für jede Stufenkombination der Effekte (abgesehen vom Kontrastfeld) angezeigt. Die Balken zeigen die Differenz zwischen den einzelnen Stufen des Kontrastfelds und dem Gesamtmittelwert, der durch eine schwarze horizontale Linie dargestellt wird.

  Bei **einfachen** Kontrasten wird ein Balkendiagramm mit dem vom Modell geschätzten Zielwert der vertikalen Achse und den Werten des Kontrastfelds auf der horizontalen Achse angezeigt. Bei Interaktionen wird ein Diagramm für jede Stufenkombination der Effekte (abgesehen vom Kontrastfeld) angezeigt. Die Balken zeigen die Differenz zwischen den einzelnen Stufen des Kontrastfelds (abgesehen von der letzten) und der letzten Stufe, die durch eine schwarze horizontale Linie dargestellt wird.

- **Tabelle.** Dieser Stil zeigt eine Tabelle des vom Modell geschätzten Zielwerts, dem zugehörigen Standardfehler und dem Konfidenzintervall für alle Stufenkombinationen der Felder im Effekt an. Alle anderen Prädiktoren werden konstant gehalten.

  Wenn Kontraste angefordert wurden, wird eine weitere Tabelle mit Schätzung, Standardfehler, Signifikanztest und Konfidenzintervall für jeden Kontrast angezeigt. Bei Interaktionen wird eine separate Gruppe von Zeilen für jede Stufenkombination der Effekte (abgesehen vom Kontrastfeld) angezeigt. Außerdem wird eine Tabelle mit den Gesamttestergebnissen angezeigt. Bei Interaktionen gibt es einen separaten Gesamttest für jede Stufenkombination der Effekte (abgesehen vom Kontrastfeld).

Konfidenz. Hierdurch wird die Anzeige der oberen und unteren Konfidenzgrenzen für die Randmittel ein- und ausgeschaltet, wobei das im Rahmen der Erstellungsoptionen angegebene Konfidenzniveau verwendet wird.

Layout. Hierdurch wird das Layout für das Diagramm der paarweisen Kontraste ein- und ausgeschaltet. Das Kreislayout zeigt weniger Kontraste als das Netzlayout, vermeidet jedoch das Überlappen von Linien.

Einstellungen: Beim Scoring des Modells sollten die ausgewählten Objekte in dieser Registerkarte produziert werden. Wenn das Modell gescort wird, werden stets der vorhergesagte Wert (für alle Ziele) und die Konfidenz (für stetige Ziele) berechnet. Die berechnete Konfidenz kann auf der Wahrscheinlichkeit des vorhergesagten Werts (die höchste vorhergesagte Wahrscheinlichkeit) oder der Differenz zwischen der höchsten vorhergesagten Wahrscheinlichkeit und der zweithöchsten vorhergesagten Wahrscheinlichkeit basieren.

- **Vorhergesagte Wahrscheinlichkeit für kategoriale Ziele.** Mit dieser Option werden die vorhergesagten Wahrscheinlichkeiten für kategoriale Ziele produziert. Für jede Kategorie wird ein Feld erstellt.
- **Propensity-Scores für Flagziele.** Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Das Modell erzeugt Raw-Propensity-Scores. Bei aktiven Partitionen erzeugt das Modell außerdem Adjusted-Propensitys-Scores anhand der Testpartition.

Cox-Knoten

Die Cox-Regression erstellt ein Vorhersagemodell für Daten, die die Zeit bis zum Eintreten des Ereignisses angeben. Das Modell erzeugt eine Überlebensfunktion, die die Wahrscheinlichkeit vorhersagt, mit der das interessierende Ereignis zu einer gegebenen Zeit *t* für vorgegebene Werte der Prädiktorvariablen aufgetreten ist. Die Form der Überlebensfunktion und die Regressionskoeffizienten für die Prädiktoren werden aus beobachteten Subjekten geschätzt. Anschließend kann das Modell auf neue Fälle angewendet werden, die Messungen für die Prädiktorvariablen enthalten. Beachten Sie, dass Informationen aus zensierten Subjekten, also Subjekten, bei denen das relevante Ereignis während der Beobachtungszeit nicht eintritt, einen nützlichen Beitrag zur Schätzung des Modells leisten.

Beispiel. Im Rahmen seiner Bemühungen zur Reduzierung der Kundenabwanderung ist ein Telekommunikationsunternehmen daran interessiert, die "Zeit bis zur Abwanderung" zu modellieren, um die Faktoren zu ermitteln, die für Kunden gelten, die rasch zu einem anderen Dienst wechseln. Dazu wird eine Zufallsstichprobe von Kunden ausgewählt und ihre Zeit als Kunden (unabhängig davon, ob sie noch immer aktive Kunden sind) und verschiedene demografische Felder werden aus der Datenbank extrahiert.

Anforderungen. Es werden mindestens ein Eingabefeld und genau ein Zielfeld benötigt und Sie müssen innerhalb des Cox-Knotens ein Feld für die Überlebenszeit angeben. Das Zielfeld sollte so codiert sein, dass der Wert "falsch" Überleben anzeigt und der Wert "wahr" anzeigt, dass das relevante Ereignis eingetreten ist. Das Feld muss ein Messniveau des Typs *Flag* mit dem Speichertyp "Zeichenfolge" oder "Ganze Zahl" aufweisen. (Der Speichertyp kann, falls erforderlich, mithilfe eines Füller- oder Ableitungsknotens konvertiert werden.) Felder, die auf *Beides* oder *Keine* gesetzt sind, werden ignoriert. Bei den im Modell verwendeten Feldern müssen die Typen vollständig instanziiert sein. Bei der Überlebenszeit kann es sich um ein beliebiges numerisches Feld handeln.

Datums- und Zeitangaben. Felder vom Typ "Datum und Uhrzeit" können nicht direkt zum Definieren der Überlebenszeit verwendet werden. Wenn Felder vom Typ "Datum und Uhrzeit" vorliegen, sollten Sie sie verwenden, um ein Feld mit Überlebenszeiten zu erstellen, das auf dem Unterschied zwischen dem Datum des Eintritts in die Studie und dem der Beobachtung basiert.

Kaplan-Meier-Analyse. Die Cox-Regression kann ohne Eingabefelder durchgeführt werden. Dies entspricht einer Kaplan-Meier-Analyse.

Feldoptionen für Cox-Knoten

Überlebenszeit. Wählen Sie ein numerisches Feld (mit dem Messniveau *Stetig*), um den Knoten ausführen zu können. Die Überlebenszeit gibt die Lebensdauer des vorherzusagenden Datensatzes an. Wenn zum Beispiel die Zeit bis zur Abwanderung von Kunden modelliert wird, wäre dies das Feld, das festhält, wie lange der Kunde schon beim Unternehmen ist. Das Beitritts- oder Abwanderungsdatum des Kunden hätte keine Auswirkungen auf das Modell. Nur die Beschäftigungsdauer des Kunden wäre relevant.

Die Überlebenszeit wird als Dauer ohne Einheiten angenommen. Sie müssen sicherstellen, dass die Eingabefelder mit der Überlebenszeit übereinstimmen. Beispielsweise würden Sie in einer Studie zur Messung des Kundenverlusts pro Monat als Eingabe den Monatsumsatz statt des Jahresumsatzes verwenden. Wenn Ihre Daten Start- und Endzeiten statt einer Dauer aufweisen, müssen Sie diese Zeiten für eine Dauer neu codieren, die vor dem Cox-Knoten liegt.

Die restlichen Felder in diesem Dialogfeld sind Standardfelder, die überall in IBM SPSS Modeler verwendet werden. Weitere Informationen finden Sie im Thema „Feldoptionen der Modellierungsknoten" auf Seite 31.

Modelloptionen für Cox-Knoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen" auf Seite 28.

Methode. Zur Eingabe von Prädiktoren in das Modell sind folgende Optionen verfügbar:

- **Einschluss.** Dies ist das Standardverfahren, bei dem alle Terme direkt in das Modell aufgenommen werden. Beim Erstellen des Modells wird keine Feldauswahl durchgeführt.
- **Schrittweise.** Bei der Methode "Schrittweise" der Feldauswahl wird, wie der Name andeutet, das Modell in Schritten erstellt. Das anfängliche Modell ist das einfachste Modell, das möglich ist. Es enthält keine Modellterme (außer der Konstanten) im Modell. Bei jedem Schritt werden die Terme, die noch nicht zum Modell hinzugefügt wurden, bewertet, und wenn der beste dieser Terme in signifikanter Weise zur Vorhersagekraft des Modells beiträgt, wird er hinzugefügt. Außerdem werden die derzeit im Modell enthaltenen Terme neu bewertet, um zu ermitteln, ob einige davon ohne signifikante Beeinträchtigung des Modells entfernt werden können. Wenn dies der Fall ist, werden sie entfernt. Der Vorgang wird wiederholt und andere Terme werden hinzugefügt und/oder entfernt. Wenn keine weiteren Terme zur Verbesserung des Modells hinzugefügt werden können und keine Terme mehr entfernt werden können, ohne das Modell zu beeinträchtigen, wird das endgültige Modell generiert.
- **Schrittweise rückwärts.** Die Methode "Schrittweise rückwärts" ist im Grunde das Gegenteil der Methode "Schrittweise". Bei dieser Methode enthält das ursprüngliche Modell alle Terme als Prädiktoren. Bei jedem Schritt werden die Terme im Modell evaluiert und alle Terme, die ohne signifikante Beeinträchtigung des Modells entfernt werden können, werden entfernt. Außerdem werden die zuvor entfernten Terme erneut evaluiert, um zu ermitteln, ob der beste dieser Terme in signifikanter Weise zur Vorhersagekraft des Modells beiträgt. Ist dies der Fall, so wird er wieder in das Modell aufgenommen. Wenn keine Terme mehr entfernt werden können, ohne das Modell wesentlich zu beeinträchtigen, und keine weiteren Terme zur Verbesserung des Modells hinzugefügt werden können, wird das endgültige Modell generiert.

Hinweis: Die automatischen Methoden, einschließlich "Schrittweise" und "Schrittweise rückwärts", sind sehr anpassungsfähige Lernmethoden und weisen eine starke Tendenz zur übermäßigen Anpassung an die Trainingsdaten auf. Bei der Verwendung dieser Methoden ist es ganz besonders wichtig, die Validität des entstehenden Modells zu überprüfen - entweder mit neuen Daten oder mithilfe einer zurückgehaltenen Teststichprobe, die mit dem Partitionsknoten erstellt wurde.

Gruppen. Die Angabe eines Gruppenfelds führt dazu, dass der Knoten separate Modelle für die einzelnen Kategorien des Felds berechnet. Es kann sich dabei um ein beliebiges kategoriales Feld (Flag oder nominal) mit dem Speichertyp "Zeichenfolge" oder "Ganze Zahl" handeln.

Modelltyp. Es gibt zwei Optionen zur Definition der Terme im Modell. **Haupteffektmodelle** beinhalten nur die einzelnen Eingabefelder und testen nicht die Interaktionen (multiplikativen Effekte) zwischen den Eingabefeldern. **Benutzerdefinierte Modelle** enthalten nur die von Ihnen angegebenen Terme (Haupteffekte und Interaktionen). Verwenden Sie bei der Auswahl dieser Option die Liste "Terme im Modell", um Terme zum Modell hinzuzufügen oder daraus zu entfernen.

Terme im Modell. Beim Erstellen eines benutzerdefinierten Modells müssen Sie die Terme im Modell explizit angeben. Die Liste zeigt die aktuelle Menge an Termen für das Modell. Mit den Schaltflächen auf der rechten Seite der Liste "Terme im Modell" können Sie Modellterme hinzufügen bzw. entfernen.

- Um Terme zum Modell hinzuzufügen, klicken Sie auf die Schaltfläche *Neue Terme im Modell hinzufügen*.
- Zum Löschen von Termen wählen Sie die gewünschten Terme aus und klicken Sie auf die Schaltfläche *Ausgewählte Terme im Modell löschen*.

Hinzufügen von Termen zu einem Cox-Regressionsmodell

Beim Anfordern eines benutzerdefinierten Modells können Sie Terme zum Modell hinzufügen, indem Sie auf der Registerkarte für das Modell auf die Schaltfläche *Neue Terme im Modell hinzufügen* klicken. Ein neues Dialogfeld wird geöffnet, in dem Sie Terme angeben können.

Typ des hinzuzufügenden Terms. Es gibt mehrere Methoden zum Hinzufügen von Termen zum Modell, je nach der Auswahl der Eingabefelder in der Liste der verfügbaren Felder.

- **Einzelne Interaktion.** Fügt den Term ein, der für die Interaktion aller ausgewählten Felder steht.
- **Haupteffekte.** Fügt für jedes ausgewählte Eingabefeld einen Haupteffekt-Term (das Feld selbst) ein.
- **Alle zweifachen Interaktionen.** Fügt für jedes mögliche Paar ausgewählter Eingabefelder einen Zweifach-Interaktions-Term (das Produkt der Eingabefelder) ein. Beispiel: Bei Auswahl der Eingabefelder *A*, *B* und *C* in der Liste der verfügbaren Felder werden bei dieser Methode die Terme *A * B*, *A * C* und *B * C* eingefügt.
- **Alle dreifachen Interaktionen.** Fügt für jede mögliche Kombination aus jeweils drei ausgewählten Eingabefeldern einen Dreifach-Interaktions-Term (das Produkt der Eingabefelder) ein. Beispiel: Bei Auswahl der Eingabefelder *A*, *B*, *C* und *D* in der Liste der verfügbaren Felder werden bei dieser Methode die Terme *A * B * C*, *A * B * D*, *A * C * D* und *B * C * D* eingefügt.
- **Alle vierfachen Interaktionen.** Fügt für jede mögliche Kombination aus jeweils vier ausgewählten Eingabefeldern einen Vierfach-Interaktions-Term (das Produkt der Eingabefelder) ein. Beispiel: Bei Auswahl der Eingabefelder *A*, *B*, *C*, *D* und *E* in der Liste der verfügbaren Felder werden bei dieser Methode die Terme *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* und *B * C * D * E* eingefügt.

Verfügbare Felder. Listet die verfügbaren Eingabefelder auf, die bei der Konstruktion der Modellterme verwendet werden sollen. Beachten Sie, dass die Liste möglicherweise Felder enthält, bei denen es sich nicht um zulässige Eingabefelder handelt. Achten Sie daher sorgfältig darauf, dass alle Modellterme nur Eingabefelder enthalten.

Vorschau. Zeigt die Terme an, die beim Klicken auf **Einfügen** zum Modell hinzugefügt werden. Dabei werden die ausgewählten Felder und der oben ausgewählte Termtyp zugrunde gelegt.

Einfügen. Fügt (auf der Grundlage der aktuellen Auswahl von Feldern und des Termtyps) Terme in das Modell ein und schließt das Dialogfeld.

Expertenoptionen für Cox-Knoten

Konvergenz. Mit diesen Optionen können Sie die Parameter für die Modellkonvergenz festlegen. Bei der Ausführung des Modells steuern die Konvergenzkriterien, wie viele Male die verschiedenen Parameter wiederholt durchlaufen werden, um zu ermitteln, wie gut sie passen. Je häufiger die Parameter durchprobiert werden, desto enger liegen die Ergebnisse beieinander (d. h. die Ergebnisse konvergieren). Weitere Informationen finden Sie im Thema „Konvergenzkriterien für Cox-Knoten".

Ausgabe. Mit diesen Optionen können Sie zusätzliche Statistiken und Plots anfordern (einschließlich der Überlebenskurve), die in der erweiterten Ausgabe des vom Knoten erstellten generierten Modells angezeigt werden. Weitere Informationen finden Sie im Thema „Cox-Knoten - Erweiterte Ausgabeoptionen".

Kriterien. Mit diesen Optionen können Sie die Kriterien zum Hinzufügen und Entfernen von Feldern mit der Schätzmethode "Schrittweise" festlegen. (Die Schaltfläche ist inaktiviert, wenn die Methode "Einschluss" ausgewählt ist.) Weitere Informationen finden Sie im Thema „Schrittkriterien für Cox-Knoten" auf Seite 213.

Konvergenzkriterien für Cox-Knoten

Maximale Iterationen. Hiermit können Sie die Maximalzahl der Iterationen für das Modell festlegen, die bestimmt, wie lange die Prozedur nach einer Lösung sucht.

Log-Likelihood-Konvergenz. Iterationen werden angehalten, wenn die relative Änderung der Log-Wahrscheinlichkeit (Log-Likelihood) kleiner als dieser Wert ist. Wenn der Wert gleich 0 ist, wird dieses Kriterium nicht angewendet.

Parameterkonvergenz. Die Iterationen werden angehalten, wenn die absolute oder relative Änderung in den Parameterschätzungen kleiner als dieser Wert ist. Wenn der Wert gleich 0 ist, wird dieses Kriterium nicht angewendet.

Cox-Knoten - Erweiterte Ausgabeoptionen

Statistik. Für die Modellparameter sind Statistiken wie Konfidenzintervalle für Exp(*B*) und Korrelation der Schätzungen verfügbar. Diese Statistiken können für jeden Schritt oder nur für den letzten Schritt angefordert werden.

Grundlinienfunktion anzeigen. Hiermit können Sie die Basishazardfunktion und die kumulative Überlebensfunktion beim Mittelwert der Kovariaten anzeigen.

Diagramme

Diagramme können ein Hilfsmittel zur Bewertung des geschätzten Modells und zur Interpretation der Ergebnisse sein. Die Überlebens-, Hazard- und Log-Minus-Log-Funktionen sowie Eins minus Überleben können grafisch dargestellt werden.

- *Überleben*. Zeigt die kumulative Überlebensfunktion auf einer linearen Skala an.
- *Hazard*. Zeigt die kumulative Hazardfunktion auf einer linearen Skala an.
- **Log minus Log.** Zeigt die kumulative Überlebensschätzung nach Anwendung der ln(-ln)-Transformation auf die Schätzung an.
- *Eins minus Überleben*. Erzeugt ein Diagramm der Werte "1 - Überlebensfunktion" auf einer linearen Skala.

Gesonderte Linie für jeden Wert darstellen. Diese Option ist nur für kategoriale Felder verfügbar.

Für Diagramme zu verwendender Wert. Da diese Funktionen von den Werten der Prädiktoren abhängen, müssen Sie konstante Werte für die Prädiktoren verwenden, um die Funktionen in Abhängigkeit von der

Zeit grafisch darzustellen. In der Standardvorgabe wird der Mittelwert der einzelnen Prädiktoren als konstanter Wert verwendet. Sie können jedoch mithilfe des Rasters Ihre eigenen Werte für den Plot eingeben. Bei kategorialen Eingaben wird Indikatorcodierung verwendet, sodass ein Regressionskoeffizient für jede Kategorie (mit Ausnahme der letzten) vorhanden ist. Kategoriale Eingaben weisen also einen Mittelwert für jeden Indikatorkontrast auf, der gleich dem Anteil an Fällen in der Kategorie ist, die zum Indikatorkontrast gehört.

Schrittkriterien für Cox-Knoten

Kriterium für Entfernen. Wählen Sie **Likelihood-Quotient** für ein robusteres Modell. Zur Verkürzung der für die Modellerstellung erforderlichen Zeit können Sie **Wald** auswählen. Es gibt die zusätzliche Option **Bedingt**, die Ausschlusstests auf der Grundlage der Wahrscheinlichkeit der Likelihood-Quotienten-Statistik ermöglicht, welche auf bedingten Parameterschätzungen beruht.

Signifikanzschwellen für LR-Kriterien. Mit dieser Option können Sie Auswahlkriterien auf der Grundlage der statistischen Wahrscheinlichkeit (*p*-Wert) angeben, die den einzelnen Feldern zugeordnet ist. Felder werden nur zum Modell hinzugefügt, wenn der zugehörige *p*-Wert kleiner ist als der Wert für **Aufnahme**, und nur dann entfernt, wenn der *p*-Wert größer ist als der Wert für **Ausschluss**. Der Wert für **Aufnahme** muss unter dem Wert für **Ausschluss** liegen.

Einstellungsoptionen für Cox-Knoten

Überleben zu zukünftigen Zeitpunkten voraussagen. Wählen Sie einen oder mehrere zukünftige Zeitpunkte aus. Das Überleben, also ob der jeweilige Fall mindestens die angegebene Zeitdauer (vom aktuellen Zeitpunkt gerechnet) überlebt, ohne dass das terminale Ereignis eintritt, wird für jeden Datensatz bei jedem Zeitwert vorhergesagt. Dabei wird jeweils eine Vorhersage pro Zeitwert erstellt. Beachten Sie, dass das Überleben der Wert "falsch" des Zielfelds ist.

- **Regelmäßige Intervalle.** Werte für die Überlebenszeit werden aus den angegebenen Werten für **Zeitintervall** und **Anzahl der zu scorenden Zeitperioden** erstellt. Wenn beispielsweise drei Zeitperioden mit dem Intervall 2 zwischen den einzelnen Zeitpunkten angefordert werden, wird das Überleben für die zukünftigen Zeitpunkte 2, 4, 6 vorhergesagt. Jeder Datensatz wird bei denselben Zeitwerten evaluiert.
- **Zeitfelder.** Für jeden Datensatz im ausgewählten Zeitfeld werden Überlebenszeiten angegeben (es wird genau ein Vorhersagefeld generiert), sodass die einzelnen Datensätze zu verschiedenen Zeitpunkten evaluiert werden können.

Vergangene Überlebenszeit. Gibt die bisherige Überlebenszeit des Datensatzes an. Beispielsweise wird die Beschäftigungsdauer eines vorhandenen Kunden als Feld angegeben. Das Scoring der Überlebenswahrscheinlichkeit zu einem zukünftigen Zeitpunkt hängt von der vergangenen Überlebenswahrscheinlichkeit ab.

Hinweis: Die Werte für zukünftige und vergangene Überlebenszeiten müssen innerhalb des Bereichs für Überlebenszeiten in den zum Trainieren des Modells verwendeten Daten liegen. Datensätze, bei denen die Zeiten außerhalb dieses Bereichs liegen, werden als "null" gescort.

Alle Wahrscheinlichkeiten anhängen. Gibt an, ob die Wahrscheinlichkeiten für die einzelnen Kategorien des Ausgabefelds zu den einzelnen vom Knoten verarbeiteten Datensätzen hinzugefügt werden. Wenn diese Option nicht ausgewählt wurde, wird nur die Wahrscheinlichkeit der vorhergesagten Kategorie hinzugefügt. Die Wahrscheinlichkeiten werden für jeden zukünftigen Zeitpunkt berechnet.

Kumulative Hazardfunktion berechnen. Gibt an, ob die Werte der kumulativen Hazardrate in jeden Datensatz aufgenommen werden sollen. Die kumulative Hazardrate wird für jeden zukünftigen Zeitpunkt berechnet.

Cox-Modellnugget

Cox-Regressionsmodelle stehen für die Gleichungen, die durch die Cox-Knoten geschätzt wurden. Sie enthalten alle Informationen, die vom Modell erfasst wurden, sowie Informationen über die Struktur und Leistung des Modells.

Wenn Sie einen Stream ausführen, der ein generiertes Cox-Regressionsmodell enthält, fügt der Knoten zwei neue Felder hinzu, die die Vorhersage des Modells und die zugehörige Wahrscheinlichkeit enthalten. Die Namen der neuen Felder werden aus dem Namen des prognostizierten Ausgabefelds abgeleitet, dem *$C-* für die vorhergesagte Kategorie und *$CP-* für die zugehörige Wahrscheinlichkeit vorangestellt ist. Die Nummer des zukünftigen Zeitintervalls bzw. der Name des Zeitfelds, das das Zeitintervall definiert, ist als Suffix angegeben. Beispiel: Für das Ausgabefeld *churn* und zwei zukünftige Zeitintervalle, die in regelmäßigen Intervallen definiert sind, erhalten die neuen Felder die Namen *$C-churn-1*, *$CP-churn-1*, *$C-churn-2* und *$CP-churn-2*. Wenn zukünftige Zeitpunkte mit dem Zeitfeld *tenure* definiert sind, erhalten die neuen Felder die Namen *$C-churn_tenure* und *$CP-churn_tenure*.

Wenn Sie im Cox-Knoten die Einstellungsoption **Alle Wahrscheinlichkeiten anhängen** ausgewählt haben, werden für jeden zukünftigen Zeitpunkt zwei zusätzliche Felder hinzugefügt, die für jeden Datensatz die Wahrscheinlichkeiten für Überleben und Versagen enthalten. Diese zusätzlichen Felder werden anhand des Namens des Ausgabefelds benannt. Diesem wird *$CP-<Falsch-Wert>-* für die Überlebenswahrscheinlichkeit und *$CP-<Wahr-Wert>-* für die Wahrscheinlichkeit, dass das Ereignis eingetreten ist, vorangestellt und die Nummer des zukünftigen Zeitintervalls nachgestellt. Beispiel: Bei einem Ausgabefeld, bei dem der Wert "falsch" 0 und der Wert "wahr" 1 ist und zwei zukünftige Zeitintervalle in regelmäßigen Intervallen definiert sind, erhalten die neuen Felder die Namen *$CP-0-1*, *$CP-1-1*, *$CP-0-2* und *$CP-1-2*. Wenn zukünftige Zeiten mit einem einzigen Zeitfeld *tenure* definiert sind, erhalten die neuen Felder die Namen *$CP-0-1* und *$CP-1-1*, da nur ein einziges zukünftiges Intervall vorhanden ist.

Wenn Sie im Cox-Knoten die Einstellungsoption **Kumulative Hazardfunktion berechnen** ausgewählt haben, wird für jeden zukünftigen Zeitpunkt ein zusätzliches Feld hinzugefügt, das für jeden Datensatz die kumulative Hazard-Funktion enthält. Diese zusätzlichen Felder werden anhand des Namens des Ausgabefelds benannt. Diesem wird *$CH-* vorangestellt und die Nummer des zukünftigen Zeitintervalls bzw. der Name, der das Zeitintervall definiert, nachgestellt. Beispiel: Für das Ausgabefeld *churn* und zwei zukünftige Zeitintervalle, die in regelmäßigen Intervallen definiert sind, erhalten die neuen Felder die Namen *$CH-churn-1* und *$CH-churn-2*. Wenn zukünftige Zeitpunkte mit dem Zeitfeld *tenure* definiert sind, erhält das neue Feld den Namen *$CH-churn-1*.

Cox-Regression - Ausgabeeinstellungen

Die Registerkarte "Einstellungen" des Nuggets enthält dieselben Steuerelemente wie die Registerkarte "Einstellungen" des Modellknotens. Die Standardwerte der Nuggetsteuerelemente richten sich nach den im Modellknoten festgelegten Werten. Weitere Informationen finden Sie im Thema „Einstellungsoptionen für Cox-Knoten” auf Seite 213.

Cox-Regression - Erweiterte Ausgabe

Die erweiterte Ausgabe für die Cox-Regression bietet detaillierte Informationen zum geschätzten Modell und dessen Leistung, einschließlich der Überlebenskurve. Die meisten der in der erweiterten Ausgabe enthaltenen Informationen weisen einen hohen Fachlichkeitsgrad auf und zur richtigen Interpretation dieser Ausgaben sind umfassende Kenntnisse der Cox-Regression erforderlich.

Kapitel 11. Clustermodelle

Clustermodelle konzentrieren sich auf die Ermittlung ähnlicher Datensätze und Beschriftung der Datensätze anhand der Gruppe, in die sie gehören. Dies geschieht ohne Vorkenntnisse zu Gruppen und ihren Eigenschaften. Vielleicht wissen Sie nicht einmal, nach wie vielen Gruppen Sie suchen sollen. Hierin unterscheiden Clustering-Modelle sich von den anderen Techniken des Maschinenlernens: Es gibt keine vordefinierte Ausgabe und kein vordefiniertes Zielfeld für das vorherzusagende Modell. Diese Modelle werden häufig als **nicht überwachte Lernmodelle** bezeichnet, da es keinen externen Standard gibt, mit dem die Klassifizierungsleistung des Modells beurteilt werden könnte. Für diese Modelle gibt es keine *richtigen* oder *falschen* Antworten. Ihr Wert wird durch die Möglichkeit bestimmt, interessante Gruppierungen in den Daten zu erfassen und sinnvolle Beschreibungen dieser Gruppierungen zu liefern.

Methoden zur Clusterbildung basieren auf dem Messen der Entfernungen zwischen Datensätzen und Clustern. Die Datensätze werden den Clustern auf eine Weise zugewiesen, die die Entfernung zwischen den Datensätzen minimiert, die demselben Cluster angehören.

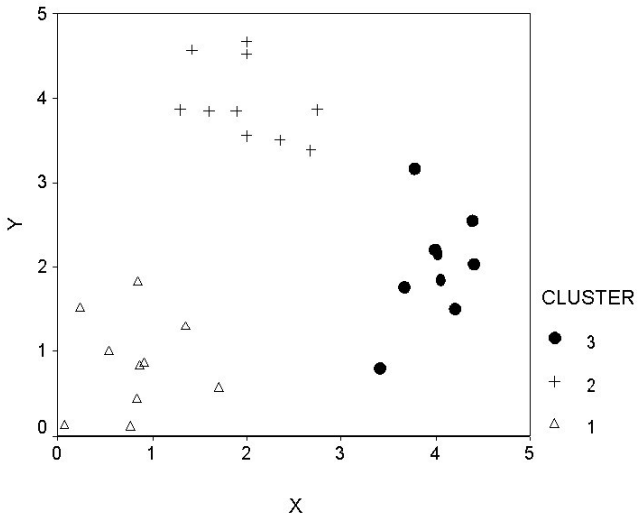


Abbildung 44. Einfaches Clustering-Modell

Drei Clustering-Methoden werden bereitgestellt:

Der K-Means-Knoten teilt das Dataset in unterschiedliche Gruppen (oder Cluster) auf. Bei dieser Methode wird eine festgelegte Anzahl von Clustern definiert, den Clustern werden iterativ Datensätze zugewiesen und die Clusterzentren werden angepasst, bis eine weitere Verfeinerung keine wesentliche Verbesserung des Modells mehr darstellen würde. Statt zu versuchen, ein Ergebnis vorherzusagen, versucht *K*-Means mithilfe eines als "nicht überwachtes Lernen" bezeichneten Prozesses Muster im Set der Eingabefelder zu entdecken.

Der TwoStep-Knoten verwendet eine aus zwei Schritten bestehende Clusterbildungsmethode. Im ersten Schritt wird ein einzelner Durchlauf durch die Daten vorgenommen, bei dem die Eingangsrohdaten zu einem verwaltbaren Set von Subclustern komprimiert werden. Im zweiten Schritt werden die Subcluster mithilfe einer hierarchischen Methode zur Clusterbildung nach und nach in immer größere Cluster zusammengeführt. TwoStep hat den Vorteil, dass die optimale Anzahl an Clustern für die Trainingsdaten automatisch geschätzt wird. Mit dem Verfahren können gemischte Feldtypen und große Datasets effizient verarbeitet werden.

Der Kohonen-Knoten erstellt eine Art von neuronalem Netz, das verwendet werden kann, um ein Clustering des Datasets in einzelne Gruppen vorzunehmen. Wenn das Netz voll trainiert ist, sollten ähnliche Datensätze auf der Ausgabekarte eng nebeneinander stehen, während Datensätze, die sich unterscheiden, weit voneinander entfernt sein sollten. Die Zahl der von jeder Einheit im Modellnugget erfassten Beobachtungen gibt Aufschluss über die starken Einheiten. Dadurch wird ein Eindruck von der ungefähren Zahl der Cluster vermittelt.

Clustermodelle werden häufig verwendet, um Cluster oder Segmente zu erstellen, die dann als Eingaben in nachfolgenden Analysen verwendet werden. Ein häufiges Beispiel dafür sind die von Marktforschern verwendeten Marktsegmente, mit denen der Gesamtmarkt in homogene Untergruppen aufgeteilt wird. Jedes Segment weist besondere Eigenschaften auf, die sich auf den Erfolg der Marktforschung auswirken. Wenn Sie Data Mining zur Optimierung Ihrer Marketingstrategie verwenden, können Sie Ihr Modell in der Regel erheblich verbessern, indem Sie die entsprechenden Segmente ermitteln und diese Segmentinformationen für Ihre Vorhersagemodelle verwenden.

Kohonen-Knoten

Kohonen-Netze stellen eine Form von neuronalen Netzen zur Clusterbildung dar. Sie sind auch bekannt unter dem Bezeichnung **K-Netz (knet)** oder **SOM (selbstorganisierende Karte)**. Mit dieser Methode können Sie ein Clustering des Datasets in einzelne Gruppen vornehmen, wenn Sie nicht wissen, wie diese Gruppen am Anfang aussehen. Datensätze werden in Gruppen zusammengefasst, wobei Datensätze innerhalb einer Gruppe oder eines Clusters ähnlich und Datensätze in verschiedenen Gruppen unterschiedlich sind.

Die Basiseinheiten sind **Neuronen** und sie sind in zwei Schichten organisiert: die **Eingabeschicht** und die **Ausgabeschicht** (auch **Ausgabezuordnung** genannt). Alle Eingabeneuronen sind mit allen Ausgabeneuronen verbunden. Mit diesen Verbindungen sind **Stärken** oder **Gewichtungen** verknüpft. Während des Trainings wetteifert jede Einheit mit allen anderen, um einen Datensatz zu "gewinnen".

Die Ausgabekarte ist ein zweidimensionales Neuronenraster ohne Verbindungen zwischen den Einheiten.

Die Eingabedaten werden der Eingabeschicht präsentiet und die Werte an die Ausgabeschicht weitergeleitet. Das Ausgabeneuron mit der stärksten Reaktion soll der **Gewinner** sein und ist die Antwort für diese Eingabe.

Anfänglich sind alle Gewichtungen zufällig. Wenn eine Einheit einen Datensatz gewinnt, werden die Gewichtungen (zusammen mit denen anderer Nachbareinheiten, die kollektiv als **Nachbarschaft** bezeichnet werden) so angepasst, dass sie dem Muster der Prädiktorwerte für diesen Datensatz besser entsprechen. Alle Eingabedatensätze werden angezeigt und die Gewichtungen entsprechend aktualisiert. Dieser Vorgang wird viele Male wiederholt, bis die Änderungen nur noch gering sind. Während des Trainings werden die Gewichtungen an den Rastereinheiten so angepasst, dass sie eine zweidimensionale "Karte" der Cluster bilden (deshalb die Bezeichnung **selbstorganisierende Karte**).

Wenn das Netz voll trainiert ist, sollten ähnliche Datensätze auf der Ausgabekarte eng nebeneinander stehen, während Datensätze, die sich stark unterscheiden, weit voneinander entfernt sein sollten.

Im Gegensatz zu den meisten Lernmethoden in IBM SPSS Modeler verwenden Kohonen-Netze *kein* Zielfeld. Diese Art des Lernens, d. h. ohne Zielfeld, wird als **nicht überwachtes Lernen** bezeichnet. Statt zu versuchen, ein Ergebnis vorherzusagen, versuchen Kohonen-Netze, Muster im Set der Eingabefelder zu entdecken. In der Regel weist ein Kohonen-Netz schließlich einige Einheiten auf, die viele Beobachtungen zusammenfassen (**starke** Einheiten), und mehrere Einheiten, die keiner Beobachtung wirklich entsprechen (**schwache** Einheiten). Die starken Einheiten (und mitunter benachbarte Einheiten im Raster) repräsentieren mögliche Clusterzentren.

Eine weitere Einsatzmöglichkeit von Kohonen-Netzen findet sich bei der **Dimensionsreduzierung**. Das räumliche Merkmal des zweidimensionalen Rasters bietet eine Zuordnung der ursprünglichen *k*-Prädiktoren zu zwei abgeleiteten Funktionen, die die Ähnlichkeitsbeziehung in den ursprünglichen Prädiktoren bewahren. In einigen Fällen kann dies ebenso vorteilhaft sein wie die Faktoranalyse oder PCA.

Beachten Sie, dass die Methode zur Berechnung der Standardgröße des Ausgaberasters sich im Vergleich zu früheren Versionen von IBM SPSS Modeler geändert hat. Mit der neuen Methode werden im Allgemeinen kleinere Ausgabeschichten erzielt, die schneller zu trainieren sind und besser verallgemeinern. Wenn Sie mit der Standardgröße schlechte Ergebnisse erzielen, erhöhen Sie den Wert für die Größe des Ausgaberasters auf der Registerkarte "Experten". Weitere Informationen finden Sie im Thema „Expertenoptionen für den Kohonen-Knoten” auf Seite 218.

Anforderungen. Zum Trainieren eines Kohonen-Netzes ist mindestens ein Feld mit der Rolle *Eingabe* erforderlich. Felder, deren Rolle auf *Ziel*, *Beides* oder *Keine* festgelegt ist, werden ignoriert.

Stärken. Um ein Kohonen-Netzmodell zu erstellen, sind keine Daten über die Gruppenzugehörigkeit erforderlich. Auch die Anzahl Gruppen muss für die Suche nicht bekannt sein. Kohonen-Netze beginnen mit einer großen Anzahl von Einheiten, und mit Fortschreiten des Trainings gravitieren die Einheiten zu natürlichen Clustern in den Daten. Die Zahl der von jeder Einheit erfassten Beobachtungen im Modellnugget gibt Aufschluss über die starken Einheiten, die einen Eindruck von der ungefähren Zahl der Cluster vermitteln.

Optionen des Kohonen-Knotenmodells

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Training des bestehenden Modells fortsetzen. Standardmäßig wird bei jeder Ausführung eines Kohonen-Knotens ein komplett neues Netz aufgebaut. Bei Auswahl dieser Option wird das Training mit dem letzten, vom Knoten erfolgreich aufgebauten Netz fortgesetzt.

Feedbackdiagramm anzeigen. Bei Auswahl dieser Option wird während des Trainings eine visuelle Darstellung des zweidimensionalen Arrays angezeigt. Die Stärke der einzelnen Knoten wird farblich dargestellt. Rot kennzeichnet eine Einheit mit vielen Datensätzen (eine **starke** Einheit), während Weiß auf eine Einheit hinweist, die wenige oder keine Datensätze enthält (eine **schwache** Einheit). Feedback wird möglicherweise nicht angezeigt, wenn die Zeit für die Modellerstellung relativ kurz ist. Diese Funktion kann die Trainingszeit verlangsamen. Inaktivieren Sie diese Option, wenn Sie die Trainingszeit beschleunigen möchten.

Stopp bei. Das Standardstoppkriterium stoppt das Training basierend auf internen Parametern. Sie können auch eine Zeit als Stoppkriterium festlegen. Geben Sie die Zeit (in Minuten) für das Training des Netzes ein.

Startwert für Zufallsgenerator festlegen. Wenn kein Startwert für den Zufallsgenerator festgelegt wurde, ist die Sequenz der Zufallswerte, mit denen die Netzgewichtungen initialisiert werden, bei jeder Ausführung des Knotens unterschiedlich. Dadurch ist es möglich, dass der Knoten bei verschiedenen Ausführungen unterschiedliche Modelle erstellt, auch wenn die Knoteneinstellungen und Datenwerte vollkommen identisch sind. Wenn Sie diese Option auswählen, können Sie den Startwert für den Zufallsgenerator auf einen bestimmten Wert festlegen, sodass das entstehende Modell genau reproduziert werden kann. Ein bestimmter Startwert für den Zufallsgenerator erzeugt immer dieselbe Sequenz der zufälligen Werte. In diesem Fall führt die Ausführung des Knotens immer zu demselben generierten Modell.

Hinweis: Bei Verwendung der Option **Startwert für Zufallsgenerator festlegen** mit Datensätzen, die aus einer Datenbank eingelesen wurden, ist möglicherweise vor der Stichprobenziehung ein Sortierknoten erforderlich, um zu gewährleisten, dass bei jeder Ausführung des Knotens dasselbe Ergebnis erzielt wird. Dies liegt daran, dass der Startwert für den Zufallsgenerator von der Reihenfolge der Datensätze abhängt, die in relationalen Datenbanken nicht unbedingt gleich bleibt.

Hinweis: Wenn Sie nominale Felder (Setfelder) in Ihr Modell integrieren möchten, jedoch Speicherprobleme bei der Modellerstellung haben oder die Modellerstellung zu viel Zeit in Anspruch nimmt, codieren Sie große Setfelder um, um die Anzahl der Werte zu verringern, oder verwenden Sie ein anderes Feld mit weniger Werten als Proxy für das große Set. Wenn Sie beispielsweise Probleme mit einem *product_id*-Feld haben, das Werte für einzelne Produkte enthält, können Sie es aus dem Modell entfernen und stattdessen ein weniger detailliertes *product_category*-Feld hinzufügen.

Optimieren. Wählen Sie Optionen aus, die die Leistung während der Modellerstellung basierend auf Ihren persönlichen Anforderungen erhöhen.

- Wählen Sie **Geschwindigkeit** aus, um den Algorithmus anzuweisen, zur Verbesserung der Leistung keinen Datenträgerüberlauf zu verwenden.
- Wählen Sie **Speicher** aus, um den Algorithmus anzuweisen, gegebenenfalls einen Datenträgerüberlauf zu verwenden und dafür Geschwindigkeitseinbußen hinzunehmen. Diese Option ist standardmäßig aktiviert.

  Hinweis: Bei der Ausführung im verteilten Modus kann diese Einstellung durch die in der Datei *options.cfg* angegebenen Administratoroptionen überschrieben werden.

Clusterbeschriftung anhängen. Diese standardmäßig für neue Modelle ausgewählte Option, die aber für Modelle aus früheren IBM SPSS Modeler-Versionen inaktiviert ist, erstellt ein einzelnes kategoriales Scorefeld desselben Typs, der sowohl vom Knoten "K-Means" als auch vom Knoten "TwoStep" erstellt wird. Dieses Zeichenfolgenfeld wird im Knoten "Autom. Cluster" verwendet, wenn Rangmessungen für die unterschiedlichen Modelltypen berechnet werden. Weitere Informationen finden Sie im Thema „Knoten "Autom. Cluster"" auf Seite 75.

Expertenoptionen für den Kohonen-Knoten

Für Anwender mit detaillierten Kenntnissen über Kohonen-Netze stehen Expertenoptionen zur Verfügung, mit denen der Trainingsprozess verfeinert wird. Für den Zugriff auf die Expertenoptionen stellen Sie den Modus auf der Registerkarte "Experten" auf **Experten** ein.

Breite und Länge. Geben Sie die Größe (Breite und Länge) der zweidimensionalen Ausgabekarte als Anzahl der Ausgabeeinheiten für jede Dimension an.

Lernratenverfall. Wählen Sie entweder den linearen oder exponentiellen Lernratenverfall. Die **Lernrate** ist ein Gewichtungsfaktor, der im Laufe der Zeit abnimmt, sodass das Netz mit der Verschlüsselung weiträumiger Funktionen der Daten beginnt und sich allmählich auf feinere Details konzentriert.

Phase 1 und Phase 2. Das Training des Kohonen-Netzes ist in zwei Phasen aufgeteilt. Phase 1 ist eine grobe Schätzphase, in der die groben Muster der Daten erfasst werden. Phase 2 ist eine Abstimmungsphase, in der die Zuordnung angepasst wird, um die feineren Merkmale der Daten zu modellieren. Für jede Phase gibt es drei Parameter:

- **Nachbarschaft.** Legt die Startgröße (Radius) der Nachbarschaft fest. Dieser Wert bestimmt die Zahl der "benachbarten" Einheiten, die beim Training zusammen mit der gewonnenen Einheit aktualisiert werden. In Phase 1 beginnt die Größe der Nachbarschaft mit *Phase 1 Nachbarschaft* und sinkt auf (*Phase 2 Nachbarschaft* + 1) ab. In Phase 2 liegt der Startwert für die Nachbarschaft bei *Phase 2 Nachbarschaft* und nimmt bis auf 1.0 ab. *Phase 1 Nachbarschaft* sollte größer sein als *Phase 2 Nachbarschaft*.

- **Anfängliches Eta.** Legt den Startwert für die Lernrate **Eta** fest. In Phase 1 beginnt der Eta-Wert bei *Phase 1 Anfängliches Eta* und sinkt auf *Phase 2 Anfängliches Eta* ab. In Phase 2 liegt der Eta-Anfangswert bei *Phase 2 Anfängliches Eta* und nimmt bis auf 0 ab. *Phase 1 Anfängliches Eta* sollte größer sein als *Phase 2 Anfängliches Eta*.
- **Zyklen.** Legt die Anzahl der Zyklen für jede Trainingsphase fest. Jede Phase hält für die Dauer der angegebenen Zahl von Durchläufen durch die Daten an.

Modellnuggets vom Typ "Kohonen"

Modellnuggets für Kohonen-Modelle enthalten alle Informationen, die vom trainierten Kohonen-Netz erfasst wurden, sowie Informationen zur Architektur des Netzes.

Wenn Sie einen Stream mit einem Modellnuggets vom Typ "Kohonen" auswählen, fügt der Knoten zwei neue Felder mit den *X*- und *Y*-Koordinaten der Einheit im Kohonen-Ausgaberaster hinzu, die am stärksten auf diesen Datensatz reagiert hat. Die neuen Feldnamen aus dem Modellnamen abgeleitet, wobei ihnen die Präfixe *$KX-* und *$KY-* vorangestellt wird. Beispiel: Wenn das Modell den Namen *Kohonen* trägt, erhalten die neuen Felder die Namen*$KX-Kohonen* und *$KY-Kohonen*.

Für ein besseres Verständnis davon, was im Kohonen-Netz codiert wurde, klicken Sie auf die Registerkarte "Modell" im Modellnugget-Browser. Dadurch wird der Cluster-Viewer angezeigt, der eine grafische Darstellung der Cluster, Felder und Wichtigkeitsgrade bietet. Weitere Informationen finden Sie im Thema „Cluster-Viewer - Registerkarte "Modell"" auf Seite 224.

Wenn Sie es vorziehen, die Cluster als Gitter zu visualisieren, können Sie die Ergebnisse des Kohonen-Netzes anzeigen, indem Sie mithilfe eines Plotknotens ein Diagramm der *$KX-* und *$KY-*-Felder erstellen. (Sie sollten **X-Bewegung** und **Y-Bewegung** im Plotknoten auswählen, um zu verhindern, dass die Datensätze der einzelnen Einheiten übereinander dargestellt werden.) Im Diagramm können Sie außerdem ein symbolisches Feld überlagern, um festzustellen, wie das Kohonen-Netz die Daten in Cluster eingeteilt hat.

Ein anderes leistungsfähiges Verfahren, um Einblicke über das Kohonen-Netz zu gewinnen, besteht darin, die Merkmale zu erkennen, die die vom Netz gefundenen Cluster unterscheiden. Weitere Informationen finden Sie im Thema „C5.0-Knoten" auf Seite 106.

Allgemeine Informationen zum Verwenden des Modellbrowsers finden Sie unter „Durchsuchen von Modellnuggets" auf Seite 42

Übersicht über das Kohonen-Modell

Auf der Registerkarte "Übersicht" für ein Kohonen-Modellnugget werden Informationen über die Architektur bzw. Topologie des Netzes angezeigt. Die Länge und Breite der zweidimensionalen Kohonen-Funktionskarte (Ausgabeschicht) werden als **$KX-** *Modellname* und **$KY-** *Modellname* angezeigt. Für Ein- und Ausgabeschicht wird jeweils die Anzahl der Einheiten in der Schicht aufgeführt.

K-Means-Knoten

Der K-Means-Knoten bietet eine Methode der **Clusteranalyse**. Mit dieser Methode können Sie ein Clustering der Datasets in einzelne Gruppen vornehmen, wenn Sie nicht wissen, wie diese Gruppen am Anfang aussehen. Im Gegensatz zu den meisten Lernmethoden in IBM SPSS Modeler verwenden K-Means-Modelle *kein* Zielfeld. Diese Art des Lernens, d. h. ohne Zielfeld, wird als **nicht überwachtes Lernen** bezeichnet. Statt zu versuchen, ein Ergebnis vorherzusagen, versuchen K-Means-Knoten, Muster im Set der Eingabefelder zu entdecken. Datensätze werden in Gruppen zusammengefasst, wobei Datensätze innerhalb einer Gruppe oder eines Clusters ähnlich und Datensätze in verschiedenen Gruppen unterschiedlich sind.

K-Means definiert einen Set von Clusterstartzentren, die von Daten abgeleitet werden. Anschließend werden die einzelnen Datensätze basierend auf ihren Eingabefeldwerten dem Cluster zugewiesen, dem sie am meisten ähneln. Nachdem alle Datensätze zugewiesen wurden, werden die Clusterzentren aktualisiert, um die neuen Datensatzsets, die den einzelnen Clustern zugewiesen wurden, wiederzugeben. Die Datensätze werden nun erneut überprüft, um festzustellen, ob sie einem anderen Cluster zugewiesen werden sollten. Der Prozess der Datensatzzuweisung bzw. Clusteriteration wird so lange fortgesetzt, bis die maximale Anzahl an Iterationen erreicht ist oder die Änderung von einer Iteration auf die nächste einen bestimmten Schwellenwert nicht überschreitet.

Hinweis: Das entstehende Modell hängt bis zu einem gewissen Grad von der Reihenfolge der Trainingsdaten ab. Eine Änderung der Datenreihenfolge und ein erneutes Erstellen des Modells kann zu einem anderen endgültigen Clustermodell führen.

Anforderungen. Zum Trainieren eines K-Means-Modells ist mindestens ein Feld mit der Rolle *Eingabe* erforderlich. Felder, deren Rolle auf *Ausgabe*, *Beides* oder *Keine* festgelegt ist, werden ignoriert.

Stärken. Um ein K-Means-Modell zu erstellen, sind keine Daten über die Gruppenzugehörigkeit erforderlich. Das K-Means-Modell stellt häufig die schnellste Methode zur Clusterbildung von großen Datasets dar.

Optionen für K-Means-Knotenmodelle

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Angegebene Anzahl der Cluster. Geben Sie die Anzahl der zu generierenden Cluster an. Der Standardwert ist 5.

Distanzfeld generieren. Wenn diese Option ausgewählt ist, enthält das Modellnugget ein Feld mit der Distanz jedes einzelnen Datensatzes vom Zentrum des entsprechend zugewiesenen Clusters.

Clusterbeschriftung. Geben Sie das Format für die Werte im generierten Feld "Clusterzugehörigkeit" an. Die Clusterzugehörigkeit kann als **Zeichenfolge** mit dem festgelegten **Beschriftungspräfix** (z. B. `"Cluster 1"`, `"Cluster 2"` usw.) oder als **Zahl** angegeben werden.

Hinweis: Wenn Sie nominale Felder (Setfelder) in Ihr Modell integrieren möchten, jedoch Speicherprobleme beim Erstellen des Modells haben oder die Modellerstellung zu viel Zeit in Anspruch nimmt, codieren Sie große Setfelder um, um die Anzahl der Werte zu reduzieren, oder verwenden Sie ein anderes Feld mit weniger Werten als Proxy für das große Set. Wenn Sie beispielsweise Probleme mit einem *product_id*-Feld haben, das Werte für einzelne Produkte enthält, können Sie es aus dem Modell entfernen und stattdessen ein weniger detailliertes *product_category*-Feld hinzufügen.

Optimieren. Wählen Sie Optionen aus, die die Leistung während der Modellerstellung basierend auf Ihren persönlichen Anforderungen erhöhen.

- Wählen Sie **Geschwindigkeit** aus, um den Algorithmus anzuweisen, zur Verbesserung der Leistung keinen Datenträgerüberlauf zu verwenden.
- Wählen Sie **Speicher** aus, um den Algorithmus anzuweisen, gegebenenfalls einen Datenträgerüberlauf zu verwenden und dafür Geschwindigkeitseinbußen hinzunehmen. Diese Option ist standardmäßig aktiviert.

  Hinweis: Bei der Ausführung im verteilten Modus kann diese Einstellung durch die in der Datei *options.cfg* angegebenen Administratoroptionen überschrieben werden.

Expertenoptionen für K-Means-Knoten

Für Anwender mit detaillierten Kenntnissen über *K*-Means-Clusterbildung stehen Expertenoptionen zur Verfügung, mit denen der Trainingsprozess verfeinert wird. Für den Zugriff auf die Expertenoptionen stellen Sie den Modus auf der Registerkarte "Experten" auf **Experten** ein.

Stopp bei. Geben Sie das Stoppkriterium für das Training des Modells an. Das Stoppkriterium **Standard** beträgt 20 Iterationen oder eine Änderung < 0,000001, je nachdem, was zuerst eintritt. Wenn Sie eigene Stoppkriterien angeben möchten, wählen Sie **Benutzerdefiniert** aus.

- **Maximale Iterationen.** Mit dieser Option können Sie das Modelltraining nach der angegebenen Anzahl von Iterationen beenden.
- **Toleranz ändern.** Mit dieser Option können Sie das Modelltraining beenden, sobald die größte Änderung der Clusterzentren für eine Iteration kleiner ist als das angegebene Niveau.

Verschlüsselungswert für Sets. Geben Sie einen Wert zwischen 0 und 1,0 für die Umcodierung von Setfeldern als Gruppe von numerischen Feldern an. Der Standardwert ist die Wurzel aus 0,5 (rund 0,707107). Dieser Wert bietet die richtige Gewichtung für umcodierte Flagfelder. Werte, die näher an 1,0 liegen, gewichten Setfelder stärker als numerische Felder.

Modellnuggets vom Typ "K-Means"

Nuggets für K-Means-Modelle enthalten alle Informationen, die vom Clustermodell erfasst wurden, sowie Informationen zu den Trainingsdaten und dem Schätzvorgang.

Wenn Sie einen Stream ausführen, der einen Modellierungsknoten vom Typ "K-Means" enthält, fügt dieser Knoten zwei neue Felder hinzu, die die Clusterzugehörigkeit und die Entfernung vom zugewiesenen Clusterzentrum für den betreffenden Datensatz enthalten. Die neuen Feldnamen werden durch Präfigierung von *$KM-* für die Clusterzugehörigkeit und *$KMD-* für die Entfernung vom Clusterzentrum aus dem Modellnamen abgeleitet. Beispiel: Wenn das Modell den Namen *Kmeans* trägt, erhalten die neuen Felder die Namen *$KM-Kmeans* und *$KMD-Kmeans*.

Ein leistungsfähiges Verfahren, um Einblicke in das K-Means-Modell zu gewinnen, besteht darin, mithilfe der Regelinduktion die Merkmale zu erkennen, die die vom Modell gefundenen Cluster unterscheiden. Weitere Informationen finden Sie im Thema „C5.0-Knoten” auf Seite 106. Dadurch wird der Cluster-Viewer angezeigt, der eine grafische Darstellung der Cluster, Felder und Wichtigkeitsgrade bietet. Weitere Informationen finden Sie im Thema „Cluster-Viewer - Registerkarte "Modell"” auf Seite 224.

Allgemeine Informationen zum Verwenden des Modellbrowsers finden Sie unter „Durchsuchen von Modellnuggets” auf Seite 42

Übersicht über das K-Means-Modell

Auf der Registerkarte "Übersicht" eines Nuggets für K-Means-Modelle finden Sie Informationen zu den Trainingsdaten, dem Schätzvorgang und den durch das Modell definierten Clustern. Die Anzahl der Cluster sowie der Iterationsverlauf werden angezeigt. Wenn Sie einen Analyseknoten ausgeführt haben, der an diesen Modellierungsknoten angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt.

TwoStep-Clusterknoten

Der TwoStep-Clusterknoten bietet eine Form der **Clusteranalyse**. Mit dieser Methode können Sie ein Clustering der Datasets in einzelne Gruppen vornehmen, wenn Sie nicht wissen, wie diese Gruppen am Anfang aussehen. Ebenso wie Kohonen-Knoten und K-Means-Knoten verwenden auch TwoStep-Clustermodelle *kein* Zielfeld. Statt zu versuchen, ein Ergebnis vorherzusagen, versuchen TwoStep-Cluster, Muster

im Set der Eingabefelder zu entdecken. Datensätze werden in Gruppen zusammengefasst, wobei Datensätze innerhalb einer Gruppe oder eines Clusters ähnlich und Datensätze in verschiedenen Gruppen unterschiedlich sind.

Beim TwoStep-Cluster handelt es sich um eine Clusterbildungsmethode in zwei Schritten. Im ersten Schritt wird ein einzelner Durchlauf durch die Daten vorgenommen, bei dem die Eingaberohdaten zu einem verwaltbaren Set von Subclustern komprimiert werden. Im zweiten Schritt wird eine hierarchische Clusterbildungsmethode verwendet, mit der die Subcluster zu immer größeren Clustern zusammengeführt werden. Dabei ist kein erneuter Durchlauf durch die Daten erforderlich. Die hierarchische Clusterbildung bietet den Vorteil, dass vorab keine Clusteranzahl ausgewählt werden muss. Bei vielen hierarchischen Methoden zur Clusterbildung werden einzelne Datensätze als Startcluster verwendet, die dann rekursiv zu noch größeren Clustern zusammengeführt werden. Diese Methoden versagen häufig bei großen Datenmengen. Durch die anfängliche Vorclusterbildung von TwoStep wird die hierarchische Clusterbildung hingegen auch für große Datasets zu einem schnellen Verfahren.

Hinweis: Das entstehende Modell hängt bis zu einem gewissen Grad von der Reihenfolge der Trainingsdaten ab. Eine Änderung der Datenreihenfolge und ein erneutes Erstellen des Modells kann zu einem anderen endgültigen Clustermodell führen.

Anforderungen. Zum Trainieren eines TwoStep-Clustermodells ist mindestens ein Feld mit der Rolle *Eingabe* erforderlich. Felder, deren Rolle auf *Ziel*, *Beides* oder *Keine* festgelegt ist, werden ignoriert. Der TwoStep-Clusteralgorithmus verarbeitet keine fehlenden Werte. Bei der Modellerstellung werden Datensätze, die in einem der Eingabefelder Leerzeichen enthalten, ignoriert.

Stärken. TwoStep-Clustering kann gemischte Feldtypen verarbeiten und ist in der Lage, große Datasets effizient zu verarbeiten. Es verfügt außerdem über die Fähigkeit, mehrere Clusterlösungen zu testen und die beste auszuwählen, sodass Sie nicht wissen müssen, wie viele Cluster Sie am Anfang abrufen müssen. TwoStep-Cluster können so eingestellt werden, dass **Ausreißer** oder äußerst unwahrscheinliche Fälle, die Ihre Ergebnisse verfälschen könnten, automatisch ausgeschlossen werden.

Optionen für TwoStep-Clusterknotenmodelle

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Numerische Felder standardisieren. Standardmäßig nimmt TwoStep eine Standardisierung aller numerischen Eingabefelder auf dieselbe Skalierung vor, d. h. ein Mittelwert von 0 und eine Abweichung von 1. Wenn Sie die ursprüngliche Skalierung der numerischen Felder beibehalten möchten, inaktivieren Sie diese Option. Symbolische Felder sind davon nicht betroffen.

Ausreißer ausschließen. Wenn Sie diese Option wählen, werden Datensätze, die nicht in ein betrachtetes Cluster zu passen scheinen, automatisch von der Analyse ausgeschlossen. So wird eine Verfälschung des Ergebnisses durch derartige Fälle verhindert.

Die Erkennung von Ausreißern erfolgt während des Schritts der Vorclusterbildung. Bei Auswahl dieser Option werden Subcluster, die im Vergleich zu anderen Subclustern wenige Datensätze enthalten, als potenzielle Ausreißer eingestuft und der Baum der Subcluster wird unter Ausschluss dieser Datensätze neu aufgebaut. Die Größe, unter der angenommen wird, dass Subcluster potenzielle Ausreißer enthalten, wird von der Option **Prozent** kontrolliert. Einige dieser potenziellen Ausreißerdatensätze können den neu erstellten Subclustern hinzugefügt werden, wenn sie den neuen Subcluster-Profilen ausreichend ähneln. Die

übrigen potenziellen Ausreißer, die nicht zusammengeführt werden können, werden als Ausreißer eingestuft, zu einem Cluster "Rauschen" hinzugefügt und vom Schritt der hierarchischen Clusterbildung ausgeschlossen.

Beim *Scoring* von Daten mit einem TwoStep-Modell, das Ausreißer verarbeitet, werden neue Fälle, die mehr als eine bestimmte Grenzdistanz (basierend auf der Log-Wahrscheinlichkeit) vom nächsten betrachteten Cluster entfernt sind, als Ausreißer eingestuft und dem Cluster "Rauschen" mit dem Namen -1 zugewiesen.

Clusterbeschriftung. Geben Sie das Format für das generierte Feld "Clusterzugehörigkeit" an. Die Clusterzugehörigkeit kann als **Zeichenfolge** mit dem festgelegten **Beschriftungspräfix** (z. B. `"Cluster 1"`, `"Cluster 2"` usw.) oder als **Zahl** angegeben werden.

Anzahl der Cluster automatisch berechnen. TwoStep-Cluster können eine große Anzahl von Clusterlösungen sehr rasch analysieren, um die optimale Anzahl von Clustern für die Trainingsdaten auszuwählen. Geben Sie einen Bereich der Lösungen an, die ausprobiert werden sollen, indem Sie die **maximale** und **minimale** Anzahl der Cluster festlegen. TwoStep ermittelt die optimale Anzahl von Clustern in einem zweistufigen Prozess. In der ersten Stufe wird eine Obergrenze für die Anzahl der Cluster in dem Modell basierend darauf, wie sich das BIC (Bayes Information Criterion) beim Hinzufügen weiterer Cluster ändert, ausgewählt. In der zweiten Stufe wird die Änderung in der Mindestdistanz zwischen Clustern für alle Modelle mit weniger Clustern gesucht als die Mindest-BIC-Lösung. Das endgültige Clustermodell wird anhand der größten Distanzänderung ermittelt.

Anzahl der Cluster angeben. Wenn Sie wissen, wie viele Cluster in Ihr Modell einzubeziehen sind, wählen Sie diese Option und geben Sie die Anzahl der Cluster ein.

Distanzmaß. Mit dieser Auswahl legen Sie fest, wie Ähnlichkeiten zwischen zwei Clustern verarbeitet werden.

- **Log-Likelihood.** Mit dem Likelihood-Maß wird eine Wahrscheinlichkeitsverteilung für die Variablen vorgenommen. Bei stetigen Variablen wird von einer Normalverteilung, bei kategorialen Variablen von einer multinomialen Verteilung ausgegangen. Bei allen Variablen wird davon ausgegangen, dass sie unabhängig sind.
- **Euklidisch.** Das Euklidische Maß bezeichnet die "gerade" Distanz zwischen zwei Clustern. Es kann nur dann verwendet werden, wenn es sich bei sämtlichen Variablen um stetige Variablen handelt.

Clusterkriterium. Mit dieser Auswahl legen Sie fest, wie die Anzahl der Cluster vom automatischen Clusteralgorithmus bestimmt wird. Angegeben werden kann entweder das Bayes-Informationskriterium (BIC) oder das Akaikes-Informationskriterium (AIC).

Modellnuggets vom Typ "TwoStep-Cluster"

Modellnuggets für TwoStep-Clustermodelle enthalten alle Informationen, die vom Clustermodell erfasst wurden, sowie Informationen zu den Trainingsdaten und dem Schätzvorgang.

Wenn Sie einen Stream ausführen, der ein Modellnugget vom Typ "TwoStep-Cluster" enthält, fügt der Knoten ein neues Feld hinzu, das die Clusterzugehörigkeit für den betreffenden Datensatz enthält. Der neue Feldname wird aus dem Modellnamen abgeleitet, dem das Präfix *$T-* hinzugefügt wird. Beispiel: Wenn das Modell den Namen *TwoStep* trägt, erhält das neue Feld den Namen *$T-TwoStep*.

Ein leistungsfähiges Verfahren, um Einblicke in das Two-Step-Modell zu gewinnen, besteht darin, mithilfe der Regelinduktion die Merkmale zu erkennen, die die vom Modell gefundenen Cluster unterscheiden. Weitere Informationen finden Sie im Thema „C5.0-Knoten” auf Seite 106. Dadurch wird der Cluster-Viewer angezeigt, der eine grafische Darstellung der Cluster, Felder und Wichtigkeitsgrade bietet. Weitere Informationen finden Sie im Thema „Cluster-Viewer - Registerkarte "Modell"” auf Seite 224.

Allgemeine Informationen zum Verwenden des Modellbrowsers finden Sie unter „Durchsuchen von Modellnuggets” auf Seite 42

Übersicht über das TwoStep-Modell

Auf der Registerkarte "Übersicht" für ein Modellnugget vom Typ "TwoStep-Cluster" werden die Anzahl der gefundenen Cluster sowie Informationen zu den Trainingsdaten, dem Schätzvorgang und den verwendeten Aufbaueinstellungen angezeigt.

Weitere Informationen finden Sie im Thema „Durchsuchen von Modellnuggets” auf Seite 42.

Cluster-Viewer

Clustermodelle werden üblicherweise verwendet, um Gruppen (oder Cluster) ähnlicher Datensätze zu finden, die auf den untersuchten Variablen basieren, wobei die Ähnlichkeit zwischen Elementen derselben Gruppe hoch und die Ähnlichkeit zwischen Elementen verschiedener Gruppen niedrig ist. Die Ergebnisse können zur Identifizierung von Zusammenhängen verwendet werden, die ansonsten nicht offensichtlich wären. So kann es zum Beispiel die Clusteranalyse von Kundenpräferenzen, Einkommensniveau und Kaufgewohnheiten ermöglichen, die Kundentypen zu identifizieren, die mit größerer Wahrscheinlichkeit auf eine bestimmte Marketingkampagne ansprechen.

Es gibt zwei Ansätze bei der Interpretierung der Ergebnisse in einer Clusterdarstellung:

- Untersuchen der Cluster, um die Merkmale zu bestimmen, die in einem Cluster eindeutig sind. *Enthält ein Cluster sämtliche Käufer mit hohem Einkommen? Enthält dieser Cluster mehr Datensätze als die anderen?*
- Untersuchen von Feldern in allen Clustern, um zu bestimmen, wie die Werte in den Clustern verteilt sind. *Ist der Bildungsstand entscheidend für die Zugehörigkeit zu einem Cluster? Spielt ein hoher Kreditrahmen eine Rolle bei der Zugehörigkeit zu einem Cluster oder einem anderen?*

Wenn Sie die Hauptansicht und die zahlreichen verknüpften Ansichten in der Clusteranzeige nutzen, lassen sich diese Fragen beantworten.

Die folgenden Clustermodellnuggets können in IBM SPSS Modeler generiert werden:

- Kohonen-Netzmodellnugget
- K-Means-Modellnugget
- TwoStep-Clustermodellnugget

Klicken Sie für weitere Informationen über die Clustermodellnuggets mit der rechten Maustaste auf den Modellknoten und wählen Sie **Durchsuchen** aus dem Kontextmenü (oder **Bearbeiten** für Knoten in einem Strom). Wenn Sie den Modellierungsknoten Auto Cluster verwenden, doppelklicken Sie auf den erforderlichen Cluster-Nugget in dem Auto Cluster Modellnugget. Weitere Informationen finden Sie im Thema „Knoten "Autom. Cluster"” auf Seite 75.

Cluster-Viewer - Registerkarte "Modell"

Die Registerkarte "Modell" für Clustermodelle zeigt eine grafische Darstellung von Übersichtsstatistiken und die Verteilung von Feldern zwischen Clustern; bekannt als **Cluster-Viewer**.

Hinweis: Die Registerkarte "Modell" ist für Modelle, die in IBM SPSS Modeler-Versionen vor Version 13 erstellt wurden, nicht verfügbar.

Die Clusteranzeige besteht aus zwei Bereichen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich. Es gibt zwei Hauptansichten:

- Modellübersicht (Standard). Weitere Informationen finden Sie im Thema „Ansicht "Modellübersicht"” auf Seite 225.
- Cluster. Weitere Informationen finden Sie im Thema „Clusteransicht” auf Seite 225.

Es gibt vier verknüpfte/Hilfsansichten:

- Prädiktoreinfluss. Weitere Informationen finden Sie im Thema „Ansicht "Prädiktoreinfluss im Cluster"" auf Seite 227.
- Clustergrößen (Standard). Weitere Informationen finden Sie im Thema „Clustergrößenansicht" auf Seite 227.
- Zellenverteilung. Weitere Informationen finden Sie im Thema „Zellverteilungsansicht" auf Seite 227.
- Clustervergleich. Weitere Informationen finden Sie im Thema „Clustervergleichsansicht" auf Seite 227.

Ansicht "Modellübersicht"

Die Ansicht "Modellübersicht" zeigt eine Momentaufnahme oder eine Übersicht des Clustermodells einschließlich eines schattierten Silhouettenmaßes der Clusterkohäsion und Clusterseparation, um schlechte, mittelmäßige und gute Ergebnisse anzuzeigen. Anhand dieser Momentaufnahme erkennen Sie schnell, ob die Qualität schlecht ist, sodass Sie dann gegebenenfalls zum Modellierungsknoten zurückkehren und die Clustermodell-Einstellungen ändern können, um ein besseres Ergebnis zu erzielen.

Die Ergebnisse "schlecht", "mittelmäßig" oder "gut" basieren auf der Arbeit von Kaufman und Rousseeuw (1990) zur Interpretation von Clusterstrukturen. In der Ansicht "Modellübersicht" entspricht ein gutes Ergebnis Daten, die von Kaufman und Rousseeuw als annehmbarer oder starker Hinweis auf eine Clusterstruktur eingestuft werden, "mittelmäßig" entspricht ihrer Einstufung als schwacher Hinweis und "schlecht" entspricht ihrer Einstufung als kein signifikanter Hinweis.

Das Silhouettenmaß ist ein Durchschnitt aller Datensätze, (B−A) / max(A,B), wobei A die Distanz der Datensatzes zum zugehörigen Clusterzentrum ist und B die Distanz des Datensatzes zum nächsten nicht zugehörigen Clusterzentrum. Ein Silhouettenkoeffizient von 1 würde bedeuten, dass alle Fälle direkt in ihren Clusterzentren liegen. Der Wert −1 würde bedeuten, dass alle Fälle in den Clusterzentren anderer Cluster liegen. Ein Wert von 0 bedeutet, dass die Fälle im Durchschnitt gleich weit entfernt von ihrem eigenen Clusterzentrum und dem nächsten benachbarten Cluster liegen.

Die Übersicht beinhaltet eine Tabelle, die folgende Daten enthält:

- **Algorithmus.** Der verwendete Clustering-Algorithmus, zum Beispiel "TwoStep".
- **Eingabemerkmale.** Die Anzahl der Felder, auch bekannt als **Eingaben** oder **Einflussgrößen**.
- **Cluster.** Die Anzahl der Cluster in der Lösung.

Clusteransicht

Die Clusteransicht enthält ein Cluster-nach-Funktionen-Raster mit Clusternamen, -größen und -profilen für jeden Cluster.

Die Spalten in der Tabelle enthalten die folgenden Informationen:

- **Cluster.** Die Clusternummern werden von dem Algorithmus erstellt.
- **Beschriftung.** Beschriftungen für jeden Cluster (ist standardmäßig leer). Doppelklicken Sie in die Zelle, um eine Beschriftung einzugeben, die den Clusterinhalt beschreibt; zum Beispiel "Käufer von Luxusautos".
- **Beschreibung.** Beschreibung des Clusterinhalts (ist standardmäßig leer). Doppelklicken Sie in die Zelle, um eine Beschreibung des Clusters einzugeben, zum Beispiel "Alter 55+, Berufstätige, Einkommen über $100.000".
- **Größe.** Die Größe jedes Clusters als Prozentsatz der gesamten Clusterstichprobe. Jede Größenzelle in der Tabelle zeigt einen vertikalen Balken, der den Größenprozentsatz innerhalb des Clusters, einen Größenprozentsatz in numerischem Format und die Clusterfallzahl anzeigt.
- **Strukturen.** Die einzelnen Eingaben oder Einflussgrößen, standardmäßig nach Gesamtwichtigkeit sortiert. Wenn Spalten die gleiche Größe aufweisen, werden sie in aufsteigender Sortierfolge ihrer Clusternummern angezeigt.

Die Gesamtwichtigkeit des Merkmals wird von der Farbe der Zellenhintergrundschattierung angezeigt; das wichtigste Merkmal ist am dunkelsten, das am wenigsten wichtige Merkmal ist unschattiert. Ein Hinweis oberhalb der Tabelle erläutert die Wichtigkeit, die jeder Merkmalszellfarbe zugewiesen ist.

Wenn Sie mit der Maus über eine Zelle fahren, wird der volle Name/die Beschriftung des Merkmals und der Wichtigkeitswert der Zelle angezeigt. Je nach Anzeige- und Merkmalstyp können auch weitere Informationen angezeigt werden. In der Ansicht "Clusterzentrum" zählen die Zellenstatistik und der Zellenwert dazu, beispielsweise "Mittelwert: 4,32". Bei kategorialen Merkmalen zeigt die Zelle den Namen der häufigsten (modalen) Kategorie und den dazugehörigen Prozentsatz an.

In der Ansicht "Cluster" können Sie verschiedene Anzeigearten für die Clusterinformationen auswählen:

- Cluster und Merkmale transponieren. Weitere Informationen finden Sie im Thema „Cluster und Merkmale transponieren".
- Merkmale sortieren. Weitere Informationen finden Sie im Thema „Merkmale sortieren".
- Cluster sortieren. Weitere Informationen finden Sie im Thema „Cluster sortieren.".
- Zelleninhalte auswählen. Weitere Informationen finden Sie im Thema „Zelleninhalt".

Cluster und Merkmale transponieren: Standardmäßig werden Cluster als Spalten und Merkmale als Zeilen angezeigt. Um die Anzeige umzudrehen, klicken Sie auf die Schaltfläche **Cluster und Merkmale transponieren** links von der Schaltfläche **Merkmale sortieren nach**. Dies kann zum Beispiel wünschenswert sein, wenn zahlreiche Cluster angezeigt werden, um den horizontalen Bildlauf bei der Datenansicht zu verringern.

Merkmale sortieren: Die Schaltflächen **Merkmale sortieren nach** ermöglichen Ihnen die Auswahl, wie Merkmalzellen angezeigt werden:

- **Gesamtwichtigkeit.** Das ist die standardmäßige Sortierfolge. Die Merkmale werden in absteigender Sortierfolge der Gesamtwichtigkeit sortiert, und die Sortierfolge ist dieselbe bei allen Clustern. Wenn Merkmale gebundene Wichtigkeitswerte aufweisen, sind die gebundenen Merkmale in aufsteigender Sortierfolge der Merkmalnamen aufgelistet.
- **Wichtigkeit innerhalb des Clusters.** Die Merkmale werden hinsichtlich ihrer Wichtigkeit für jeden Cluster sortiert. Wenn Merkmale gebundene Wichtigkeitswerte aufweisen, sind die gebundenen Merkmale in aufsteigender Sortierfolge der Merkmalnamen aufgelistet. Wenn diese Option ausgewählt wird, variiert üblicherweise die Sortierfolge in den Clustern.
- **Name.** Die Merkmale werden nach Namen in alphabetischer Reihenfolge sortiert.
- **Datenfolge.** Die Merkmale werden nach ihrer Reihenfolge im Datensatz sortiert.

Cluster sortieren.: Standardmäßig werden Cluster ihrer Größe nach absteigend sortiert. Mit den Schaltflächen **Cluster sortieren nach** können Sie die Cluster nach Namen in alphabetischer Reihenfolge sortieren, oder, wenn Sie eindeutige Beschriftungen erstellt haben, stattdessen auch in alphanumerischer Beschriftungsreihenfolge.

Merkmale mit derselben Beschriftung werden nach Clustername sortiert. Wenn die Cluster nach Beschriftung sortiert sind und Sie die Beschriftung eines Clusters bearbeiten, wird die Sortierfolge automatisch aktualisiert.

Zelleninhalt: Mit den Schaltflächen **Zellen** können Sie die Anzeige der Zelleninhalte für Merkmal- und Evaluierungsfelder ändern.

- **Clusterzentren.** Standardmäßig zeigen Zellen Namen/Beschriftungen und die Lage (zentrale Tendenz) für jede Cluster/Merkmal-Kombination an. Für stetige Felder wird der Mittelwert angezeigt und für kategorische Felder der Modus (die am häufigsten auftretende Kategorie) mit Kategorieprozentsatz.
- **Absolute Verteilungen.** Zeigt die Merkmalnamen/-beschriftungen und die absoluten Verteilungen der Merkmale in jedem Cluster. Bei kategorischen Merkmalen werden Balkendiagramme angezeigt, mit überlagerter Anzeige der Kategorien, die nach ihren Datenwerten aufsteigend geordnet sind. Bei steti-

gen Merkmalen stellt die Anzeige ein gleichmäßiges Dichtediagramm dar, bei dem die gleichen Endpunkte und Intervalle für jeden Cluster verwendet werden.

Die intensiv rote Anzeige stellt die Clusterverteilung dar, wogegen die blassere Anzeige die Gesamtdaten repräsentiert.

- **Relative Verteilungen.** Zeigt die Merkmalnamen/-beschriftungen und die relativen Verteilungen in den Zellen. Im Allgemeinen sind die Anzeigen vergleichbar mit denen für absolute Verteilungen, nur dass stattdessen die relativen Verteilungen dargestellt sind.

  Die intensiv rote Anzeige stellt die Clusterverteilung dar, wogegen die blassere Anzeige die Gesamtdaten repräsentiert.
- **Basisansicht.** Bei sehr vielen Clustern kann es schwierig sein, sämtliche Details ohne Bildlauf zu sehen. Wählen Sie diese Ansicht, um den Bildlauf einzuschränken und die Anzeige auf eine kompaktere Version der Tabelle zu ändern.

Ansicht "Prädiktoreinfluss im Cluster"

Die Ansicht "Prädiktoreinfluss" zeigt die relative Wichtigkeit jedes Felds bei Schätzung des Modells.

Clustergrößenansicht

Die Clustergrößenansicht zeigt ein Kreisdiagramm, das sämtliche Cluster enthält. In jedem Ausschnitt wird die prozentuale Größe des Clusters angezeigt; fahren Sie mit der Maus über einen Ausschnitt, um den Zahlwert in diesem Ausschnitt anzuzeigen.

Unterhalb des Diagramms sind in einer Tabelle die folgenden Informationen aufgelistet:

- Größe des kleinsten Clusters (als Zahlwert und Prozentsatz des Ganzen).
- Größe des größten Clusters (als Zahlwert und Prozentsatz des Ganzen).
- Verhältnis der Größe des größten Clusters zum kleinsten Cluster.

Zellverteilungsansicht

Die Zellverteilungsansicht zeigt ein erweitertes, detaillierteres Diagramm der Datenverteilung für jede Merkmalszelle, die Sie in der Tabelle im Clusterhauptbereich auswählen.

Clustervergleichsansicht

Die Clustervergleichsansicht ist eine tabellarische Grafik, bei der die Merkmale in den Zeilen und die ausgewählten Cluster in den Spalten dargestellt werden. Mit dieser Ansicht lassen sich die Faktoren besser verstehen, die die Cluster ausmachen; außerdem hilft sie dabei, die Unterschiede zwischen den Clustern zu erkennen - nicht nur im Vergleich zum Gesamtdatensatz, sondern auch untereinander.

Zum Auswählen der Cluster für die Ansicht klicken Sie oben auf die Clusterspalte im Clusterhauptbereich. Wenn Sie die Steuertaste oder die Umschalttaste beim Klicken gedrückt halten, können Sie mehrere Cluster zum Vergleich auswählen oder wieder aus der Auswahl entfernen.

Hinweis: Sie können bis zu fünf Cluster zur Anzeige auswählen.

Die Cluster werden in der Reihenfolge ihrer Auswahl angezeigt, während die Reihenfolge der Felder mit der Option **Merkmale sortieren nach** festgelegt wird. Wenn Sie **Wichtigkeit innerhalb der Cluster** auswählen, werden die Felder immer nach ihrer Gesamtwichtigkeit sortiert.

Die Hintergrunddiagramme zeigen die Gesamtverteilungen der Merkmale:

- Kategorische Merkmale sind als Punktdiagramme dargestellt, wobei die Größe des Punktes die häufigste/typische Kategorie für jeden Cluster (nach Merkmal) anzeigt.
- Stetige Merkmale sind als Boxplots angezeigt, die die Gesamtmediane und die Interquartilbereiche anzeigen.

Vor diesen Hintergrundansichten sind Boxplots für ausgewählte Cluster dargestellt:

- Bei stetigen Merkmalen geben quadratische Punktmarkierungen und horizontale Linien den Median- und Interquartilbereich für die einzelnen Cluster an.
- Jeder Cluster ist mit einer anderen Farbe gekennzeichnet, die oben an der Ansicht angezeigt wird.

Navigieren in der Clusteranzeige

Bei der Clusteranzeige handelt es sich um eine interaktive Anzeige. Sie verfügen über folgende Möglichkeiten:

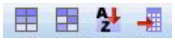
- Auswählen eines Felds oder eines Clusters für weitere Details
- Vergleichen von Clustern, um die Objekte von Interesse auszuwählen
- Verändern der Anzeige
- Achsen transponieren.
- Erzeugen von Knoten zum Ableiten, Filtern und Auswählen unter Verwendung des Menüs "Erzeugen".

Verwendung der Symbolleisten

Sie können die Informationen, die in den Panels links und rechts angezeigt werden, mithilfe der Symbolleistenoptionen steuern. Mit der Symbolleistensteuerung können Sie die Ausrichtung der Anzeige ändern (oben-unten, links-rechts oder rechts-links). Außerdem können Sie die Clusteranzeige auf die Standardeinstellungen zurücksetzen und ein Dialogfeld öffnen, um den Inhalt der Clusteransicht im Hauptbereich zu spezifizieren.

Die Optionen **Merkmale sortieren nach**, **Cluster sortieren nach**, **Zellen** und **Anzeige** sind nur verfügbar, wenn Sie die Ansicht **Cluster** im Hauptbereich auswählen. Weitere Informationen finden Sie im Thema „Clusteransicht" auf Seite 225.

Tabelle 12. Symbolleistensymbole.

Symbol	Thema
	Siehe Cluster und Merkmale transponieren
	Siehe Merkmale sortieren nach
	Siehe Cluster sortieren nach
	Siehe Zellen

Erzeugen von Knoten aus Clustermodellen

Mit dem Menü "Erzeugen" können Sie auf der Basis des Clustermodells neue Knoten erstellen. Diese Option ist auf der Registerkarte "Modell" des erzeugten Modells verfügbar und ermöglicht es, Knoten auf der Basis der aktuellen Anzeige oder einer Auswahl zu erzeugen (das heißt, alle sichtbaren oder alle ausgewählten Cluster). Sie können zum Beispiel ein einzelnes Merkmal auswählen und anschließend einen Filterknoten erstellen, um alle anderen (nicht sichtbaren) Merkmale zu verwerfen. Die erzeugten Knoten werden unzusammenhängend im Erstellungsbereich platziert. Sie können außerdem eine Kopie des Modellnuggets erstellen und zur Modellpalette hinzufügen. Denken Sie daran, vor der Ausführung die Knoten zu verknüpfen und alle erwünschten Änderungen vorzunehmen.

- **Modellierungsknoten generieren.** Erzeugt einen Modellierungsknoten im Streamerstellungsbereich. Das könnte zum Beispiel nützlich sein, wenn Sie bei einem Stream diese Modelleinstellungen verwenden möchten, aber nicht mehr über den Modellierungsknoten verfügen, um sie zu erzeugen.
- **Modell in Palette.** Erstellt ein Nugget auf der Modellpalette. Das ist nützlich, wenn Sie von einem Kollegen einen Stream, der das Modell enthält, jedoch nicht das Modell selbst erhalten.

- **Filterknoten.** Erstellt einen neuen Filterknoten, um Felder zu filtern, die von dem Clustermodell nicht verwendet werden und/oder in der aktuellen Ansicht der Clusteranzeige nicht sichtbar sind. Wenn es von diesem Clusterknoten einen vorgelagerten Typenknoten gibt, werden Felder mit der Rolle *Ziel* von dem erzeugten Filterknoten verworfen.
- **Filterknoten (aus Auswahl).** Erzeugt einen neuen Filterknoten, um die Felder auf der Basis der Auswahl der Clusteranzeige zu filtern. Mehrere Felder können Sie auswählen, indem Sie bei gedrückter Steuertaste klicken auf die Felder klicken. Die in der Clusteranzeige ausgewählten Felder werden nachgelagert verworfen, doch Sie können diese Einstellung ändern, indem Sie den Filterknoten vor dem Ausführen bearbeiten.
- **Auswahlknoten.** Erstellt einen neuen Auswahlknoten, um Datensätze basierend auf ihrer Zugehörigkeit zu den Clustern, die in der aktuellen Ansicht der Clusteranzeige sichtbar sind, auszuwählen. Eine Auswahlbedingung wird automatisch generiert.
- **Auswahlknoten (aus Auswahl).** Erstellt einen neuen Auswahlknoten, um Datensätze basierend auf ihrer Zugehörigkeit zu Clustern, die in der Clusteranzeige ausgewählt wurden, auszuwählen. Wählen Sie mehrere Cluster aus, indem Sie bei gedrückter Steuertaste klicken.
- **Ableitungsknoten.** Erzeugt einen neuen Ableitungsknoten, der ein Flagfeld erstellt, das Datensätzen einen Wert *Wahr* oder *Falsch* zuweist, basierend auf der Zugehörigkeit zu allen in der Clusteranzeige sichtbaren Clustern. Eine Ableitungsbedingung wird automatisch generiert.
- **Ableitungsknoten (aus Auswahl).** Erstellt einen neuen Ableitungsknoten, der ein Flagfeld basierend auf der Zugehörigkeit zu Clustern, die in der Clusteranzeige ausgewählt wurden, ableitet. Wählen Sie mehrere Cluster aus, indem Sie bei gedrückter Steuertaste auf die Cluster klicken.

Im Menü "Erzeugen" können Sie nicht nur Knoten, sondern auch Diagramme erstellen. Weitere Informationen finden Sie im Thema „Erzeugen von Diagrammen aus Clustermodellen”.

Anzeige "Clusteransicht steuern"

Um zu steuern, was in der Clusteransicht im Hauptbereich angezeigt wird, klicken Sie auf die Schaltfläche **Anzeige**. Der Anzeigedialog wird geöffnet.

Strukturen. Standardmäßig ausgewählt. Inaktivieren Sie das Kästchen, um alle Eingabemerkmale auszublenden.

Evaluierungsfelder. Wählen Sie die anzuzeigenden Evaluierungsfelder aus (Felder, die nicht für die Erstellung des Clustermodells verwendet, sondern an die Modellanzeige zur Evaluierung der Cluster gesendet werden); standardmäßig werden keine angezeigt. *Hinweis*: Dieses Kontrollkästchen ist nicht verfügbar, wenn keine Evaluierungsfelder verfügbar sind.

Clusterbeschreibungen. Standardmäßig ausgewählt. Inaktivieren Sie das Kontrollkästchen, um alle Clusterbeschreibungszellen auszublenden.

Clustergrößen. Standardmäßig ausgewählt. Inaktivieren Sie das Kontrollkästchen, um alle Clustergrößenzellen auszublenden.

Maximale Anzahl an Kategorien. Geben Sie die maximale Anzahl an Kategorien an, die in den Diagrammen der kategorischen Merkmale angezeigt werden sollen; der Standard ist 20.

Erzeugen von Diagrammen aus Clustermodellen

Clustermodelle beinhalten viele Informationen; sie sind jedoch häufig in einem Format, auf das Fachanwender nicht so einfach zugreifen können. Zur Bereitstellung der Daten auf eine Art, die problemlos in Geschäftsberichte, Präsentationen u.s.w. integriert werden kann, können aus ausgewählten Daten Diagramme erstellt werden. In der Clusteranzeige können Sie zum Beispiel ein Diagramm für einen ausgewählten Cluster erstellen, wobei das Diagramm nur für die Fälle in diesem Cluster erzeugt wird.

Hinweis: Sie können ein Diagramm nur über den Cluster-Viewer erstellen, wenn das Modellnugget an andere Knoten in einem Stream angehängt ist.

Generieren eines Diagramms

1. Öffnen Sie das Modellnugget, das die Clusteranzeige enthält.
2. Wählen Sie auf der Registerkarte "Modell" in der Dropdown-Liste *Ansicht* die Option **Cluster** aus.
3. Wählen Sie in der Hauptansicht den oder die Cluster, für die Sie ein Diagramm erstellen möchten, aus.
4. Wählen Sie im Menü "Erzeugen" den Menüpunkt **Diagramm (aus der Auswahl)** aus; die Registerkarte "Basis" auf der Diagrammtafel wird angezeigt.

   Hinweis: Wenn Sie die Diagrammtafel auf diese Art anzeigen, sind nur die Registerkarten "Basis" und "Details" verfügbar.
5. Mithilfe der Einstellungen auf den Registerkarten "Basis" oder "Details" können Sie die Details angeben, die auf dem Diagramm angezeigt werden sollen.
6. Klicken Sie auf "OK", um das Diagram zu erstellen.

Die Diagrammüberschrift identifiziert den Modelltyp und den oder die Cluster, die eingeschlossen wurden.

Kapitel 12. Assoziationsregeln

Assoziationsregeln ordnen eine bestimmte Schlussfolgerung (beispielsweise den Kauf eines bestimmten Produkts) einer Menge von Bedingungen (beispielsweise dem Kauf mehrerer anderer Produkte) zu. Beispiel: Die Regel

```
Bier <= -Dosengemüse & TK-Fertiggericht (173, 17,0 %, 0,84)
```

besagt, dass *Bier* häufig vorkommt, wenn *Dosengemüse* und *TK-Fertiggericht* zusammen vorkommen. Die Regel ist zu 84 % zuverlässig und auf 17 % der Daten, also 173 Datensätze, anwendbar. Algorithmen für Assoziationsregeln finden automatisch die Zuordnungen, die Sie manuell finden könnten, wenn Sie Visualisierungstechniken, wie den Netzdiagrammknoten, anwenden.

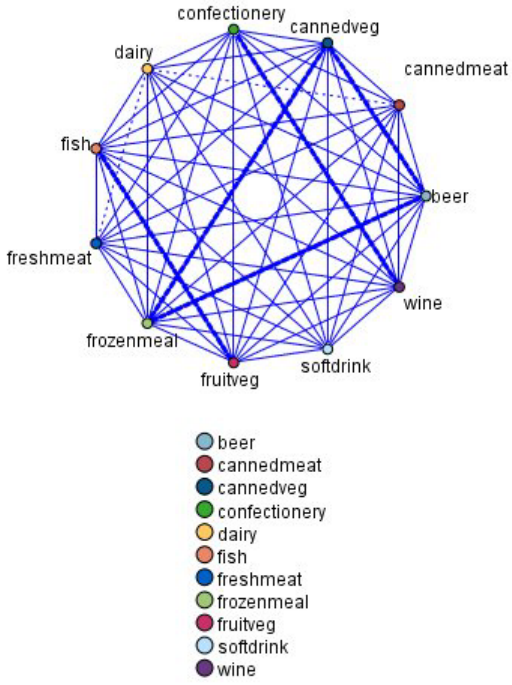


Abbildung 45. Netzdiagrammknoten, der Assoziationen zwischen Elementen des Warenkorbs anzeigt

Der Vorteil von Algorithmen für Assoziationsregeln im Vergleich zu Algorithmen für Standardentscheidungsbäume (C5.0 und C&R-Bäume) besteht darin, dass zwischen *beliebigen* Attributen Verbindungen bestehen können. Ein Entscheidungsbaumalgorithmus erstellt Regeln mit nur einer Schlussfolgerung, während Assoziationsalgorithmen viele Regeln zu finden versuchen, von denen jede zu einer anderen Schlussfolgerung kommen kann.

Der Nachteil von Assoziationsregeln besteht darin, dass sie versuchen, Muster innerhalb eines potenziell sehr großen Suchbereichs zu finden, also mehr Zeit für die Ausführung in Anspruch nehmen können als ein Entscheidungsbaumalgorithmus. Die Algorithmen verwenden eine Methode vom Typ **Generieren und Testen** zum Auffinden von Regeln, bei der einfache Regeln erstellt und mit dem Dataset verglichen werden. Die "guten" Regeln werden gespeichert und alle Regeln, die verschiedenen Bedingungen unterworfen sind, werden anschließend spezialisiert. **Spezialisierung** ist der Prozess, bei dem einer Regel Bedingungen hinzugefügt werden. Diese neuen Regeln werden dann mit den Daten verglichen und validiert. Die "besten" oder interessantesten Regeln werden dann gespeichert. Der Benutzer legt normalerweise einen Grenzwert für die mögliche Anzahl Antezedenzien in einer Regel fest. Es werden

außerdem verschiedene Techniken basierend auf der Informationstheorie oder effiziente Indizierungsschemata verwendet, um den potenziell großen Suchbereich zu reduzieren.

Am Ende der Verarbeitung wird eine Tabelle mit den besten Regeln ausgegeben. Im Gegensatz zu einem Entscheidungsbaum kann dieser Satz mit Assoziationsregeln nicht direkt dazu verwendet werden, Vorhersagen auf eine Weise zu machen, wie dies mit einem Standardmodell (z. B. einem Entscheidungsbaum oder neuronalen Netz) möglich ist. Dies ist auf die vielen möglichen Schlussfolgerungen für die Regeln zurückzuführen. Es ist eine weitere Stufe der Transformation erforderlich, um die Assoziationsregeln in ein Klassifizierungsregelset umzuwandeln. Deshalb sind die von Assoziationsalgorithmen erstellten Assoziationsregeln bekannt als **nicht verfeinerte Modelle**. Obwohl der Benutzer diese nicht verfeinerten Modelle durchsuchen kann, können Sie nicht ausdrücklich als Klassifizierungsmodelle verwendet werden, es sei denn, der Benutzer weist das System an, aus dem nicht verfeinerten Modell ein Klassifizierungsmodell zu generieren. Dies geschieht mit dem Browser über die Menüoption "Generieren".

Es werden zwei Algorithmen für Assoziationsregeln unterstützt:

Der Apriori-Knoten extrahiert ein Regelset aus den Daten und daraus die Regeln mit dem höchsten Informationsgehalt. Apriori bietet fünf verschiedene Methoden zur Auswahl von Regeln und verwendet ein ausgereiftes Indizierungsschema zur effizienten Verarbeitung großer Datasets. Bei großen Problemen ist Apriori in der Regel schneller zu trainieren, es gibt keine willkürliche Begrenzung für die Anzahl der Regeln, die beibehalten werden können, und es können Regeln mit bis zu 32 Vorbedingungen verarbeitet werden. Bei Apriori müssen alle Ein- und Ausgabefelder kategorial sein; dafür bietet es jedoch eine bessere Leistung, da es für diesen Datentyp optimiert ist.

Der Sequenzknoten erkennt Assoziationsregeln in sequenziellen oder zeitorientierten Daten. Eine Sequenz ist eine Liste mit Elementsets, die in einer vorhersagbaren Reihenfolge auftreten. Beispiel: Ein Kunde, der einen Rasierer und After-Shave-Lotion kauft, kauft möglicherweise beim nächsten Einkauf Rasiercreme. Der Sequenzknoten basiert auf dem CARMA-Assoziationsregelalgorithmus, der eine effiziente bidirektionale Methode zum Suchen von Sequenzen verwendet.

Tabellendaten im Vergleich zu Transaktionsdaten

Die von Assoziationsregelmodellen verwendeten Daten können im Transaktionsformat oder in Tabellenform (siehe unten) vorliegen. Hierbei handelt es sich um allgemeine Beschreibungen; die konkreten Anforderungen können abweichen, wie in der Dokumentation für die einzelnen Modelltypen erörtert. Beachten Sie: Beim Scoring von Modellen müssen die zu scorenden Daten dasselbe Format aufweisen wie die bei der Modellerstellung verwendeten Daten. Mit Tabellendaten erstellte Modelle können ausschließlich zum Scoring von Tabellendaten verwendet werden; mit Transaktionsdaten erstellte Modelle dienen ausschließlich zum Scoring von Transaktionsdaten.

Transaktionsformat

Transaktionsdaten weisen einen eigenen Datensatz für jede Transaktion bzw. jedes Element auf. Wenn ein Kunde beispielsweise mehrere Einkäufe tätigt, handelt es sich bei jedem um einen eigenen Datensatz, wobei die zugehörigen Elemente durch eine Kunden-ID verknüpft sind. Dies wird auch manchmal als **Kassenrollen**-Format bezeichnet.

Kunde	Kauf
E	Marmelade
Z	Milch
3	Marmelade
3	Brot

Kunde	Kauf
4	Marmelade
4	Brot
4	Milch

Apriori-, CARMA- und Sequenzknoten können jeweils Transaktionsdaten verwenden.

Tabellendaten

In Tabellendaten (auch als **Warenkorb**- oder **Wahrheitstabellen**-Daten bezeichnet) werden die Elemente durch gesonderte Flags dargestellt, wobei jedes Flagfeld für das Vorliegen bzw. die Abwesenheit eines bestimmten Elements steht. Jeder Datensatz steht für ein komplettes Set zugehöriger Elemente. Prinzipiell können Flagfelder kategorial oder numerisch sein; bei manchen Modellen können jedoch genauere Anforderungen gelten.

Kunde	Marmelade	Brot	Milch
E	T	F	F
Z	F	F	T
3	T	T	F
4	T	T	T

Apriori-, CARMA-, und Sequenzknoten können jeweils Tabellendaten verwenden.

Apriori-Knoten

Der Apriori-Knoten erkennt Assoziationsregeln in den Daten. Apriori bietet fünf verschiedene Methoden zur Auswahl von Regeln und verwendet ein ausgereiftes Indizierungsschema zur effizienten Verarbeitung großer Datasets.

Anforderungen. Um ein Apriori-Regelset zu erstellen, benötigen Sie mindestens ein *Eingabe*- und ein *Ziel*-Feld. Ein- und Ausgabefelder (mit der Rolle *Eingabe*, *Ziel* oder *Beides*) müssen symbolisch sein. Felder mit der Rolle *Keine* werden ignoriert. Feldtypen müssen vollständig instanziiert werden, bevor der Knoten ausgeführt wird. Die Daten können als Tabellen- oder Transaktionsdaten vorliegen. Weitere Informationen finden Sie im Thema „Tabellendaten im Vergleich zu Transaktionsdaten” auf Seite 232.

Stärken. Bei großen Problemen ist Apriori in der Regel schneller zu trainieren. Außerdem gibt es keine willkürliche Begrenzung für die Anzahl der Regeln, die beibehalten werden können, und es können Regeln mit max. 32 Vorbedingungen verarbeitet werden. Apriori bietet fünf verschiedene Trainingsmethoden und somit mehr Flexibilität bei der Anpassung der Data-Mining-Methode an das aktuelle Problem.

Modelloptionen für den Apriori-Knoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Minimale Antezedens-Unterstützung. Sie können ein Stützkriterium angeben, um die Regeln im Regelset beizubehalten. **Unterstützung** bezieht sich auf den Prozentwert der Datensätze in den Trainingsdaten, für die die Antezedenzien (der "Wenn"-Teil der Regel) wahr sind. (Beachten Sie, dass diese Definition von Unterstützung von der für CARMA- und Sequenzknoten verwendeten abweicht. Weitere Informationen finden Sie im Thema „Modelloptionen für den Sequenzknoten” auf Seite 249.) Wenn Sie Regeln erhalten, die für sehr kleine Subsets der Daten gelten, sollten Sie diese Einstellung erhöhen.

Hinweis: Die Definition von Unterstützung für Apriori basiert auf der Anzahl der Datensätze mit den Antezedenzien. Dies steht im Gegensatz zu den CARMA- und Sequenzalgorithmen, für die die Definition von Unterstützung auf der Anzahl der Datensätze mit allen Elementen in einer Regel basiert (d. h. sowohl Antezedenzien als auch Sukzedenzien). Die Ergebnisse der Assoziationsmodelle zeigen sowohl die Maße für die Antezedens- als auch die Regelunterstützung.

Minimale Regelkonfidenz. Sie können auch ein Konfidenzkriterium angeben. **Konfidenz** basiert auf den Datensätzen, für die die Antezedenzien der Regel wahr sind, und stellt den Prozentwert der Datensätze dar, für die auch die Sukzedenzien wahr sind. Mit anderen Worten, es ist der prozentuale Anteil der Vorhersagen, die, basierend auf der Regel, richtig sind. Regeln mit einer niedrigeren Konfidenz als das angegebene Kriterium werden verworfen. Wenn Sie zu viele Regeln oder uninteressante Regeln erhalten, sollten Sie diese Einstellung erhöhen. Wenn Sie zu wenige (oder überhaupt keine) Regeln erhalten, sollten Sie diese Einstellung reduzieren.

Maximale Anzahl von Antezedenzien. Sie können die maximale Anzahl an Vorbedingungen für jede beliebige Regel festlegen. Auf diese Weise können Sie die Komplexität der Regeln begrenzen. Wenn die Regeln zu komplex oder zu spezifisch sind, sollten Sie diese Einstellung reduzieren. Diese Einstellung hat auch einen großen Einfluss auf die Trainingszeit. Wenn das Training des Regelsets zu viel Zeit in Anspruch nimmt, sollten Sie diese Einstellung reduzieren.

Nur wahre Werte für Flags. Wenn diese Option für Daten in tabellarischer Form (Wahrheitswerttabelle) ausgewählt ist, enthalten die resultierenden Regeln lediglich wahre Werte. Auf diese Weise können Sie die Regeln einfacher verstehen. Die Option gilt nicht für Daten im Transaktionsformat. Weitere Informationen finden Sie im Thema „Tabellendaten im Vergleich zu Transaktionsdaten” auf Seite 232.

Optimieren. Wählen Sie Optionen aus, die die Leistung während der Modellerstellung basierend auf Ihren persönlichen Anforderungen erhöhen.

- Wählen Sie **Geschwindigkeit** aus, um den Algorithmus anzuweisen, zur Verbesserung der Leistung keinen Datenträgerüberlauf zu verwenden.
- Wählen Sie **Speicher** aus, um den Algorithmus anzuweisen, gegebenenfalls einen Datenträgerüberlauf zu verwenden und dafür Geschwindigkeitseinbußen hinzunehmen. Diese Option ist standardmäßig aktiviert. *Hinweis*: Bei der Ausführung im verteilten Modus kann diese Einstellung durch die in der Datei *options.cfg* angegebenen Administratoroptionen überschrieben werden. Weitere Informationen finden Sie im *IBM SPSS Modeler Server-Administratorhandbuch*.

Expertenoptionen für den Apriori-Knoten

Für Personen mit umfassenden Kenntnissen von Apriori ermöglichen die folgenden Expertenoptionen die Feinabstimmung des Induktionsvorgangs. Für den Zugriff auf die Expertenoptionen stellen Sie den Modus auf der Registerkarte "Experten" auf **Experten** ein.

Evaluierungsmaß. Apriori unterstützt fünf Methoden zur Auswertung möglicher Regeln.

- **Regelkonfidenz.** Die Standardmethode verwendet die Konfidenz (oder Genauigkeit) der Regel zur Regelevaluierung. Für dieses Maß ist die Option **Untergrenze der Auswertungsmaßnahme** inaktiviert, da sie aufgrund der Option **Minimale Regelkonfidenz** auf der Registerkarte "Modell" redundant ist. Weitere Informationen finden Sie im Thema „Modelloptionen für den Apriori-Knoten” auf Seite 233.
- **Konfidenzdifferenz.** (Auch als **absolute Konfidenz-Differenz zum Vorgänger** bezeichnet.) Dieses Evaluierungsmaß stellt die absolute Differenz zwischen der Regelkonfidenz und der vorherigen Konfidenz dar. Diese Option vermeidet dort einen Bias, wo die Ergebnisse nicht gleichverteilt sind. Dadurch kann verhindert werden, dass "offensichtliche" Regeln beibehalten werden. Zum Beispiel kann es der Fall sein, dass 80 % der Kunden Ihr beliebtestes Produkt kaufen. Eine Regel, die den Kauf eines beliebten Produkts mit Genauigkeit von 85 % vorhersagt, hilft Ihnen nicht viel weiter, auch wenn eine Genauigkeit von 85 % auf einer absoluten Skala sehr gut erscheint. Setzen Sie die Untergrenze des Evaluierungsmaßes auf die Mindestkonfidenzdifferenz, für die Regeln beibehalten werden sollen.

- **Konfidenzverhältnis.** (Auch als **Differenz des Konfidenzquotienten zur 1** bezeichnet.) Bei diesem Evaluierungsmaß handelt es sich um das Verhältnis der Regelkonfidenz zur vorherigen Konfidenz (oder, wenn das Verhältnis größer als 1 ist, der reziproke Wert) subtrahiert von 1. Wie bei der Konfidenzdifferenz berücksichtigt diese Methode andere Verteilungen als die Gleichverteilung. Sie eignet sich besonders gut, um Regeln zu finden, die seltene Ereignisse vorhersagen. Angenommen, es gibt eine seltene Krankheit, die nur bei 1 % der Patienten vorkommt. Eine Regel, mit der diese Bedingung in 10 % der Fälle vorhergesagt werden kann, ist im Gegensatz zur groben Schätzung eine erhebliche Verbesserung, auch wenn eine Genauigkeit von 10 % auf einer absoluten Skala nicht sehr beeindruckend erscheint. Setzen Sie die Untergrenze des Evaluierungsmaßes auf die Differenz, für die Regeln eingehalten werden sollen.
- **Informationsdifferenz.** (Auch als **Informationsdifferenz zum Vorgänger** bezeichnet.) Dieses Maß basiert auf dem Maß **Informationsgewinn**. Wenn die Wahrscheinlichkeit eines bestimmten Antezedens als logischer Wert betrachtet wird (ein **Bit**), ist der Informationsgewinn der Anteil dieses Bits, der basierend auf den Antezedenzien ermittelt werden kann. Die Informationsdifferenz ist die Differenz zwischen dem Informationsgewinn, gegeben die Antezedenzien, und dem Informationsgewinn, gegeben lediglich die vorher bestehende Konfidenz des Sukzedens. Ein wichtiges Merkmal dieser Methode ist, dass die Unterstützung berücksichtigt wird, sodass die Regeln, die für mehr Datensätze gelten, für ein bestimmtes Konfidenzniveau bevorzugt werden. Setzen Sie die Untergrenze des Evaluierungsmaßes auf die Informationsdifferenz, für die Regeln eingehalten werden sollen.

  Hinweis: Da die Skala für dieses Maß etwas weniger intuitiv ist als die anderen Skalen, müssen Sie möglicherweise mit verschiedenen Untergrenzen experimentieren, um ein zufriedenstellendes Regelset zu erhalten.
- **Normalisiertes Chi-Quadrat.** (Auch als **normalisiertes Chi-Quadrat-Maß** bezeichnet.) Dieses Maß ist ein statistischer Assoziationsindex zwischen Antezedenzien und Sukzedenzien. Dieses Maß wird auf Werte zwischen 0 und 1 normiert. Dieses Maß ist noch stärker abhängig von der Unterstützung als das Informationsdifferenzmaß. Setzen Sie die Untergrenze des Evaluierungsmaßes auf die Informationsdifferenz, für die Regeln eingehalten werden sollen.

  Hinweis: Da die Skala für dieses Maß wie beim Informationsdifferenzmaß etwas weniger intuitiv als die anderen Skalen ist, müssen Sie möglicherweise mit verschiedenen Untergrenzen experimentieren, um ein zufriedenstellendes Regelset zu erhalten.

Regeln ohne Antezedenzien zulassen. Wählen Sie diese Option aus, um Regeln zuzulassen, die lediglich das Sukzedens (Element oder Elementset) enthalten. Dies ist sinnvoll, wenn Sie an der Ermittlung häufig verwendeter Elemente oder Elementsets interessiert sind. Zum Beispiel ist `Konservengemüse` eine Regel mit nur einem Element ohne Antezedens, die angibt, dass der Kauf von *Konservengemüse* in den Daten häufig vorkommt. In manchen Fällen können Sie derartige Regeln aufnehmen, wenn Sie nur an den wahrscheinlichsten Vorhersagen interessiert sind. Diese Option ist standardmäßig inaktiviert. In der Regel wird die Antezedens-Unterstützung für Regeln ohne Antezedens als 100 % ausgedrückt. Die Regelunterstützung ist gleich der Konfidenz.

CARMA-Knoten

Der CARMA-Knoten verwendet einen Erkennungsalgorithmus für Regeln, um Assoziationsregeln in den Daten zu erkennen. Assoziationsregeln sind Anweisungen in der Form

falls *Antezedens* **dann** *Sukzedens*

Wenn ein Webkunde z. B. eine drahtlose Karte und einen drahtlosen High-end-Router kauft, wird er wahrscheinlich auch einen drahtlosen Musikserver kaufen, falls ihm einer angeboten wird. Beim CARMA-Modell wird ein Regelset aus den Daten extrahiert, ohne dass Sie Eingabe- oder Zielfelder angeben müssen. Dies bedeutet, dass die generierten Regeln für einen breiteren Anwendungsbereich verwendet werden können. So können Sie z. B. von diesem Knoten generierte Regeln verwenden, um eine Liste mit Produkten und Dienstleistungen (Antezedenzien) zu finden, deren Sukzedens das Element darstellt, das Sie in der Ferienzeit desselben Jahres bewerben möchten. Mit IBM SPSS Modeler können Sie ermitteln, welche Kunden die Vorgängerprodukte gekauft haben, und eine Marketing-Kampagne für das Nachfolger-Produkt ins Leben rufen.

Anforderungen. Im Gegensatz zu Apriori sind für den CARMA-Knoten die Felder *Eingabe* oder *Ziel* nicht erforderlich. Dies ist ein integraler Bestandteil der Funktionsweise des Algorithmus und entspricht dem Erstellen eines Apriori-Modells, wobei alle Felder auf *Beides* gesetzt sind. Sie können beschränken, welche Elemente nur als Antezedenzien oder Sukzedenzien aufgelistet werden, indem Sie das Modell filtern, nachdem es erstellt wurde. So können Sie z. B. den Modellbrowser verwenden, um eine Liste mit Produkten und Dienstleistungen (Antezedenzien) zu finden, deren Nachfolger (Sukzedens) das Element darstellt, das Sie in der Ferienzeit desselben Jahres bewerben möchten.

Um ein CARMA-Regelset zu erstellen, müssen Sie ein ID-Feld und mindestens ein Inhaltsfeld angeben. Das ID-Feld kann eine beliebige Rolle oder ein beliebiges Messniveau aufweisen. Felder mit der Rolle *Keine* werden ignoriert. Feldtypen müssen vollständig instanziiert werden, bevor der Knoten ausgeführt wird. Wie bei Apriori können die Daten als Tabellen- oder als Transaktionsdaten vorliegen. Weitere Informationen finden Sie im Thema „Tabellendaten im Vergleich zu Transaktionsdaten" auf Seite 232.

Stärken. Der CARMA-Knoten basiert auf dem CARMA-Assoziationsregelalgorithmus. Im Gegensatz zu Apriori bietet der CARMA-Knoten Erstellungseinstellungen für die Regelunterstützung (Unterstützung für Antezedens und Sukzedens) und nicht für die Antezedens-Unterstützung. CARMA erlaubt auch Regeln mit mehreren Sukzedenzien. Wie bei Apriori können von einem CARMA-Knoten generierte Modelle zum Erstellen von Vorhersagen in einen Datenstream eingefügt werden. Weitere Informationen finden Sie im Thema „Modellnuggets" auf Seite 37.

Feldoptionen für den CARMA-Knoten

Vor der Ausführung eines CARMA-Knotens müssen Sie Eingabefelder auf der Registerkarte "Felder" des CARMA-Knotens angeben. Während die meisten Modellierungsknoten dieselben Optionen auf der Registerkarte "Felder" aufweisen, enthält der CARMA-Knoten mehrere einzigartige Optionen. Sämtliche Optionen werden nachfolgend beschrieben.

Typknoteneinstellungen verwenden. Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

Benutzerdefinierte Einstellungen verwenden. Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option die Felder unten danach an, ob die Daten im Transaktions- oder im Tabellenformat gelesen werden.

Transaktionsformat verwenden. Mit dieser Option werden die Feldsteuerelemente im übrigen Dialogfeld danach geändert, ob die Daten im Transaktions- oder im Tabellenformat vorliegen. Wenn Sie mehrere Felder für Transaktionsdaten verwenden, stellen die in diesen Feldern für einen bestimmten Datensatz angegebenen Elemente solche Elemente dar, die in einer Einzeltransaktion mit einer einzelnen Zeitmarke gefunden wurden. Weitere Informationen finden Sie im Thema „Tabellendaten im Vergleich zu Transaktionsdaten" auf Seite 232.

Tabellendaten

Wenn **Transaktionsformat verwenden** nicht ausgewählt ist, werden die folgenden Felder angezeigt.

- **Eingaben.** Wählen Sie das/die Eingabefeld(er) aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.
- **Partition.** In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer anderen Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln. Wenn mehrere Partitionsfelder mithilfe von Typ- oder Partitionsknoten erstellt wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.) Beachten Sie au-

ßerdem, dass die Partitionierung auch auf der Registerkarte "Modelloptionen" des Knotens aktiviert sein muss, wenn die ausgewählte Partitionierung in Ihrer Analyse angewendet werden soll. (Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung ohne Änderung der Feldeinstellungen inaktiviert werden.)

Transaktionsdaten

Wenn Sie **Transaktionsformat verwenden** auswählen, werden die folgenden Felder angezeigt.

- **ID.** Wählen Sie für Transaktionsdaten ein ID-Feld aus der Liste aus. Als ID-Feld können numerische oder symbolische Felder verwendet werden. Jeder eindeutige Wert in diesem Feld sollte eine bestimmte Analyseeinheit darstellen. Bei einer Warenkorbanwendung könnte z. B. jede ID einen einzelnen Kunden darstellen. Für eine Webprotokoll-Analyseanwendung könnte jede ID einen Computer (nach IP-Adresse) oder einen Benutzer (nach Anmeldedaten) darstellen.
- **IDs sind zusammenhängend.** (Nur Apriori- und CARMA-Knoten) Wenn Ihre Daten vorsortiert sind, sodass alle Datensätze mit derselben ID im Datenstream zusammengefasst sind, wählen Sie diese Option, um die Verarbeitung zu beschleunigen. Wenn Ihre Daten nicht vorsortiert sind (oder Sie nicht sicher sind), lassen Sie diese Option inaktiviert. Die Daten werden dann vom Knoten automatisch sortiert.

  Hinweis: Wenn Ihre Daten nicht sortiert sind und Sie diese Option auswählen, erhalten Sie möglicherweise ungültige Ergebnisse in Ihrem Modell.
- **Inhalt.** Geben Sie die Inhaltsfelder für das Modell an. Diese Felder enthalten die Elemente, die für die Assoziationsmodellierung interessant sind. Sie können mehrere Flagfelder angeben (falls die Daten in tabellarischer Form vorliegen) oder ein einzeln nominales Feld (falls die Daten im Transaktionsformat vorliegen).

Modelloptionen für den CARMA-Knoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Minimale Regelunterstützung (%). Sie können auch ein Stützkriterium angeben. **Regelunterstützung** bezieht sich auf den Anteil von IDs in den Trainingsdaten, die die gesamte Regel enthalten. (Beachten Sie, dass diese Definition von der für Apriori-Knoten verwendeten Antezedens-Unterstützung abweicht.) Wenn Sie weitere gemeinsame Regeln wünschen, erhöhen Sie diese Einstellung.

Minimale Regelkonfidenz (%). Sie können ein Konfidenzkriterium angeben, um die Regeln im Regelset beizubehalten. **Konfidenz** bezieht sich auf den prozentualen Anteil von IDs, für die eine richtige Vorhersage gemacht wird (aus allen IDs, für die die Regel eine Vorhersage macht). Sie wird aus der Anzahl von IDs berechnet, für die die gesamte Regel gefunden wird, dividiert durch die Anzahl der IDs, für die die Vorgänger gefunden werden, basierend auf den Trainingsdaten. Regeln mit einer niedrigeren Konfidenz als das angegebene Kriterium werden verworfen. Wenn Sie uninteressante oder zu viele Regeln erhalten, sollten Sie diese Einstellung erhöhen. Wenn Sie zu wenige Regeln erhalten, sollten Sie diese Einstellung reduzieren.

Maximale Regelgröße. Sie können die maximale Anzahl unterschiedlicher *Elementsets* (im Gegensatz zu *Elementen*) in einer Regel festlegen. Wenn die gewünschten Regeln relativ kurz sind, können Sie diese Einstellung reduzieren, um die Erstellung des Regelsets zu beschleunigen.

Expertenoptionen für den CARMA-Knoten

Für Personen mit umfassenden Kenntnissen über die Operation des CARMA-Knotens ermöglichen die folgenden Expertenoptionen die Feinabstimmung des Modellierungsvorgangs. Um auf die Expertenoptionen zuzugreifen, legen Sie auf der Registerkarte "Experten" den Modus **Experten** fest.

Regeln mit mehreren Sukzedenzien ausschließen. Wählen Sie diese Option aus, um "doppelköpfige" Sukzendenzien auszuschließen, d. h. Sukzendenzien mit zwei Elementen. Zum Beispiel enthält die Regel `Brot & Käse & Fisch -> Wein&Obst` ein doppelköpfiges Sukzedens, `Wein&Obst`. Standardmäßig werden solche Regeln eingeschlossen.

Reduzierungswert festlegen. Um Speicherplatz zu sparen, entfernt (**reduziert**) der verwendete CARMA-Algorithmus in regelmäßigen Abständen während der Verarbeitung seltene Elementsets aus seiner Liste möglicher Elemente. Wählen Sie diese Option, um die Reduzierungshäufigkeit anzupassen. Die von Ihnen angegebene Zahl ermittelt die Reduktionshäufigkeit. Geben Sie einen kleinen Wert ein, um die Speicheranforderungen des Algorithmus zu reduzieren (möglicherweise aber die erforderliche Trainingszeit zu erhöhen), oder geben Sie einen hohen Wert ein, um die Trainingsgeschwindigkeit zu erhöhen (möglicherweise aber die Speicheranforderungen zu erhöhen). Der Standardwert ist 500.

Unterstützung variieren. Wählen Sie diese Option aus, um die Effizienz zu erhöhen, indem Sie seltene Elementsets ausführen, die den Eindruck erwecken, als wären sie häufig, wenn Sie ungleich verteilt vorkommen. Dies erreichen Sie durch ein höheres Unterstützungsniveau und durch die Reduktion auf den auf der Registerkarte "Modell" angegebenen Wert. Geben Sie einen Wert für **Geschätzte Anzahl an Transaktionen** ein, um anzugeben, wie schnell das Unterstützungsniveau reduziert werden soll.

Regeln ohne Antezedenzien zulassen. Wählen Sie diese Option aus, um Regeln zuzulassen, die lediglich das Sukzedens (Element oder Elementset) enthalten. Dies ist sinnvoll, wenn Sie an der Ermittlung häufig verwendeter Elemente oder Elementsets interessiert sind. Zum Beispiel ist `Konservengemüse` eine Regel mit nur einem Element ohne Antezedens, die angibt, dass der Kauf von *Konservengemüse* in den Daten häufig vorkommt. In manchen Fällen können Sie derartige Regeln aufnehmen, wenn Sie nur an den wahrscheinlichsten Vorhersagen interessiert sind. Diese Option ist standardmäßig inaktiviert.

Assoziationsregelmodellnuggets

Assoziationsregelmodellnuggets enthalten die Regeln, die von einem der folgenden Modellierungsknoten für Assoziationsregeln entdeckt wurden:

- Apriori
- CARMA

Die Modellnuggets enthalten Informationen zu den Regeln, die bei der Modellerstellung aus den Daten extrahiert wurden.

Anzeigen von Ergebnissen

Sie können die von den Assoziationsmodellen (Apriori und CARMA) und den Sequenzmodellen generierten Regeln mithilfe der Registerkarte "Modell" im Dialogfeld durchsuchen. Beim Durchsuchen eines Modellnuggets erhalten Sie Informationen zu den Regeln sowie Optionen zum Filtern und Sortieren der Ergebnisse vor der Erstellung neuer Knoten oder der Bewertung des Modells.

Bewerten des Modells

Verfeinerte Modellnuggets (Apriori, CARMA und Sequenz) können zum Stream hinzugefügt und für die Bewertung verwendet werden. Weitere Informationen finden Sie im Thema „Verwendung von Modellnuggets in Streams” auf Seite 48. Für Scoring verwendete Modelnuggets enthalten eine zusätzliche Registerkarte "Einstellungen" in den entsprechenden Dialogfeldern. Weitere Informationen finden Sie im Thema „Einstellungen beim Assoziationsregelmodellnugget” auf Seite 242.

Ein nicht verfeinertes Modellnugget kann nicht im Rohformat für die Bewertung verwendet werden. Stattdessen können Sie ein Regelset erstellen und für die Bewertung verwenden. Weitere Informationen finden Sie im Thema „Generieren eines Regelsets aus einem Assoziationsmodellnugget” auf Seite 243.

Nähere Informationen zum Assoziationsregelmodellnugget

Auf der Registerkarte "Modell" eines Assoziationsregelmodellnuggets wird eine Tabelle angezeigt, die die vom Algorithmus extrahierten Regeln enthält. Jede Zeile in der Tabelle steht für eine Regel. Die erste Spalte enthält die Sukzedenzien (den "dann"-Teil der Regel), während die nächste Spalte die Antezedenzien (den "wenn"-Teil der Regel) enthält. Die weiteren Spalten enthalten Informationen zur Regel, wie Konfidenz, Unterstützung und Lift.

Assoziationsregeln werden häufig in einem Format wie in der folgenden Tabelle angezeigt.

Tabelle 13. Beispiel einer Assoziationsregel

Sukzedens	Antezedent
`Medikament = MedikamentY`	`Geschlecht = W` `BD = HOCH`

Die Beispielregel wird wie folgt interpretiert: *Wenn Geschlecht = "W" und BD = "HOCH", ist das Medikament wahrscheinlich MedikamentY*, oder anders ausgedrückt: *Bei Datensätzen mit Geschlecht = "W" und BD = "HOCH" ist das Medikament wahrscheinlich MedikamentY*. Mit der Symbolleiste des Dialogfelds können Sie weitere Informationen anzeigen, beispielsweise Konfidenz, Unterstützung und Instanzen.

Menü "Sortieren". Mit der Schaltfläche des Menüs "Sortieren" in der Symbolleiste wird die Sortierung der Regeln gesteuert. Die Sortierrichtung (aufsteigend oder absteigend) lässt sich mit der Schaltfläche für die Sortierrichtung (nach unten bzw. oben zeigender Pfeil) ändern.

Regeln können nach folgenden Faktoren sortiert werden:

- Unterstützung
- Konfidenz
- Regelunterstützung
- Sukzedens
- Lift
- Bereitstellbarkeit

Menü "Anzeigen/Ausblenden". Das Menü "Anzeigen/Ausblenden" (Symbolleistenschaltfläche "Kriterien") steuert die Optionen für die Anzeige der Regeln.

Abbildung 46. Schaltfläche "Anzeigen/Ausblenden"

Die folgenden Anzeigeoptionen sind verfügbar:

- **Regel-ID** zeigt die während der Modellerstellung zugewiesene Regel-ID an. Mit einer Regel-ID können Sie identifizieren, welche Regeln für eine bestimmte Vorhersage angewendet werden. Mit Regel-IDs können Sie außerdem zusätzliche Regelinformationen zu einem späteren Zeitpunkt zusammenführen, beispielsweise Verwendbarkeit, Produktinformationen oder Antezedenzien.
- **Instanzen** zeigt Informationen über die Anzahl der eindeutigen IDs an, auf die die Regel zutrifft, d. h., für die die Antezedenzien wahr sind. Bei der Regel `Brot -> Käse` beispielsweise wird die Anzahl der Datensätze in den Trainingsdaten, die das Antezedens *Brot* enthalten, als **Instanzen** bezeichnet.
- **Unterstützung** zeigt die Antezedens-Unterstützung basierend auf den Trainingsdaten an, also den Teil der IDs, für den die Antezedenzien wahr sind. Beispiel: Wenn 50 % der Trainingsdaten den Kauf von Brot beinhalten, hat die Regel `Brot -> Käse` eine Antezedens-Unterstützung von 50 %. *Hinweis*: Unterstützung entspricht in diesem Kontext den Instanzen, wird jedoch als Prozentsatz dargestellt.

- **Konfidenz** zeigt das Verhältnis von Regelunterstützung zu Antezedens-Unterstützung an. Es wird also der Anteil der IDs mit den angegebenen Antezedenzien angegeben, bei denen auch das Sukzedens (bzw. die Sukzedenzien) wahr ist. Wenn beispielsweise 50 % der Trainingsdaten "Brot" enthalten (Antezedens-Unterstützung), aber nur 20 % sowohl Brot als auch Käse enthalten (Regelunterstützung), wäre die Konfidenz für die Regel `Brot -> Käse` `Regelunterstützung / Antezedens-Unterstützung`, in diesem Fall 40 %.
- **Regelunterstützung** zeigt den Teil der IDs an, für die die gesamte Regel, die Antezedenzien und das Sukzedens (bzw. die Sukzedenzien) wahr sind. Wenn beispielsweise 20 % der Trainingsdaten sowohl Brot als auch Käse enthalten, beträgt die Regelunterstützung für die Regel `Brot -> Käse` 20 %.
- **Lift** zeigt das Verhältnis zwischen der Konfidenz für die Regel und der A-priori-Wahrscheinlichkeit für das Sukzedens an. Beispiel: Wenn 10 % der Gesamtbevölkerung Brot kauft, weist eine Regel, die mit einer Konfidenz von 20 % vorhersagt, ob die Leute Brot kaufen, einen Lift von 20/10 = 2 auf. Wenn eine andere Regel mit einer Konfidenz von 11 % besagt, dass die Leute Brot kaufen, weist diese Regel einen Lift von annähernd 1 auf, was bedeutet, dass die Antezedenzien keinen großen Unterschied für die Wahrscheinlichkeit des Sukzedens aufweisen. Im Allgemeinen sind Regeln mit einem Lift, der sich von 1 unterscheidet, interessanter als Regeln mit einem Lift von annähernd 1.
- **Verwendbarkeit** ist ein Maß dafür, welcher Prozentsatz der Trainingsdaten die Bedingungen des Antezedens erfüllen, nicht jedoch die Bedingungen des Sukzedens. Beim Einkauf von Produkten bedeutet dies im Grunde, welcher Prozentsatz des Kundenstamms das Produkt aus dem Antezedens besitzt (bzw. erworben hat), jedoch noch nicht das im Sukzedens festgelegte Produkt gekauft hat. Die Verwendbarkeitsstatistik ist definiert als `([Antezedens-Unterstützung in Anzahl der Datensätze - Regelunterstützung in Anzahl der Datensätze] / Anzahl der Datensätze) * 100`. Dabei ist *Antezedens-Unterstützung* die Anzahl der Datensätze, bei denen die Antezedenzien wahr sind, und *Regelunterstützung* die Anzahl der Datensätze, bei denen sowohl die Antezedenzien als auch das Sukzedens wahr sind.

Schaltfläche "Filter". Mit der Schaltfläche "Filter" (Trichtersymbol) im Menü wird der untere Teil des Dialogfelds erweitert und ein Fenster mit aktiven Regelfiltern wird angezeigt. Filter werden verwendet, um die Anzahl der auf der Registerkarte "Modelle" angezeigten Regeln einzugrenzen.

Abbildung 47. Schaltfläche "Filter"

Zum Erstellen von Filtern klicken Sie auf das Filtersymbol rechts neben dem erweiterten Fenster. Dadurch wird ein separates Dialogfeld geöffnet, in dem Sie Bedingungen für die Anzeige von Regeln eingeben können. Beachten Sie, dass die Filterschaltfläche häufig in Verbindung mit dem Menü "Generieren" verwendet wird, um zunächst die Regeln zu filtern und anschließend ein Modell zu erstellen, das das betreffende Subset der Regeln enthält. Weitere Informationen finden Sie unter „Angeben von Filtern für Regeln” auf Seite 241.

Schaltfläche "Regel suchen". Die Schaltfläche "Regel suchen" (Fernglassymbol) ermöglicht das Durchsuchen der angezeigten Regeln nach einer angegebenen Regel-ID. Im angrenzenden Dialogfeld wird angegeben, wie viele der verfügbaren Regeln derzeit angezeigt werden. Regel-IDs werden vom Modell in der Reihenfolge ihrer Entdeckung zugewiesen und werden während der Bewertung zu den Daten hinzugefügt.

Abbildung 48. Schaltfläche "Regel suchen"

So können Sie Regel-IDs neu ordnen:

1. Sie können die Regel-IDs in IBM SPSS Modeler neu anordnen, indem Sie zuerst die Regelanzeigetabelle gemäß dem gewünschten Maß sortieren, beispielsweise "Konfidenz" oder "Lift".
2. Anschließend erstellen Sie mit den Optionen aus dem Menü "Generieren" ein gefiltertes Modell.
3. Wählen Sie im Dialogfeld "Gefiltertes Modell" die Option **Regeln neu nummerieren, beginnend mit** und geben Sie eine Startnummer an.

Weitere Informationen finden Sie im Thema „Erstellen eines gefilterten Modells” auf Seite 244.

Angeben von Filtern für Regeln

Standardmäßig können Regelalgorithmen wie "Apriori", "CARMA" und "Sequenz" sehr große und umständlich zu handhabende Mengen von Regeln generieren. Zugunsten einer größeren Klarheit beim Durchsuchen bzw. zur Rationalisierung der Regelbewertung sollten Sie in Erwägung ziehen, die Regeln zu filtern, sodass die für Sie relevanten Sukzedenzien und Antezedenzien deutlicher zu sehen sind. Mithilfe der Filteroptionen auf der Registerkarte "Modell" eines Regelbrowsers können Sie ein Dialogfeld zur Angabe der Filterbedingungen öffnen.

Sukzedenzien. Mit der Option **Filter aktivieren** können Sie Optionen für das Filtern von Regeln definieren, die auf der Aufnahme bzw. dem Ausschluss angegebener Sukzedenzien beruhen. Wählen Sie **Mindestens eines einschließen aus**, um einen Filter zu erstellen, bei dem die Regeln mindestens einen der angegebenen Sukzedenzien enthalten. Wählen Sie alternativ **Ausschließen**, um einen Filter zu erstellen, der die angegebenen Sukzedenzien ausschließt. Sie können die Sukzedenzien mithilfe des Auswahlsymbols rechts neben dem Listenfeld auswählen. Dadurch wird ein Dialogfeld geöffnet, das alle Sukzedenzien auflistet, die in den generierten Regeln vorliegen.

Hinweis: Sukzedenzien können mehrere Elemente enthalten. Mit den Filtern wird nur überprüft, ob ein Sukzedens eines der angegebenen Elemente enthält.

Antezedenzien. Mit der Option **Filter aktivieren** können Sie Optionen für das Filtern von Regeln definieren, die auf der Aufnahme bzw. dem Ausschluss angegebener Antezedenzien beruhen. Sie können die gewünschten Elemente mithilfe des Auswahlsymbols rechts neben dem Listenfeld auswählen. Dadurch wird ein Dialogfeld geöffnet, das alle Antezedenzien auflistet, die in den generierten Regeln vorliegen.

- Wählen Sie die Option **Alle einschließen aus**, um den Filter als Einschlussfilter festzulegen, bei dem alle angegebenen Antezedenzien in einer Regel enthalten sein müssen.
- Wählen Sie **Mindestens eines einschließen aus**, um einen Filter zu erstellen, bei dem die Regeln mindestens einen der angegebenen Antezedenzien enthalten.
- Wählen Sie **Ausschließen** aus, um einen Filter zu erstellen, der Regeln ausschließt, die ein angegebenes Antezedens enthalten.

Konfidenz. Mit der Option **Filter aktivieren** können Sie Optionen für das Filtern von Regeln definieren, die auf dem Konfidenzniveau der jeweiligen Regel beruhen. Mit den **Min**- und **Max**-Steuerelementen können Sie einen Konfidenzbereich angeben. Beim Durchsuchen der generierten Modelle wird die Konfidenz als Prozentsatz angegeben. Bei der Bewertung von Ausgaben wird die Konfidenz als Zahl zwischen 0 und 1 angegeben.

Antezedens-Unterstützung. Mit der Option **Filter aktivieren** können Sie Optionen für das Filtern von Regeln definieren, die auf dem Niveau der Antezedens-Unterstützung für die jeweilige Regel beruhen. Die Antezedens-Unterstützung gibt den Anteil der Trainingsdaten an, die dieselben Antezedenzien wie die aktuelle Regel beinhalten, ähnlich einem Popularitätsindex. Mit den **Min**- und **Max**-Steuerelementen können Sie einen Bereich angeben, der zum Filtern der Regel anhand des Unterstützungsniveaus verwendet wird.

Lift. Mit der Option **Filter aktivieren** können Sie Optionen für das Filtern von Regeln definieren, die auf dem Liftmaß für die jeweilige Regel beruhen. *Hinweis*: Die Liftfilterung ist nur für Assoziationsmodelle verfügbar, die nach Release 8.5 erstellt wurden, oder für frühere Modelle, die eine Liftmessung enthalten. Bei Sequenzmodellen ist diese Option nicht verfügbar.

Klicken Sie auf **OK**, um alle Filter anzuwenden, die in diesem Dialogfeld aktiviert wurden.

Erzeugen von Diagrammen für Regeln

Die Assoziationsknoten bieten eine Menge an Informationen, jedoch sind diese nicht unbedingt immer in einem leicht zugänglichen Format für Fachanwender. Zur Bereitstellung der Daten auf eine Art, die problemlos in Geschäftsberichte, Präsentationen u.s.w. integriert werden kann, können aus ausgewählten Daten Diagramme erstellt werden. auf der Registerkarte "Modell" können Sie ein Diagramm für eine ausgewählte Regel erzeugen und damit ein Diagramm nur für die Fälle dieser Regel erstellen.

1. Wählen Sie auf der Registerkarte "Modell" die Regel aus, die Sie interessiert.
2. Wählen Sie im Menü "Generieren" den Befehl **Diagramm (von Auswahl)**. Die Registerkarte "Einfach" der Diagrammtafel wird angezeigt.

   Hinweis: Wenn Sie die Diagrammtafel auf diese Art anzeigen, sind nur die Registerkarten "Basis" und "Details" verfügbar.
3. Mithilfe der Einstellungen auf den Registerkarten "Basis" oder "Details" können Sie die Details angeben, die auf dem Diagramm angezeigt werden sollen.
4. Klicken Sie auf "OK", um das Diagram zu erstellen.

Die Überschrift des Diagramms identifiziert die Regel und Antezedensdetails, die im Diagramm berücksichtigt werden.

Einstellungen beim Assoziationsregelmodellnugget

Die Registerkarte "Einstellungen" wird zur Angabe der Scoring-Optionen für Zuordnungsmodelle ("Apriori" und "CARMA") verwendet. Diese Registerkarte ist erst dann verfügbar, wenn das Modellnugget zum Zwecke des Scorings zu einem Stream hinzugefügt wurde.

Hinweis: Das Dialogfeld zum Durchsuchen eines nicht verfeinerten Modells enthält nicht die Registerkarte "Einstellungen", da es nicht gescort werden kann. Um das "nicht verfeinerte" Modell zu scoren, müssen Sie zunächst ein Regelset generieren. Weitere Informationen finden Sie im Thema „Generieren eines Regelsets aus einem Assoziationsmodellnugget” auf Seite 243.

Maximale Anzahl an Vorhersagen. Dient zur Angabe der maximalen Anzahl an Vorhersagen, die für jedes Set von Warenkorbelementen aufgenommen werden. Diese Option wird zusammen mit "Regelkriterium" (unten) verwendet, um die "Top" Vorhersagen zu erzeugen. Dabei steht *Top*, wie unten angegeben, für das höchste Niveau an Konfidenz, Lift usw.

Regelkriterium. Dient zur Auswahl des Maßes, das zur Ermittlung der Stärke der Regeln verwendet wird. Die Regeln werden nach der Stärke der hier ausgewählten Kriterien sortiert, um die Top-Vorhersagen für ein Elementset auszugeben. Folgende Kriterien sind verfügbar:

- Konfidenz
- Unterstützung
- Regelunterstützung (Unterstützung * Konfidenz)
- Lift
- Bereitstellbarkeit

Wiederholungsvorhersagen zulassen. Wählen Sie diese Option aus, um mehrere Regeln mit demselben Sukzedens beim Scoring aufzunehmen. Bei Auswahl dieser Option können beispielsweise folgende Regeln gescort werden:

```
Brot & Käse -> Wein
Käse & Obst -> Wein
```

Inaktivieren Sie diese Option, um beim Scoring Wiederholungsvorhersagen auszuschließen.

Hinweis: Regeln mit mehreren Sukzedenzien (`Brot & Käse & Obst -> Wein & Pastete`) werden nur als Wiederholungsvorhersagen betrachtet, wenn alle Sukzedenzien (`Wein & Pastete`) zuvor vorhergesagt wurden.

Nicht zugeordnete Warenkorbelemente ignorieren. Wählen Sie diese Option, um das Vorliegen zusätzlicher Elemente im Elementset zu ignorieren. Wenn diese Option beispielsweise für einen Warenkorb ausgewählt wurde, der `[Zelt & Schlafsack & Wasserkessel]` enthält, gilt die Regel `Zelt & Schlafsack -> Gaskocher` trotz des zusätzlichen Elements (`Wasserkessel`) im Warenkorb.

Es kann Umstände geben, unter denen zusätzliche Elemente ausgeschlossen werden sollten. So ist es beispielsweise wahrscheinlich, dass jemand, der ein Zelt, einen Schlafsack und einen Wasserkessel kauft, bereits einen Gaskocher besitzt, worauf der Wasserkessel hindeutet. Anders ausgedrückt: Ein Gaskocher ist möglicherweise nicht die beste Vorhersage. In solchen Fällen sollten Sie die Auswahl von **Nicht zugeordnete Warenkorbelemente ignorieren** aufheben, um sicherzustellen, dass die Antezedenzien der Regel genau mit dem Inhalt eines Warenkorbs übereinstimmen. Standardmäßig werden nicht übereinstimmende Elemente ignoriert.

Sicherstellen, dass sich keine Vorhersagen im Warenkorb befinden. Wählen Sie diese Option aus, um sicherzustellen, dass die Elemente aus den Sukzedenzien nicht ebenfalls im Warenkorb vorhanden sind. Beispiel: Wenn das Ziel des Scorings darin besteht, eine Produktempfehlung für Möbel abzugeben, ist es unwahrscheinlich, dass bei einem Warenkorb, der bereits einen Esszimmertisch enthält, ein weiterer erworben wird. In solchen Fällen, sollten Sie diese Option auswählen. Andererseits können bei verderblichen Produkten und Einwegartikeln (wie Käse, Säuglingsnahrung oder Papiertaschentüchern) Regeln, bei denen das Sukzedens bereits im Warenkorb vorhanden ist, sinnvoll sein. Im letzteren Fall ist die nützlichste Option möglicherweise **Warenkorb nicht auf Vorhersagen prüfen** (siehe unten).

Sicherstellen, dass sich Vorhersagen im Warenkorb befinden. Wählen Sie diese Option aus, um sicherzustellen, dass die Elemente aus den Sukzedenzien auch im Warenkorb vorhanden sind. Dieser Ansatz ist sinnvoll, wenn Sie versuchen, einen Einblick in bestehende Kunden oder Transaktionen zu gewinnen. Sie könnten beispielsweise die Regeln mit dem höchsten Lift ermitteln und dann untersuchen, auf welche Kunden diese Regeln zutreffen.

Warenkorb nicht auf Vorhersagen prüfen. Wählen Sie diese Option aus, um beim Scoring alle Regeln einzuschließen, unabhängig vom Vorhandensein oder Nichtvorhandensein der Sukzedenzien im Warenkorb.

Übersicht über das Assoziationsregelmodellnugget

Auf der Registerkarte "Übersicht" für ein Sequenzregelmodellnugget werden die Anzahl der ermittelten Regeln sowie die Mindest- und Höchstwerte für Unterstützung, Lift, Konfidenz und Bereitstellbarkeit für die Regeln angegeben.

Generieren eines Regelsets aus einem Assoziationsmodellnugget

Assoziationsmodellnuggets wie "Apriori" oder "CARMA" können zur direkten Speicherung von Daten verwendet werden. Alternativ können Sie zunächst ein Subset von Regeln erstellen, ein sogenanntes **Regelset**. Regelsets sind besonders nützlich bei der Arbeit mit einem nicht verfeinerten Modell, das nicht direkt zum Scoren verwendet werden kann. Weitere Informationen finden Sie im Thema „Nicht verfeinerte Modelle” auf Seite 52.

Um ein Regelset zu generieren, wählen Sie die Option **Regelset** aus dem Menü "Generieren" im Browser für Modellnuggets aus. Folgende Optionen für die Übersetzung der Regeln in ein Regelset können angegeben werden:

Regelsatzname. Ermöglicht die Angabe eines Namens für den neu generierten Regelsetknoten.

Knoten erstellen auf. Steuert den Standort des neu generierten Regelsetknotens. Wählen Sie **Erstellungsbereich**, **Generierte Modelle** oder **Beides** aus.

Zielfeld. Legt fest, welches Ausgabefeld für den generierten Regelsetknoten verwendet wird. Wählen Sie ein einzelnes Ausgabefeld in der Liste aus.

Minimale Unterstützung. Geben Sie die minimale Unterstützung für Regeln an, die im generierten Regelset erhalten bleiben sollen. Regeln, deren Unterstützung unter dem angegebenen Wert liegt, werden nicht in das neue Regelset aufgenommen.

Minimale Konfidenz. Geben Sie die minimale Konfidenz für Regeln an, die im generierten Regelset erhalten bleiben sollen. Regeln, deren Konfidenz unter dem angegebenen Wert liegt, werden nicht in das neue Regelset aufgenommen.

Standardwert. Ermöglicht die Angabe eines Standardwerts für das Zielfeld, der gescorten Datensätzen zugewiesen wird, auf die keine Regel zutrifft.

Erstellen eines gefilterten Modells

Um ein gefiltertes Modell aus einem Assoziationsmodellnugget wie einem Regelsetknoten vom Typ "Apriori", "CARMA" oder "Sequenz" zu erstellen, wählen Sie im Browser für Modellnuggets im Menü "Generieren" die Option **Gefiltertes Modell**. Dadurch wird ein Subsetmodell erstellt, das nur die Regeln enthält, die derzeit im Browser angezeigt werden. *Hinweis*: Sie können keine gefilterten Modelle für nicht verfeinerte Modelle erstellen.

Folgende Optionen sind für das Filtern von Regeln verfügbar:

Name für neues Modell. Ermöglicht die Angabe eines Namens des neuen Knotens vom Typ "Gefiltertes Modell".

Knoten erstellen auf. Steuert den Standort des neuen Knotens vom Typ "Gefiltertes Modell". Wählen Sie **Erstellungsbereich**, **Generierte Modelle** oder **Beides** aus.

Regelnummerierung. Dient zur Angabe, wie die Regel-IDs im Regelsubset, das in das gefilterte Modell eingeschlossenen ist, nummeriert werden sollen.

- **Ursprüngliche Regel-ID-Nummern beibehalten.** Wählen Sie diese Option aus, um die ursprüngliche Nummerierung der Regeln beizubehalten. Standardmäßig erhalten die Regeln eine ID, die der Reihenfolge Ihrer Entdeckung durch den Algorithmus entspricht. Diese Reihenfolge kann je nach dem verwendeten Algorithmus variieren.
- **Regeln neu nummerieren, beginnend mit.** Wählen Sie diese Option, um den gefilterten Regeln neue Regel-IDs zuzuweisen. Neue IDs werden auf der Grundlage der in der Regelbrowsertabelle auf der Registerkarte "Modell" angezeigten Sortierreihenfolge zugewiesen. Die Nummerierung beginnt mit der Zahl, die Sie hier angeben. Sie können die Startnummer für IDs mithilfe der Pfeile auf der rechten Seite angeben.

Scoren von Assoziationsregeln

Die Scores, die erzielt werden, wenn neue Daten ein Assoziationsregelmodellnugget durchlaufen, werden in separaten Feldern ausgegeben. Für jede Vorhersage werden drei neue Felder hinzugefügt. Dabei steht *P* für die Vorhersage, *C* für die Konfidenz und *I* für die Regel-ID. Die Organisation dieser Ausgabefelder ist abhängig davon, ob die Ausgabedaten im Transaktions- oder im Tabellenformat vorliegen. Einen Überblick über diese Formate finden Sie unter „Tabellendaten im Vergleich zu Transaktionsdaten” auf Seite 232.

Angenommen, Sie nehmen das Scoring von Warenkorbdaten mithilfe eines Modells vor, bei dem Vorhersagen auf der Grundlage der folgenden drei Regeln erzeugt werden:

```
Regel_15  Brot&Wein -> Fleisch (Konfidenz 54%)
Regel_22  Käse -> Obst (Konfidenz 43%)
Regel_5  Brot&Käse -> TK-Gemüse (Konfidenz 24 %)
```

Tabellendaten. Bei Tabellendaten werden die drei Vorhersagen (3 ist der Standardwert) in einem einzelnen Datensatz ausgegeben.

Tabelle 14. Scores im Tabellenformat.

ID	Brot	Wein	Käse	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	E	E	E	Fleisch	0,54	15	Obst	0,43	22	TK-Gemüse	0,24	5

Transaktionsdaten. Bei Transaktionsdaten wird je ein separater Datensatz für die einzelnen Vorhersagen erzeugt. Die Vorhersagen werden ebenfalls in separaten Spalten hinzugefügt, die Scores werden jedoch so ausgegeben, wie sie berechnet werden. Dies führt zu Datensätzen mit unvollständigen Vorhersagen (vgl. das nachstehende Ausgabebeispiel). Die zweite und dritte Vorhersage (P2 und P3) im ersten Datensatz sind leer, ebenso wie die zugehörigen Konfidenzen und Regel-IDs. Wenn die Scores ausgegeben werden, enthält der endgültige Datensatz jedoch alle drei Vorhersagen.

Tabelle 15. Scores im Transaktionsformat.

ID	Item	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	Brot	Fleisch	0,54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	Käse	Fleisch	0,54	14	Obst	0,43	22	$null$	$null$	$null$
Fred	Wein	Fleisch	0,54	14	Obst	0,43	22	TK-Gemüse	0,24	5

Um für die Berichterstellung oder Bereitstellung nur vollständige Vorhersagen aufzunehmen, wählen Sie die vollständigen Datensätze mithilfe eines Auswahlknotens aus.

Hinweis: Die in diesen Beispielen verwendeten Feldnamen sind zur Vereinfachung abgekürzt. Während der tatsächlichen Verwendung werden Ergebnisfelder für Assoziationsmodelle wie in der folgenden Tabelle benannt.

Tabelle 16. Namen der Ergebnisfelder für Assoziationsmodelle.

Neues Feld	Beispiel für Feldnamen
Vorhersage	*$A-TRANSAKTIONSNUMMER-1*
Konfidenz (oder anderes Kriterium)	*$AC-TRANSAKTIONSNUMMER-1*
Regel-ID	*$A-Regel_ID-1*

Regeln mit mehreren Sukzedenzien

Der CARMA-Algorithmus lässt Regeln mit mehreren Sukzedenzien zu, beispielsweise:

```
Brot -> Wein&Käse
```

Beim Scoren solcher "doppelköpfiger" Regeln werden Vorhersagen in dem in der folgenden Tabelle angezeigten Format zurückgegeben.

Tabelle 17. Scoring-Ergebnisse einschließlich einer Vorhersage mit mehreren Sukzedenzien.

ID	Brot	Wein	Käse	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	E	E	E	Fleisch &Gemüse	0,54	16	Obst	0,43	22	TK-Gemüse	0,24	5

In einigen Fällen müssen Sie derartige Scores vor der Bereitstellung aufteilen. Zur Aufteilung einer Vorhersage mit mehreren Sukzedenzien müssen Sie das Feld mithilfe der CLEM-Zeichenfolgefunktionen analysieren.

Bereitstellung von Assoziationsmodellen

Beim Scoring von Assoziationsmodellen werden die Vorhersagen und die Konfidenzen in separaten Spalten ausgegeben. (*P* steht dabei für die Vorhersage, *K* für die Konfidenz und *I* für die Regel-ID.) Dies gilt für Eingabedaten sowohl im Tabellenformat als auch im Transaktionsformat. Weitere Informationen finden Sie im Thema „Scoren von Assoziationsregeln" auf Seite 244.

Bei der Vorbereitung von Scores für die Bereitstellung stellt sich gegebenenfalls heraus, dass die Ausgabedaten in ein Format transponiert werden müssen, bei dem die Vorhersagen nicht in Spalten ausgegeben werden, sondern in Zeilen (je eine Vorhersage pro Zeile, auch als "Kassenrollenformat" bezeichnet).

Transponieren von Scores im Tabellenformat

Sie können Scores im Tabellenformat mithilfe einer Reihe von Schritten in IBM SPSS Modeler von Spalten in Zeilen transponieren (siehe nachfolgende Schritte).

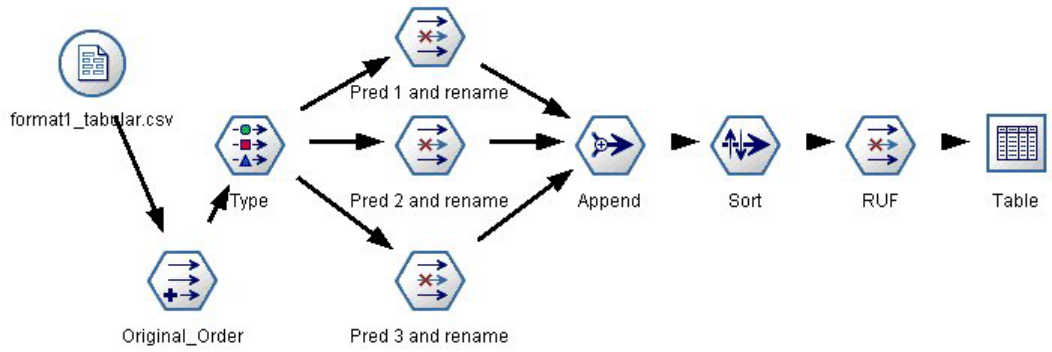


Abbildung 49. Beispielstream für die Transposition von Tabellendaten in das Kassenrollenformat

1. Überprüfen Sie die gegenwärtige Reihenfolge der Vorhersagen mithilfe der Funktion `@INDEX` in einem Ableitungsknoten und speichern Sie diesen Indikator in einem neuen Feld, wie beispielsweise *Original_order*.
2. Fügen Sie einen Typknoten hinzu, um sicherzustellen, dass alle Felder instanziiert sind.
3. Mit einem Filterknoten können Sie die Standardfelder für Vorhersage, Konfidenz und ID (*P1, C1, I1*) in gewöhnliche Felder umbenennen, wie beispielsweise *Prog*, *Krit* und *Regel-ID*, die später an die Datensätze angehängt werden. Für jede generierte Vorhersage wird jeweils ein Filterknoten benötigt.

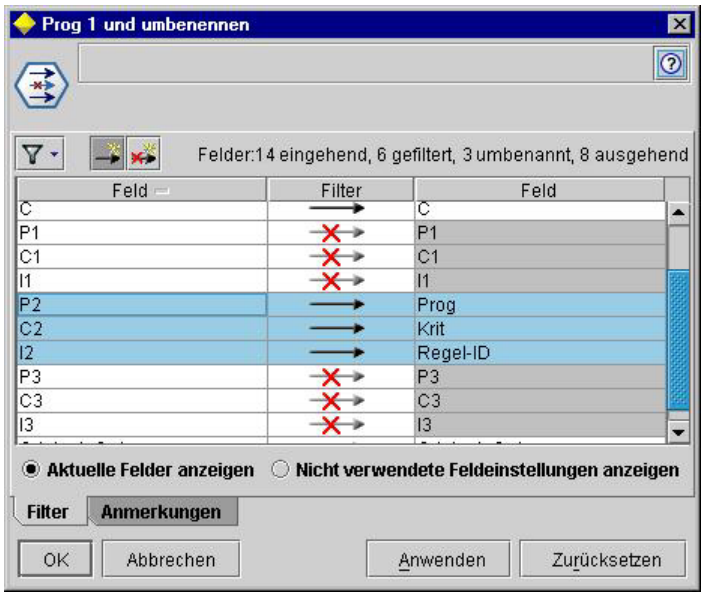


Abbildung 50. Filtern der Felder für die Vorhersagen 1 und 3 bei Umbenennung der Felder für Vorhersage 2.

4. Verwenden Sie einen Anhangknoten, um Scores für die freigegebenen Elemente *Prog*, *Krit* und *Regel-ID* anzuhängen.
5. Hängen Sie einen Sortierknoten an, um die Datensätze für das Feld *Original_order* in aufsteigender Reihenfolge und für *Crit* in absteigender Reihenfolge zu sortieren. Bei "Crit" handelt es sich um das Feld, das zum Sortieren der Vorhersagen nach Kriterien wie Konfidenz, Lift und Unterstützung verwendet wird.
6. Filtern Sie mithilfe eines weiteren Filterknotens das Feld *Original_order* aus der Ausgabe.

Nun können die Daten bereitgestellt werden.

Transponieren von Scores im Transaktionsformat

Für die Transposition von Transaktionsscores wird ein ähnlicher Prozess verwendet. Beispiel: Der unten dargestellte Stream transponiert Scores in ein Format, bei dem jede Zeile eine einzelne Vorhersage erhält, wie für die Bereitstellung erforderlich.

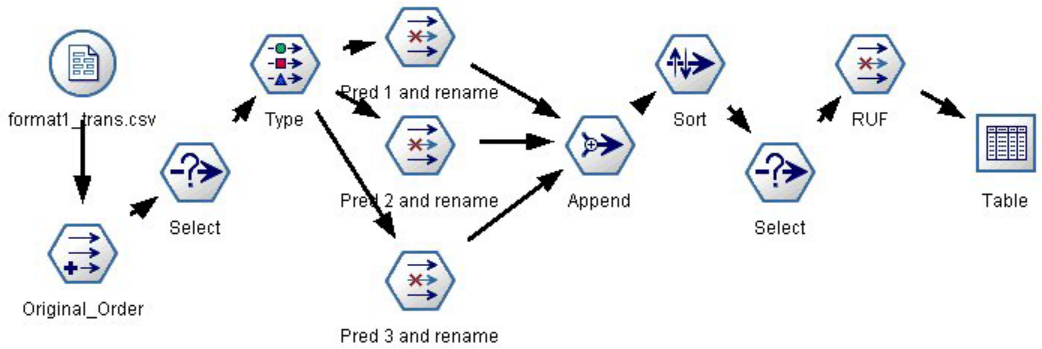


Abbildung 51. Beispielstream für die Transposition von Transaktionsdaten in das Kassenrollenformat

Abgesehen von zwei weiteren Auswahlknoten ist der Vorgang mit dem zuvor für Tabellendaten beschriebenen identisch.

- Der erste Auswahlknoten wird verwendet, um Regel-IDs in angrenzenden Datensätzen zu vergleichen und nur eindeutige nicht definierte Datensätze aufzunehmen. Dieser Auswahlknoten verwendet den CLEM-Ausdruck zum Auswählen von Datensätzen: `ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`.
- Der zweite Auswahlknoten wird verwendet, um überflüssige Regeln oder Regeln, bei denen `Regel-ID` einen Nullwert aufweist, zu verwerfen. Dieser Auswahlknoten verwendet den folgenden CLEM-Ausdruck zum Verwerfen von Datensätzen: `not(@NULL(Rule_ID))`.

Weitere Informationen zum Transponieren von Scores für die Bereitstellung erhalten Sie beim Technical Support.

Sequenzknoten

Der Sequenzknoten erkennt Muster in sequenziellen oder zeitorientierten Daten, und zwar im Format `Brot -> Käse`. Die Elemente einer Sequenz sind **Elementsets**, die eine einzelne Transaktion ausmachen. Beispiel: Wenn eine Person in den Supermarkt geht und Brot und Milch kauft und dann ein paar Tage später zurückkehrt und Käse kauft, kann das Kaufverhalten dieser Person als zwei Elementsets dargestellt werden. Der erste Elementset enthält Brot und Milch, der zweite Käse. Eine **Sequenz** ist eine Liste mit Elementsets, die in einer vorhersagbaren Reihenfolge auftreten. Der Sequenzknoten erkennt häufige Sequenzen und erstellt einen generierten Modellknoten, der für Vorhersagen verwendet werden kann.

Anforderungen. Um ein Sequenzregelset zu erstellen, müssen Sie ein ID-Feld, ein optionales Zeitfeld und mindestens ein Inhaltsfeld angeben. Beachten Sie, dass diese Einstellungen auf der Registerkarte "Felder" des Modellierungsknotens vorgenommen werden müssen. Sie können nicht aus einem aufwärts liegenden Typknoten gelesen werden. Das ID-Feld kann eine beliebige Rolle oder ein beliebiges Messniveau aufweisen. Wenn Sie ein Zeitfeld angeben, kann es jede beliebige Rolle aufweisen, muss jedoch in numerischem, Datums-, Uhrzeit- oder Zeitmarkenformat gespeichert werden. Wenn Sie kein Zeitfeld angeben, verwendet der Sequenzknoten eine implizierte Zeitmarke, wobei als Zeitwerte Zeilennummern verwendet werden. Inhaltsfelder können ein beliebiges Messniveau und eine beliebige Rolle aufweisen, sämtliche Inhaltsfelder müssen jedoch denselben Typ aufweisen. Wenn es sich um numerische Felder handelt, müssen es Bereiche ganzer Zahlen (keine reellen Bereiche) sein.

Stärken. Der Sequenzknoten basiert auf dem CARMA-Assoziationsregelalgorithmus, der eine effiziente bidirektionale Methode zum Suchen von Sequenzen verwendet. Außerdem kann der von einem Sequenzknoten generierte Modellknoten in einen Datenstream eingefügt werden, um Vorhersagen zu erstellen. Der generierte Modellknoten kann auch Superknoten zum Erkennen und Zählen spezifischer Sequenzen und zum Erstellen von Vorhersagen basierend auf bestimmten Sequenzen generieren.

Feldoptionen für den Sequenzknoten

Vor der Ausführung eines Sequenzknotens müssen Sie ID- und Inhaltsfelder auf der Registerkarte "Felder" des Sequenzknotens angeben. Wenn Sie ein Zeitfeld verwenden möchten, müssen Sie auch dieses hier angeben.

ID-Feld. Wählen Sie ein ID-Feld aus der Liste aus. Als ID-Feld können numerische oder symbolische Felder verwendet werden. Jeder eindeutige Wert in diesem Feld sollte eine bestimmte Analyseeinheit darstellen. Bei einer Warenkorbanwendung könnte z. B. jede ID einen einzelnen Kunden darstellen. Für eine Webprotokoll-Analyseanwendung könnte jede ID einen Computer (nach IP-Adresse) oder einen Benutzer (nach Anmeldedaten) darstellen.

- **IDs sind zusammenhängend.** Wenn Ihre Daten vorsortiert sind, sodass alle Datensätze mit derselben ID im Datenstream zusammengefasst sind, wählen Sie diese Option, um die Verarbeitung zu beschleunigen. Wenn Ihre Daten nicht vorsortiert sind (oder Sie nicht sicher sind), lassen Sie diese Option inaktiviert. Die Daten werden dann vom Sequenzknoten automatisch sortiert.

Hinweis: Wenn Ihre Daten nicht sortiert sind und Sie diese Option auswählen, erhalten Sie möglicherweise ungültige Ergebnisse in Ihrem Sequenzmodell.

Zeitfeld. Wenn Sie ein Feld in den Daten verwenden möchten, um die Uhrzeiten von Ereignissen anzugeben, wählen Sie **Zeitfeld verwenden** und geben Sie das zu verwendende Feld an. Das Zeitfeld muss numerische, Datums-, Uhrzeit- oder Zeitmarkenwerte enthalten. Wenn kein Zeitfeld angegeben ist, wird davon ausgegangen, dass die Datensätze in sequenzieller Reihenfolge aus der Datenquelle ankommen und Datensatznummern als Zeitwerte verwendet werden (der erste Datensatz kommt zur Uhrzeit "1" vor; der zweite zur Uhrzeit "2" usw.).

Inhaltsfelder. Geben Sie die Inhaltsfelder für das Modell an. Diese Felder enthalten die in der Sequenzmodellierung interessanten Ereignisse.

Der Sequenzknoten kann Daten im Tabellenformat und im Transaktionsformat verarbeiten. Wenn Sie mehrere Felder für Transaktionsdaten verwenden, stellen die in diesen Feldern für einen bestimmten Datensatz angegebenen Elemente solche Elemente dar, die in einer Einzeltransaktion mit einer einzelnen Zeitmarke gefunden wurden. Weitere Informationen finden Sie im Thema „Tabellendaten im Vergleich zu Transaktionsdaten" auf Seite 232.

Partition. In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer anderen Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln. Wenn mehrere Partitionsfelder mithilfe von Typ- oder Partitionsknoten erstellt wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.) Beachten Sie außerdem, dass die Partitionierung auch auf der Registerkarte "Modelloptionen" des Knotens aktiviert sein muss, wenn die ausgewählte Partitionierung in Ihrer Analyse angewendet werden soll. (Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung ohne Änderung der Feldeinstellungen inaktiviert werden.)

Modelloptionen für den Sequenzknoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Minimale Regelunterstützung (%). Sie können auch ein Stützkriterium angeben. **Regelunterstützung** bezieht sich auf den Anteil von IDs in den Trainingsdaten, die die gesamte Sequenz enthalten. Wenn Sie weitere gemeinsame Sequenzen wünschen, erhöhen Sie diese Einstellung.

Minimale Regelkonfidenz (%). Sie können ein Konfidenzkriterium angeben, um die Sequenzen im Sequenzset beizubehalten. **Konfidenz** bezieht sich auf den prozentualen Anteil von IDs, für die eine richtige Vorhersage gemacht wird, aus allen IDs, für die die Regel eine Vorhersage macht. Sie wird aus der Anzahl von IDs berechnet, für die die gesamte Sequenz gefunden wird, dividiert durch die Anzahl der IDs, für die die Antezedenzien gefunden werden, basierend auf den Trainingsdaten. Sequenzen mit einer niedrigeren Konfidenz als das angegebene Kriterium werden verworfen. Wenn Sie zu viele Sequenzen oder uninteressante Sequenzen erhalten, sollten Sie diese Einstellung erhöhen. Wenn Sie zu wenige Sequenzen erhalten, sollten Sie diese Einstellung reduzieren.

Maximale Sequenzgröße. Sie können die maximale Anzahl unterschiedlicher *Elementsets* (im Gegensatz zu *Elementen*) in einer Sequenz festlegen. Wenn die gewünschten Sequenzen relativ kurz sind, können Sie diese Einstellung reduzieren, um die Erstellung des Sequenzsets zu beschleunigen.

Vorhersagen, die zum Stream hinzugefügt werden sollen. Geben Sie die Anzahl der Vorhersagen an, die dem Stream vom resultierenden generierten Modellknoten hinzugefügt werden sollen. Weitere Informationen finden Sie im Thema „Sequenzmodellnuggets" auf Seite 251.

Expertenoptionen für den Sequenzknoten

Für Personen mit umfassenden Kenntnissen über die Funktionsweise des Sequenzknotens ermöglichen die folgenden Expertenoptionen die Feinabstimmung des Modellerstellungsvorgangs. Für den Zugriff auf die Expertenoptionen stellen Sie den Modus auf der Registerkarte "Experten" auf **Experten** ein.

Maximale Zeitdauer. Falls diese Option ausgewählt ist, werden die Sequenzen auf solche mit einer Dauer (die Zeit zwischen dem ersten und letzten Elementset) beschränkt, die dem angegebenen Wert entspricht oder darunter liegt. Wenn Sie kein Zeitfeld angegeben haben, wird die Zeitdauer in den Rohdaten in Zeilen (Datensätzen) ausgedrückt. Wenn es sich bei dem verwendeten Zeitfeld um ein Uhrzeit-, Datums- oder Zeitmarkenfeld handelt, wird die Zeitdauer in Sekunden ausgedrückt. Im Falle von numerischen Feldern wird die Zeitdauer in denselben Einheiten ausgedrückt wie das Feld selbst.

Reduzierungswert festlegen. Der im Sequenzknoten verwendete CARMA-Algorithmus entfernt (**reduziert**) während der Verarbeitung seltene Elementsets aus seiner Liste potenzieller Elementsets. Wählen Sie diese Option, um die Reduzierungshäufigkeit anzupassen. Die angegebene Zahl legt die Reduzierungshäufigkeit fest. Geben Sie einen kleinen Wert ein, um die Speicheranforderungen des Algorithmus zu reduzieren (möglicherweise aber die erforderliche Trainingszeit zu erhöhen), oder geben Sie einen hohen Wert ein, um die Trainingsgeschwindigkeit zu erhöhen (möglicherweise aber die Speicheranforderungen zu erhöhen).

Maximale Anzahl von Sequenzen im Speicher. Falls diese Option ausgewählt ist, beschränkt der CARMA-Algorithmus seinen Speicher mit möglichen Sequenzen auf die Anzahl der angegebenen Sequenzen. Wählen Sie diese Option, wenn IBM SPSS Modeler während der Erstellung von Sequenzmodellen zu viel Speicher belegt. Beachten Sie, dass der von Ihnen hier angegebene maximale Sequenzenwert der Anzahl der möglichen Sequenzen entspricht, die intern bei der Modellerstellung aufgezeichnet wird. Diese Zahl sollte viel größer sein als die Zahl der Sequenzen, die Sie im endgültigen Modell erwarten.

Zeitspannen zwischen Bestandteilen von Sequenzen begrenzen. Mit dieser Option können Sie Beschränkungen der Zeitspannen festlegen, die zwischen verschiedenen Elementsets liegen. Falls ausgewählt, werden die Elementsets mit Zeitspannen, die unter der angegebenen **Minimimalen Zeitspanne** oder über der angegebenen **Maximalen Zeitspanne** liegen, nicht als Teil einer Sequenz betrachtet. Verwenden Sie diese Option, um zu vermeiden, dass Sequenzen gezählt werden, die lange Zeitintervalle enthalten oder die in einer sehr kurzen Zeitspanne auftreten.

Hinweis: Wenn es sich bei dem verwendeten Zeitfeld um ein Uhrzeit-, Datums- oder Zeitmarkenfeld handelt, wird die Zeitspanne in Sekunden festgelegt. Im Falle von numerischen Feldern wird die Zeitdauer in denselben Einheiten ausgedrückt wie das Zeitfeld.

Betrachten Sie beispielsweise die folgende Liste mit Transaktionen.

Tabelle 18. Beispieltransaktionsliste.

ID	Zeit	Inhalt
1001	E	Äpfel
1001	Z	Brot
1001	5	Käse
1001	6	Dressing

Wenn Sie mit diesen Daten ein Modell bilden, wobei der minimale Abstand auf 2 gesetzt ist, erhalten Sie die folgenden Sequenzen:

```
Äpfel -> Käse

Äpfel -> Dressing

Brot -> Käse

Brot -> Dressing
```

Sequenzen wie `Äpfel->Brot` erscheinen nicht, weil der Abstand zwischen `Äpfel` und `Brot` kleiner als der Mindestabstand ist. Betrachten Sie auch die folgenden alternativen Daten.

Tabelle 19. Beispieltransaktionsliste.

ID	Zeit	Inhalt
1001	E	Äpfel
1001	Z	Brot
1001	5	Käse
1001	20	Dressing

Wenn die maximale Lücke auf 10 gesetzt wurde, erhalten Sie keine Sequenzen mit `Dressing`, da die Lücke zwischen `Käse` und `Dressing` zu groß ist, um als Teil derselben Sequenz betrachtet zu werden.

Sequenzmodellnuggets

Sequenzmodellnuggets stellen die Sequenzen dar, die in einem bestimmten, vom Sequenzknoten ermittelten Ausgabefeld gefunden wurden und zum Erstellen von Vorhersagen einem Stream hinzugefügt werden können.

Bei der Ausführung eines Streams, der einen Sequenzknoten enthält, fügt der Knoten ein Felderpaar hinzu, das die Vorhersagen und die zugeordneten Konfidenzwerte für die einzelnen Vorhersagen aus dem Sequenzmodell den Daten hinzufügt. Standardmäßig werden drei Felderpaare mit den drei Top-Vorhersagen (und den zugehörigen Konfidenzwerten) hinzugefügt. Sie können die Anzahl der generierten Vorhersagen ändern, wenn Sie das Modell durch Festlegung der Sequenzknotenmodelloptionen zum Erstellungszeitpunkt erstellen, oder auch auf der Registerkarte "Einstellungen", nachdem das Modellnugget einem Stream hinzugefügt wurde. Weitere Informationen finden Sie im Thema „Sequenzmodellnugget - Einstellungen” auf Seite 254.

Die neuen Feldnamen werden aus dem Modellnamen abgeleitet. Die Feldnamen lauten *$S-sequence-n* für das Vorhersagefeld (dabei gibt *n* die *n*-te Vorhersage an) und *$SC-sequence-n* für das Konfidenzfeld. In einem Stream mit mehreren Sequenzregelknoten in einer Reihe enthalten die neuen Feldnamen Zahlen im Präfix, damit sie auseinander gehalten werden können. Beim ersten Sequenzset-Knoten im Stream werden die üblichen Namen verwendet, beim zweiten Knoten Namen, die mit *$S1-* und *$SC1-* beginnen, beim dritten Knoten Namen mit *$S2-* und *$SC2-* usw. Die Vorhersagen werden nach Konfidenz geordnet angezeigt, sodass *$S-sequence-1* die Vorhersage mit der höchsten Konfidenz enthält, *$S-sequence-2* die Vorhersage mit der zweithöchsten Konfidenz usw. Bei Datensätzen, bei denen die Anzahl der verfügbaren Vorhersagen kleiner ist als die Anzahl der angeforderten Vorhersagen, enthalten die restlichen Vorhersagen den Wert `$null$`. Beispiel: Wenn nur zwei Vorhersagen für einen bestimmten Datensatz vorgenommen werden können, weisen *$S-sequence-3* und *$SC-sequence-3* den Wert `$null$` auf.

Bei jedem Datensatz werden die Regeln im Modell mit der Menge der Transaktionen verglichen, die bisher für die aktuelle ID verarbeitet wurden, einschließlich des aktuellen Datensatzes und aller vorangegangenen Datensätze mit derselben ID und früheren Zeitmarken. Die *k* Regeln mit den höchsten Konfidenzwerten, die für dieses Set von Transaktionen gelten, werden verwendet, um die *k* Vorhersagen für den Datensatz zu generieren. Dabei ist *k* die Anzahl der Vorhersagen, die nach dem Hinzufügen des Modells zum Stream auf der Registerkarte "Einstellungen" angegeben wurden. (Wenn mehrere Regeln dasselbe Er-

gebnis für das Transaktionsset vorhersagen, wird nur die Regel mit der höchsten Konfidenz verwendet.) Weitere Informationen finden Sie im Thema „Sequenzmodellnugget - Einstellungen" auf Seite 254.

Wie bei anderen Arten von Assoziationsregelmodellen muss das Datenformat mit dem Format übereinstimmen, das beim Aufbau des Sequenzmodells verwendet wurde. Mit Modellen, die mithilfe von Tabellendaten erstellt wurden, können entsprechend nur Tabellendaten gescort werden. Weitere Informationen finden Sie im Thema „Scoren von Assoziationsregeln" auf Seite 244.

Hinweis: Beim Scoring von Daten mithilfe eines generierten Sequenzsetknotens in einem Stream, werden alle Toleranz- oder Lückeneinstellungen, die Sie beim Erstellen des Modells ausgewählt haben, beim Scoring ignoriert.

Vorhersagen aus Sequenzregeln

Der Knoten bearbeitet die Datensätze in zeitabhängiger Weise (bzw. in Abhängigkeit von der Reihenfolge, wenn beim Erstellen des Modells kein Zeitmarkenfeld verwendet wurde). Die Datensätze sollten nach dem ID-Feld und dem Zeitmarkenfeld (sofern vorhanden) sortiert werden. Die Vorhersagen sind jedoch nicht an die Zeitmarke des Datensatzes gebunden, dem sie hinzugefügt werden. Sie beziehen sich einfach auf die Elemente, die unter Berücksichtigung des Transaktionsverlaufs für die aktuelle ID bis zum aktuellen Datensatz mit der größten Wahrscheinlichkeit *irgendwann in der Zukunft* auftreten.

Beachten Sie, dass die Vorhersagen für die einzelnen Datensätze nicht unbedingt von den Transaktionen des betreffenden Datensatzes abhängen. Wenn die Transaktionen des aktuellen Datensatzes keine spezifische Regel auslösen, werden die Regeln anhand der vorangegangenen Transaktionen für die aktuelle ID ausgewählt. Anders ausgedrückt: Wenn der aktuelle Datensatz keine verwertbaren Vorhersageinformationen zur Sequenz hinzufügt, wird die Vorhersage aus der letzten nützlichen Transaktion für diese ID auf den aktuellen Datensatz übertragen.

Beispiel: Angenommen Sie haben ein Sequenzmodell mit nur einer einzigen Regel:

```
Marmelade -> Brot (0,66)
```

und Sie übergeben es an die folgenden Datensätze.

Tabelle 20. Beispieldatensätze.

ID	Kauf	Vorhersage
001	Marmelade	Brot
001	Milch	Brot

Der erste Datensatz generiert, wie zu erwarten, eine Vorhersage für *Brot*. Der zweite Datensatz enthält ebenfalls eine Vorhersage für *Brot*, da keine Regel für *Marmelade* gefolgt von *Milch* vorliegt; daher fügt die Transaktion *Milch* keine verwertbaren Informationen hinzu und die Regel `Marmelade -> Brot` gilt weiterhin.

Erzeugen neuer Knoten

Im Menü "Generieren" können Sie anhand des Sequenzmodells neue Superknoten erstellen.

- **Regelsuperknoten.** Erstellt einen Superknoten, der die Vorkommen von Sequenzen in den gescorten Daten ermitteln und zählen kann. Diese Option ist inaktiviert, wenn keine Regel ausgewählt wurde. Weitere Informationen finden Sie im Thema „Generieren eines Regelsuperknotens aus einem Sequenzmodellnugget" auf Seite 255.
- **Modell in Palette.** Gibt das Modell an die Modellpalette zurück. Das ist nützlich, wenn Sie von einem Kollegen einen Stream, der das Modell enthält, jedoch nicht das Modell selbst erhalten.

Nähere Informationen zum Sequenzmodellnugget

Auf der Registerkarte "Modell" eines Sequenzmodellnuggets werden die Regeln angezeigt, die durch den Algorithmus extrahiert wurden. Jede Zeile in der Tabelle steht für eine Regel, bei der das Antezedens (der "Wenn"-Teil der Regel) in der ersten Spalte jeweils vom Sukzedens (dem "Dann"-Teil der Regel) in der zweiten Spalte gefolgt wird.

Jede Regel wird im folgenden Format angezeigt.

Tabelle 21. Regelformat

Antezedens	Sukzedens
Bier und Dosengemüse	Bier
Fisch Fisch	Fisch

Die erste Beispielregel wird wie folgt interpretiert: *Bei IDs, bei denen "Bier" und "Dosengemüse" in derselben Transaktion vorkamen, wird "Bier" mit hoher Wahrscheinlichkeit ein weiteres Mal vorkommen.* Die zweite Beispielregel kann wie folgt interpretiert werden: *Bei IDs, bei denen "Fisch" in einer Transaktion und anschließend "Fisch" in einer anderen Transaktion vorkam, wird "Fisch" mit hoher Wahrscheinlichkeit ein weiteres Mal vorkommen.* In der ersten Regel werden *Bier* und *Dosengemüse* gleichzeitig eingekauft; in der zweiten Regel wird *Fisch* in zwei separaten Transaktionen erworben.

Menü "Sortieren". Mit der Schaltfläche des Menüs "Sortieren" in der Symbolleiste wird die Sortierung der Regeln gesteuert. Die Sortierrichtung (aufsteigend oder absteigend) lässt sich mit der Schaltfläche für die Sortierrichtung (nach unten bzw. oben zeigender Pfeil) ändern.

Regeln können nach folgenden Faktoren sortiert werden:

- Unterstützung %
- Konfidenz %
- Regelunterstützung %
- Sukzedens
- Erstes Antezedens
- Letztes Antezedens
- Anzahl der Elemente (Antezedenzien)

Beispiel: Die nachstehende Tabelle wird in absteigender Reihenfolge nach der Anzahl der Elemente sortiert. Regeln mit mehreren Elementen im Antezedenssatz haben Vorrang vor Regeln mit weniger Elementen.

Tabelle 22. Nach Anzahl der Elemente sortierte Regeln

Antezedens	Sukzedens
Bier und Dosengemüse und TK-Fertiggericht	TK-Fertiggericht
Bier und Dosengemüse	Bier
Fisch Fisch	Fisch
Softdrink	Softdrink

Kriterien anzeigen/ausblenden. Das Menü "Anzeigen/Ausblenden" der Schaltfläche "Kriterien" (Rastersymbol) steuert die Optionen für die Anzeige der Regeln. Die folgenden Anzeigeoptionen sind verfügbar:

- **Instanzen** zeigt Informationen zur Anzahl der eindeutigen IDs an, für die die *vollständige Sequenz* vorkommt (also sowohl Antezedenzien als auch Sukzedenzien). (Hier besteht ein Unterschied zu den As-

soziationsmodellen, bei denen die Anzahl der Instanzen die Anzahl der IDs bezeichnet, bei denen *ausschließlich* die Antezedenzien gelten.) Bei der Regel `Brot -> Käse` beispielsweise wird die Anzahl der IDs in den Trainingsdaten, die sowohl *Brot* als auch *Käse* enthalten, als **Instanzen** bezeichnet.

- **Unterstützung** zeigt den Anteil von IDs in den Trainingsdaten, für den die Antezedenzien wahr sind. Beispiel: Wenn 50 % der Trainingsdaten das Antezedens *Brot* enthalten, hat die Regel `Brot -> Käse` eine Antezedens-Unterstützung von 50 %. (Im Gegensatz zu den Assoziationsmodellen beruht die Unterstützung *nicht* auf der Anzahl der Instanzen, wie zuvor beschrieben.)
- **Konfidenz** zeigt den prozentualen Anteil von IDs, für die eine richtige Vorhersage gemacht wird, aus allen IDs, für die die Regel eine Vorhersage macht. Sie wird aus der Anzahl von IDs berechnet, für die die gesamte Sequenz gefunden wird, dividiert durch die Anzahl der IDs, für die die Antezedenzien gefunden werden, basierend auf den Trainingsdaten. Beispiel: Wenn 50 % der Trainingsdaten `Dosengemüse` enthalten (Antezedens-Unterstützung), jedoch nur 20 % sowohl `Dosengemüse` als auch `TK-Fertiggericht`, dann ist die Konfidenz für die Regel `Dosengemüse -> TK-Fertiggericht` gleich `Regelunterstützung/Antezedens-Unterstützung`, in diesem Fall also 40 %.
- **Regelunterstützung** für Sequenzmodelle beruht auf den Instanzen und zeigt den Teil der Trainings-Datensätze an, für die die gesamte Regel, die Antezedenzien und das Sukzedens (bzw. die Sukzedenzien) wahr sind. Beispiel: Wenn 20 % der Trainingsdaten sowohl *Brot* als auch *Käse* enthalten, beträgt die Regelunterstützung für die Regel `Brot -> Käse` 20 %.

Beachten Sie, dass die Anteile auf gültigen Transaktionen beruhen (Transaktionen mit mindestens einem beobachteten Element oder Wahr-Wert) und nicht auf der Gesamtzahl der Transaktionen. Ungültige Transaktionen, Transaktionen ohne Elemente oder Wahr-Werte, werden für diese Berechnungen verworfen.

Schaltfläche "Filter". Mit der Schaltfläche "Filter" (Trichtersymbol) im Menü wird der untere Teil des Dialogfelds erweitert und ein Fenster mit aktiven Regelfiltern wird angezeigt. Filter werden verwendet, um die Anzahl der auf der Registerkarte "Modelle" angezeigten Regeln einzugrenzen.

Abbildung 52. Schaltfläche "Filter"

Zum Erstellen von Filtern klicken Sie auf das Filtersymbol rechts neben dem erweiterten Fenster. Dadurch wird ein separates Dialogfeld geöffnet, in dem Sie Bedingungen für die Anzeige von Regeln eingeben können. Beachten Sie, dass die Filterschaltfläche häufig in Verbindung mit dem Menü "Generieren" verwendet wird, um zunächst die Regeln zu filtern und anschließend ein Modell zu erstellen, das das betreffende Subset der Regeln enthält. Weitere Informationen finden Sie unter „Angeben von Filtern für Regeln” auf Seite 241.

Sequenzmodellnugget - Einstellungen

Auf der Registerkarte "Einstellungen" eines Sequenzmodellnuggets werden die Scoring-Optionen für das Modell angezeigt. Diese Registerkarte ist erst dann verfügbar, nachdem das Modell zum Zwecke des Scoring zum Streamerstellungsbereich hinzugefügt wurde.

Maximale Anzahl an Vorhersagen. Dient zur Angabe der maximalen Anzahl an Vorhersagen, die für jedes Set von Warenkorbelementen aufgenommen werden. Anhand der Regeln mit den höchsten Konfidenzwerten, die für dieses Set von Transaktionen gelten, werden die Vorhersagen für den Datensatz bis zum angegebenen Grenzwert generiert.

Sequenzmodellnugget - Übersicht

Auf der Registerkarte "Übersicht" für ein Sequenzregelmodellnugget werden die Anzahl der ermittelten Regeln sowie die Mindest- und Höchstwerte für Unterstützung und Konfidenz für die Regeln angegeben.

Wenn Sie einen Analyseknoten ausgeführt haben, der an diesen Modellierungsknoten angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt.

Weitere Informationen finden Sie im Thema „Durchsuchen von Modellnuggets" auf Seite 42.

Generieren eines Regelsuperknotens aus einem Sequenzmodellnugget

So generieren Sie einen Regelsuperknoten auf der Grundlage einer Sequenzregel:

1. Klicken Sie auf der Registerkarte "Modell" für das Sequenzregelmodellnugget auf die Zeile in der Tabelle, in der die gewünschte Regel verzeichnet ist.
2. Wählen Sie im Regelbrowser in den Menüs die folgende Befehlsfolge:

   Generieren > **Regelsuperknoten**

Wichtig: Zur Verwendung des generierten Superknotens müssen Sie die Daten nach ID-Feld (und Zeitfeld, falls vorhanden) sortieren, bevor Sie sie an den Superknoten weiterleiten. In nicht sortierten Daten kann der Superknoten die Sequenzen nicht ordnungsgemäß erkennen.

Folgende Optionen sind für das Generieren von Regelsuperknoten verfügbar:

Erkennen. Gibt an, wie für die an den Superknoten übergebenen Daten Übereinstimmungen definiert sind.

- **Nur Antezedenzien.** Der Superknoten ermittelt eine Übereinstimmung jedes Mal, wenn er die Antezedenzien für die ausgewählte Regel in der richtigen Reihenfolge in einem Set von Datensätzen mit derselben ID findet, unabhängig davon, ob das Sukzedens ebenfalls gefunden wird. Beachten Sie, dass hierbei nicht die Beschränkungseinstellungen für die Zeitmarkentoleranz oder die Elementlücken aus dem ursprünglichen Sequenzmodellierungsknoten berücksichtigt werden. Wenn das letzte Antezedens-Elementset im Stream ermittelt wird (und alle anderen Antezedenzien in der richtigen Reihenfolge gefunden wurden), enthalten alle nachfolgenden Datensätze mit der aktuellen ID die unten ausgewählte Übersicht.
- **Gesamte Sequenz.** Der Superknoten ermittelt eine Übereinstimmung jedes Mal, wenn er die Antezedenzien und das Sukzedens für die ausgewählte Regel in der richtigen Reihenfolge in einem Set von Datensätzen mit derselben ID findet. Hierbei werden nicht die Beschränkungseinstellungen für die Zeitmarkentoleranz oder die Elementlücken aus dem ursprünglichen Sequenzmodellierungsknoten berücksichtigt. Wenn das letzte Sukzedens im Stream ermittelt wird (und alle Antezedenzien ebenfalls in der richtigen Reihenfolge gefunden wurden), enthalten der aktuelle Datensatz und alle nachfolgenden Datensätze mit der aktuellen ID die unten ausgewählte Übersicht.

Anzeigen. Steuert, wie Übereinstimmungsübersichten zu den Daten in der Ausgabe des Regelsuperknotens hinzugefügt werden.

- **Sukzedenswert für erstes Vorkommen.** Bei dem zu den Daten hinzugefügten Wert handelt es sich um den Wert des Sukzedens, der auf der Grundlage des ersten Vorkommens der Übereinstimmung vorhergesagt wurde. Die Werte werden als neues Feld mit der Bezeichnung *rule_n_consequent* hinzugefügt. Dabei ist *n* die Regelnummer (gemäß der Erstellungsreihenfolge der Regelsuperknoten im Stream).
- **Echten Wert für erstes Vorkommen.** Der zu den Daten hinzugefügte Wert ist wahr, wenn mindestens eine Übereinstimmung für die ID vorliegt, und falsch, wenn keine Übereinstimmung vorliegt. Die Werte werden als neues Feld mit der Bezeichnung *rule_n_flag* hinzugefügt.
- **Anzahl der Vorkommen.** Bei dem zu den Daten hinzugefügten Wert handelt es sich um die Anzahl der Übereinstimmungen für die ID. Die Werte werden als neues Feld mit der Bezeichnung *rule_n_count* hinzugefügt.
- **Regelnummer.** Der hinzugefügte Wert ist die Regelnummer für die ausgewählte Regel. **Regelnummern** werden auf der Grundlage der Reihenfolge zugewiesen, in der der Superknoten zum Stream hinzugefügt wurde. Der erste Regelsuperknoten beispielsweise wird als *rule 1* betrachtet, der zweite als *rule 2* usw. Diese Option ist sinnvoll, wenn Sie mehrere Superknoten in Ihrem Stream berücksichtigen möchten. Die Werte werden als neues Feld mit der Bezeichnung *rule_n_number* hinzugefügt.

- **Konfidenzzahlen einschließen.** Bei Auswahl dieser Option wird die Regelkonfidenz zum Datenstream hinzugefügt, ebenso wie die ausgewählten Übersichtsdaten. Die Werte werden als neues Feld mit der Bezeichnung *rule_n_confidence* hinzugefügt.

Kapitel 13. Zeitreihenmodelle

Wozu dienen Vorhersagen?

Bei einer Vorhersage werden die Werte für eine oder mehrere Reihen im zeitlichen Verlauf vorhergesagt. Beispiel: Sie möchten vorhersagen, wie die erwartete Nachfrage für eine Produktlinie oder eine Dienstleistung aussehen wird, um Ressourcen für die Fertigung oder Distribution zuzuordnen. Da die Implementierung der Planung von Entscheidungen zeitaufwendig ist, bilden Vorhersagen bei vielen Planungsprozessen ein wichtiges Tool.

Die Methoden zur Modellierung von Zeitreihen gehen davon aus, dass sich die Geschichte wiederholt, wenn nicht genau, dann doch genau genug, dass eine Untersuchung der Vergangenheit genauere Entscheidungen in der Zukunft ermöglicht. Um z. B. die Verkaufszahlen für das nächste Jahr vorherzusagen, würden Sie wahrscheinlich damit beginnen, die Verkaufszahlen dieses und früherer Jahre zu untersuchen und eventuell vorhandene Trends oder Muster herauszuarbeiten, die sich in früheren Jahren entwickelt haben. Es kann sich aber als schwierig erweisen, Muster zu beurteilen. Wenn Ihre Verkaufszahlen beispielsweise über mehrere Wochen in Folge ansteigen, handelt es sich dann um den Beginn eines saisonbedingten Zyklus oder um den Anfang eines langfristigen Trends?

Durch statistische Modellierungsverfahren können Sie die in Ihren Vergangenheitsdaten vorhandenen Muster analysieren und projizieren, um einen Bereich zu ermitteln, in dem zukünftige Werte der Reihen wahrscheinlich liegen werden. Als Ergebnis erhalten Sie genauere Vorhersagen, auf deren Grundlage Sie Entscheidungen treffen können.

Zeitreihendaten

Bei einer **Zeitreihe** handelt es sich um eine sortierte Sammlung von Messdaten, die in regelmäßigen Zeitabständen ermittelt wurden - beispielsweise tägliche Lagerkosten oder wöchentliche Verkaufsdaten. Die Messungen können sich auf alles beziehen, was für Sie von Interesse ist. Alle Reihen können wie folgt klassifiziert werden:

- **Abhängig.** Eine Reihe, die Sie vorhersagen möchten.
- **Prädiktor.** Eine Reihe, die bei der Erklärung des Ziel hilfreich sein kann - wenn z. B. ein Anzeigebudget verwendet wird, um den Absatz vorherzusagen. Prädiktoren können nur mit ARIMA-Modellen verwendet werden.
- **Ereignis.** Eine spezielle Prädiktorenreihe, die verwendet wird, um vorhersagbare wiederholt auftretende Vorfälle, beispielsweise Werbeaktionen, zu berücksichtigen.
- **Intervention.** Eine spezielle Prädiktorenreihe, die verwendet wird, um einmalige Vorfälle in der Vergangenheit zu berücksichtigen, wie beispielsweise einen Stromausfall oder einen Streik.

Bei den Intervallen kann es sich um beliebige Zeiträume handeln. Das Intervall muss jedoch für alle Messungen identisch sein. Darüber hinaus müssen alle Intervalle, für die keine Messungen vorliegen, als fehlende Werte festgelegt werden. Demnach definiert die Anzahl der Intervalle mit Messungen (einschließlich derer mit fehlenden Werten) den Umfang der historischen Spannweite der Daten.

Merkmale von Zeitreihen

Die Untersuchung der früheren Ergebnisse einer Reihe erleichtert das Auffinden von Mustern und sorgt für bessere Vorhersagen. Bei der Darstellung als Diagramm zeigen viele Zeitreihen eines oder mehrere der folgenden Merkmale:

- Trends
- Saisonale und nicht saisonale Zyklen
- Impulse und Schritte

- Ausreißer

Trends

Ein **Trend** ist eine allmähliche Aufwärts- oder Abwärtsveränderung der Ebene von Reihen oder der Tendenz von Reihenwerten, im Laufe der Zeit zu steigen oder zu sinken.

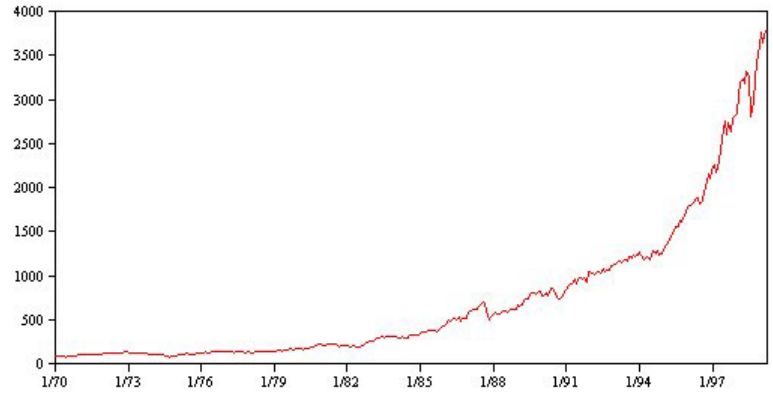


Abbildung 53. Trend

Trends sind entweder **lokal** oder **global**, eine einzelne Reihe kann jedoch beide Typen aufweisen. Reihendiagramme des Aktienindexes zeigen historisch einen globalen Aufwärtstrend. In Zeiten der Rezession tauchen lokale Abwärtstrends und in Zeiten der Hochkonjunktur lokale Aufwärtstrends auf.

Trends können außerdem entweder **linear** oder **nicht linear** sein. Lineare Trends sind positive oder negative additive Erhöhungen der Ebene von Reihen, vergleichbar mit dem Effekt der einfachen Kapitalverzinsung. Nicht lineare Trends sind häufig multiplikativ, mit Inkrementen, die sich proportional zu den vorherigen Reihenwerten verhalten.

Globale lineare Trends werden angepasst und liefern gute Vorhersagen sowohl bei Modellen für das exponentielle Glätten als auch mit ARIMA-Modellen. Beim Erstellen von ARIMA-Modellen werden Reihen, die Trends aufweisen, in der Regel unterschieden, um die Auswirkung des Trends zu beseitigen.

Saisonale Zyklen

Ein **saisonaler Zyklus** ist ein sich wiederholendes, vorhersagbares Muster in den Reihenwerten.

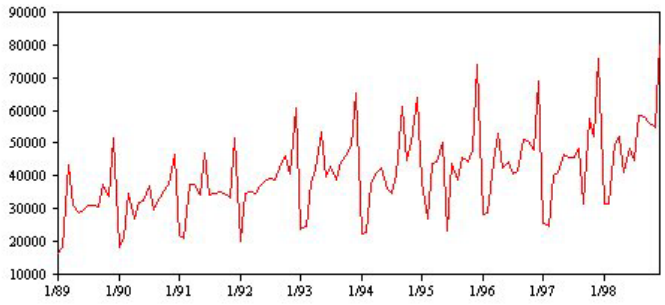


Abbildung 54. Saisonaler Zyklus

Saisonale Zyklen sind mit dem Intervall ihrer Reihen verbunden. Monatsdaten weisen z. B. in der Regel Zyklen über Quartale und Jahre auf. Eine Monatsreihe kann einen signifikanten vierteljährlichen Zyklus aufweisen, der im ersten Quartal gering ausfällt, oder einen jährlichen Zyklus mit einem Spitzenwert im Dezember. Reihen, die einen saisonalen Zyklus enthalten, zeigen eine **Saisonalität**.

Saisonale Muster sind hilfreich, um brauchbare Anpassungen und Vorhersagen zu erhalten. Die Saisonalität wird durch Modelle für das exponentielle Glätten sowie durch ARIMA-Modelle erfasst.

Nicht saisonale Zyklen

Ein **nicht saisonaler Zyklus** ist ein sich wiederholendes, möglicherweise vorhersagbares Muster in den Reihenwerten.

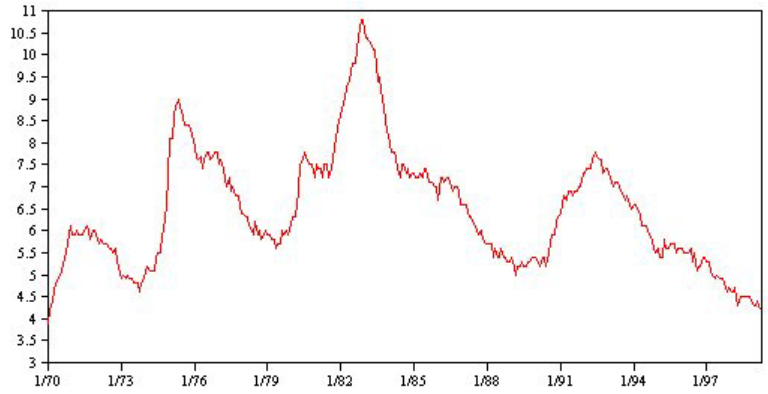


Abbildung 55. Nicht saisonaler Zyklus

Manche Reihen, wie die Arbeitslosenquote, zeigen einen deutlich zyklischen Verlauf. Die Periodizität des Zyklus verändert sich jedoch im Laufe der Zeit, weshalb es schwierig vorhersagbar ist, wann ein Höchst- oder Tiefstwert eintreten wird. Andere Reihen besitzen möglicherweise vorhersagbare Zyklen, passen aber nicht wirklich in den Gregorianischen Kalender oder haben Zyklen, die länger als ein Jahr sind. Die Gezeiten folgen z. B. dem Mondkalender, internationale Reise- und Handelsaktivitäten im Zusammenhang mit den Olympischen Spielen steigen alle vier Jahre an und es gibt viele religiöse Feiertage, die jedes Jahr auf ein anderes Datum fallen.

Nicht saisonale zyklische Muster sind schwierig zu modellieren und erhöhen in der Regel die Unsicherheit der Vorhersagen. Der Aktienmarkt bietet beispielsweise eine Vielzahl von Beispielen für Reihen, die sich den Anstrengungen beim Erstellen von Vorhersagen widersetzt haben. Trotzdem müssen nicht saisonale Muster berücksichtigt werden, wenn sie vorhanden sind. In vielen Fällen können Sie immer noch ein Modell identifizieren, das halbwegs gut zu den historischen Daten passt und Ihnen die beste Chance bietet, die Unsicherheit der Vorhersage möglichst zu minimieren.

Impulse und Schritte

In vielen Reihen treten plötzliche Ebenenänderungen auf. Diese gehören in der Regel zu zwei Typen:

- Eine plötzliche *vorübergehende* Änderung oder ein **Impuls** in der Ebene der Reihe
- Eine plötzliche *permanente* Änderung oder ein **Impuls** in der Ebene der Reihe

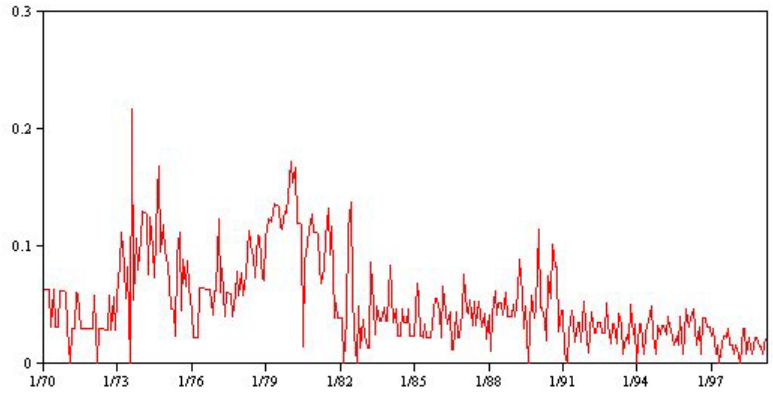


Abbildung 56. Reihen mit einem Impuls

Wenn Schritte oder Impulse beobachtet werden, ist es wichtig, eine plausible Erklärung zu finden. Zeitreihenmodelle sind so ausgelegt, dass sie allmähliche Änderungen berücksichtigen, keine plötzlichen. Demzufolge tendieren sie dazu, Impulse unterzubewerten und durch Schritte ruiniert zu werden, was zu einer schlechten Modellanpassungsgüte und unsicheren Vorhersagen führt. (Bei einigen Instanzen der Saisonalität kann eine scheinbare plötzliche Änderung der Ebene vorliegen, während die Ebene von einem saisonalen Zeitraum zum nächsten aber konstant ist.)

Wenn eine Störung erklärt werden kann, kann sie mithilfe einer **Intervention** oder eines **Ereignisses** modelliert werden. Im August 1973 hat beispielsweise ein Erdölembargo der Organisation der Erdöl exportierenden Länder (OPEC) eine drastische Veränderung der Inflationsrate ausgelöst, die dann in den darauffolgenden Monaten wieder auf einen normalen Stand zurückkehrte. Durch die Angabe einer **Punkt-Intervention** für den Monat des Embargos können Sie die Anpassungsgüte Ihres Modells verbessern und Ihre Vorhersagen so indirekt verbessern. Ein Einzelhandelsgeschäft stellt z. B. fest, dass der Absatz an einem Tag, an dem alle Artikel mit 50 % Rabatt gekennzeichnet waren, viel höher als gewöhnlich ist. Indem die 50 %-Rabatt-Aktion als wiederkehrendes **Ereignis** festgelegt wird, können Sie die Anpassungsgüte Ihres Modells verbessern und den Effekt der Wiederholung der Aktion in der Zukunft schätzen.

Ausreißer

Veränderungen der Ebene einer Zeitreihe, die nicht erklärt werden können, werden als **Ausreißer** bezeichnet. Diese Beobachtungen sind nicht mit dem Rest der Reihen konsistent und können sich drastisch auf die Analyse auswirken und dementsprechend die Vorhersagefähigkeit des Modells beeinflussen.

Die folgende Abbildung zeigt verschiedene Typen von Ausreißern, die häufig in Zeitreihen auftreten. Die blaue Linie stellt eine Reihe ohne Ausreißer dar. Die roten Linien zeigen ein Muster, das vorliegen kann, wenn die Reihen Ausreißer enthalten. Diese Ausreißer werden als **deterministisch** klassifiziert, da sie sich nur auf die durchschnittliche Ebene der Reihen auswirken.

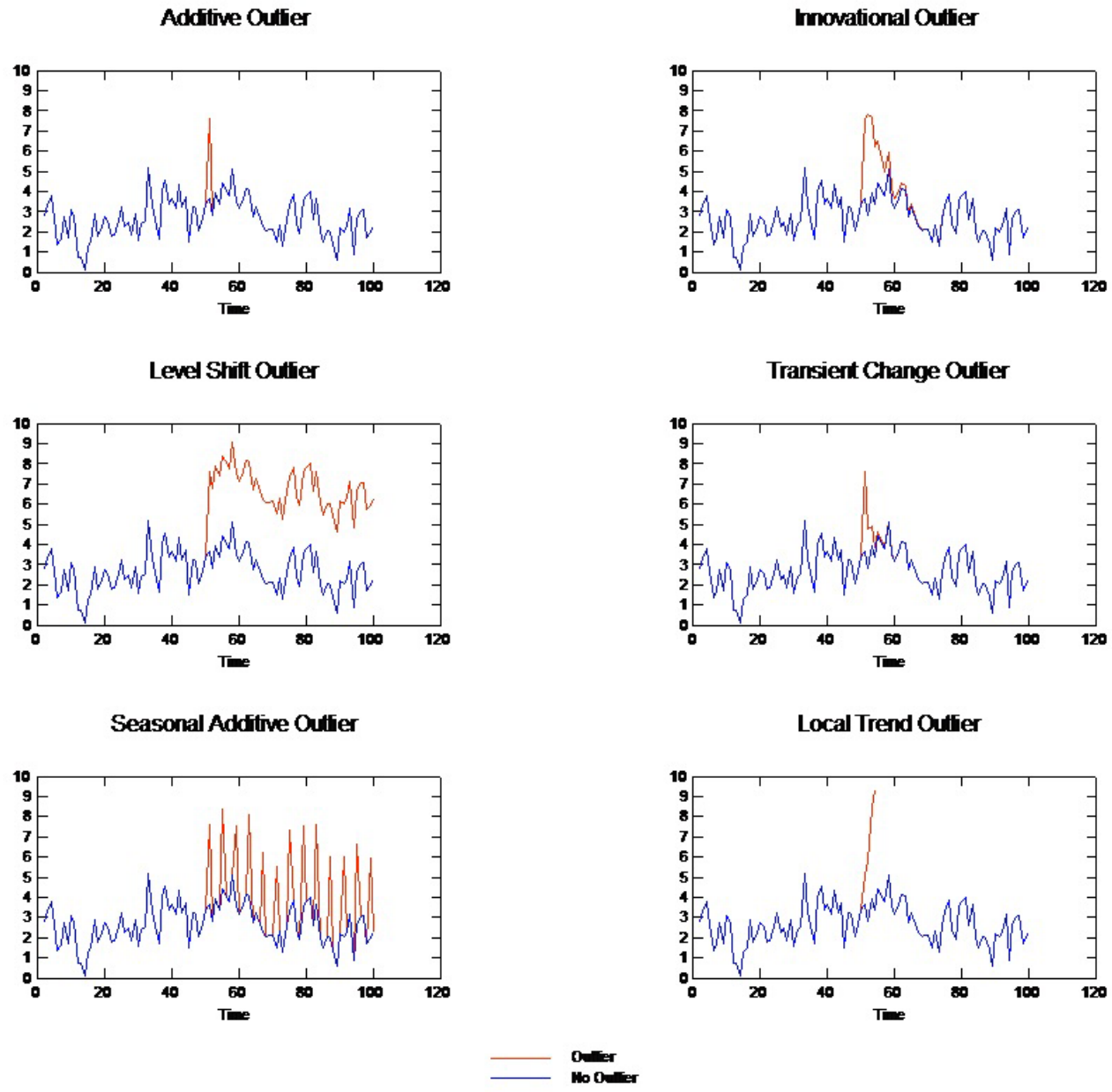


Abbildung 57. Ausreißertypen

- **Additiver Ausreißer.** Ein additiver Ausreißer tritt als überraschend großer oder kleiner Wert auf, der für eine einzelne Beobachtung erscheint. Nachfolgende Beobachtungen sind nicht durch den additiven Ausreißer beeinflusst. Aufeinanderfolgende additive Ausreißer werden in der Regel als **additive Ausreißerpatchs** bezeichnet.
- **Innovativer Ausreißer.** Ein innovatorischer Ausreißer ist gekennzeichnet durch eine anfängliche Auswirkung, deren Effekt bei nachfolgenden Beobachtungen fortbesteht. Der Einfluss des Ausreißers kann im zeitlichen Verlauf ansteigen.
- **Im Niveau verschobener Ausreißer.** Bei einer Ebenenänderung verschieben sich alle nach dem Ausreißer liegenden Beobachtungen auf eine neue Ebene. Im Gegensatz zu additiven Ausreißern wirkt sich ein Ebenen ändernder Ausreißer auf viele Beobachtungen aus und besitzt einen permanenten Effekt.

- **Ausreißer mit vorübergehender Änderung.** Ausreißer mit vorübergehender Änderung sind ähnlich wie Ausreißer mit Ebenenänderung, der Effekt der Ausreißer verringert sich aber bei den nachfolgenden Beobachtungen exponentiell. Die Reihe kehrt schließlich auf ihre normale Ebene zurück.
- **Saisonal additiver Ausreißer.** Ein saisonal additiver Ausreißer tritt als überraschend großer oder kleiner Wert auf, der wiederholt in regelmäßigen Intervallen erscheint.
- **Ausreißer mit lokalem Trend.** Ein Ausreißer mit lokalem Trend verursacht in den Reihen eine allgemeine Tendenz, die nach dem Entstehen des ersten Ausreißers durch ein Muster in den Ausreißern verursacht wird.

Beim Erkennen von Ausreißern in Zeitreihen müssen die Position, der Typ und der Umfang aller vorhandenen Ausreißer festgestellt werden. Tsay (1988) hat ein iteratives Verfahren zum Erkennen der durchschnittlichen Ebenenänderung vorgeschlagen, mit dem deterministische Ausreißer identifiziert werden. Bei diesem Prozess wird ein Zeitreihenmodell, in dem angenommen wird, dass keine Ausreißer vorhanden sind, mit einem anderen Modell verglichen, das Ausreißer berücksichtigt. Die Differenzen zwischen den Modellen ergeben Schätzungen dafür, welchen Effekt die Behandlung eines beliebigen Punkts als Ausreißer besitzt.

Autokorrelation und partielle Autokorrelationsfunktionen

Die Autokorrelation und die partielle Autokorrelation sind Maßstäbe für die Beziehung zwischen aktuellen und vergangenen Reihenwerten, die anzeigen, welche vergangenen Reihenwerte bei der Vorhersage zukünftiger Werte am nützlichsten sind. Mit diesem Wissen können Sie die Reihenfolge der Prozesse in einem ARIMA-Modell festlegen. Spezieller

- **Autokorrelationsfunktion (ACF).** Mit Lag *k* ist dies die Korrelation zwischen Reihenwerten, die *k* Intervalle entfernt sind.
- **Partielle Autokorrelationsfunktion (PACF).** Mit Lag *k* ist dies die Korrelation zwischen Reihenwerten, die *k* Intervalle entfernt sind, wobei die Werte der dazwischenliegenden Intervalle berücksichtigt werden.

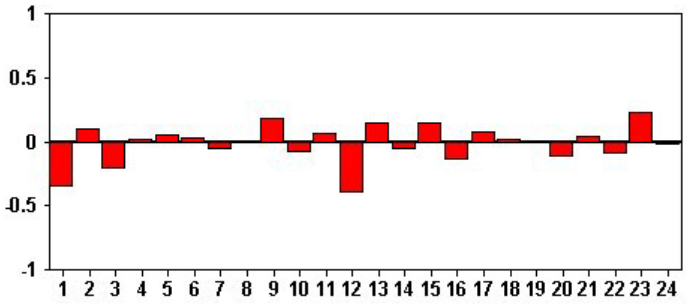


Abbildung 58. ACF-Plot für eine Reihe

Die *X*-Achse des ACF-Plots gibt den Lag an, bei dem die Autokorrelation berechnet wird. Die *Y*-Achse gibt den Wert der Korrelation an (zwischen –1 und 1). Beispiel: Eine Spitze bei Lag 1 in einem ACF-Plot zeigt eine starke Korrelation zwischen allen Reihenwerten und dem vorherigen Wert an. Eine Spitze bei Lag 2 zeigt eine starke Korrelation zwischen allen Werten und dem Wert an, der zwei Punkte vorher auftrat, usw.

- Eine positive Korrelation zeigt an, dass große aktuelle Werte großen Werten zum angegebenen Abstand entsprechen. Eine negative Korrelation zeigt an, dass große aktuelle Werte kleinen Werten zum angegebenen Abstand entsprechen.
- Der absolute Wert einer Korrelation ist ein Maßstab für die Stärke der Verbindung, wobei größere absolute Werte eine stärkere Beziehung anzeigen.

Reihentransformationen

Transformationen sind häufig nützlich, um eine Reihe zu stabilisieren, bevor Modelle geschätzt werden. Dies ist insbesondere für ARIMA-Modelle wichtig, für die eine Reihe **feststehend** sein muss, bevor Modelle geschätzt werden. Eine Reihe ist feststehend, wenn die globale Ebene (Mittelwert) und die durchschnittliche Abweichung von der Ebene (Varianz) über die Reihen hinweg konstant sind.

Obwohl die meisten Reihen nicht feststehend sind, ist ARIMA effektiv, solange Reihen durch Transformationen, wie natürlicher Logarithmus, Differenz oder saisonale Differenz, feststehend gemacht werden können.

Varianz stabilisierende Transformation. Reihen, in denen sich die Varianz mit der Zeit ändert, können häufig mit einer natürlichen Logarithmus- oder einer Quadratwurzel-Transformation feststehend gemacht werden. Diese werden auch als funktionale Transformationen bezeichnet.

- **Natürlicher Logarithmus.** Der natürliche Logarithmus wird auf Reihenwerte angewendet.
- **Quadratwurzel.** Die Quadratwurzelfunktion wird auf die Reihenwerte angewendet.

Natürliche Logarithmus- und Quadratwurzel-Transformationen können für Reihen mit negativen Werten verwendet werden.

Ebene stabilisierende Transformationen. Ein langsamer Rückgang von Werten in der ACF zeigt an, dass alle Reihenwerte eine starke Korrelation mit dem vorherigen Wert besitzen. Durch die Analyse der Veränderung der Reihenwerte erhalten Sie eine stabile Ebene.

- **Einfache Differenzbildung.** Die Differenzen zwischen den einzelnen Werten und dem vorherigen Wert in der Reihe werden berechnet, wobei der älteste Wert der Reihe ausgenommen wird. Dies bedeutet, dass die Differenzreihen einen Wert weniger als die Originalreihen besitzen.
- **Saisonale Differenzbildung.** Identisch mit der einfachen Differenz, außer, dass die Differenzen zwischen den einzelnen Werten und den vorherigen saisonalen Werten berechnet werden.

Wenn entweder die einfache oder die saisonale Differenz gleichzeitig mit der Log- oder Quadratwurzel-Transformation eingesetzt wird, wird immer zuerst die Varianz stabilisierende Transformation angewendet. Wenn sowohl die einfache als auch die saisonale Differenz angewendet wird, erhalten Sie, unabhängig von der Reihenfolge, in der die Differenz gebildet wird, immer dieselben resultierenden Reihenwerte.

Prädiktorreihen

Prädiktorreihen enthalten verwandte Daten, die das Ergebnis von Reihen erklären können, für die eine Vorhersage erstellt werden soll. Beispiel: Ein über das Internet oder mit Katalogen arbeitender Einzelhändler kann den Absatz anhand der versendeten Kataloge, der Anzahl verfügbarer Telefonverbindungen oder der Anzahl der Klicks auf die Website des Unternehmens vorhersagen.

Alle Reihen können als Prädiktor verwendet werden, wenn die Reihen so weit in die Zukunft reichen, wie sie eine Vorhersage erstellen möchten, und wenn vollständige Daten ohne fehlende Werte vorliegen.

Seien Sie vorsichtig, wenn Sie Prädiktoren zu einem Modell hinzufügen. Wenn eine große Anzahl von Prädiktoren hinzugefügt wird, steigt die für die Schätzung von Modellen erforderliche Zeit. Während das Hinzufügen von Prädiktoren die Anpassungsgüte des Modells für historische Daten verbessern kann, bedeutet dies nicht zwingend, dass das Modell bessere Vorhersagen liefert. Aus diesem Grund kann es sein, dass sich diese zusätzliche Komplexität nicht lohnt. Idealerweise sollte das Ziel die Identifizierung des einfachsten Modells darstellen, das eine gute Vorhersage ermöglicht.

Eine allgemeine Regel besagt, dass die Anzahl der Prädiktoren geringer sein sollte als der durch 15 geteilte Stichprobenumfang (höchstens ein Prädiktor für 15 Fälle).

Prädiktoren mit fehlenden Daten. Prädiktoren mit unvollständigen oder fehlenden Daten können nicht für Vorhersagen verwendet werden. Dies gilt sowohl für historische Daten als auch für zukünftige Werte. In einigen Fällen können Sie diese Einschränkung umgehen, indem Sie die Schätzspanne des Modells so einstellen, dass die ältesten Werte bei der Schätzung von Modellen ausgeschlossen werden.

Zeitreihen - Modellierungsknoten

Der Zeitreihenknoten berechnet Schätzungen für exponentielle Glättung, univariate ARIMA-Modelle (Autoregressive Integrated Moving Average - autoregressiver integrierter gleitender Durchschnitt) und multivariate ARIMA-Modelle (Transferfunktionsmodelle) für Zeitreihendaten und erstellt Vorhersagen auf der Grundlage der Zeitreihendaten.

Exponentielles Glätten ist eine Vorhersagemethode, bei der gewichtete Werte aus früheren Beobachtungen der Zeitreihe verwendet werden, um zukünftige Werte vorherzusagen. An sich beruht das exponentielle Glätten nicht auf einem theoretischen Verständnis der Daten. Es wird jeweils ein Punkt vorhergesagt und die Vorhersagen werden angepasst, wenn neue Daten verfügbar sind. Diese Methode ist nützlich für die Vorhersage von Zeitreihen, die Trend und/oder Saisonalität aufweisen. Dabei können Sie zwischen verschiedenen Modellen mit exponentiellem Glätten wählen, die sich hinsichtlich der Behandlung von Trends und Saisonalität unterscheiden.

ARIMA-Modelle bieten feinere Methoden für die Modellierung von Trend- und saisonalen Komponenten als die Modelle der exponentiellen Glättung und weisen insbesondere den zusätzlichen Vorteil auf, dass unabhängige Variablen (Prädiktorvariablen) in das Modell integriert werden können. Hierfür müssen die die Ordnung der Autoregression, die Ordnung des gleitenden Durchschnitts und der Grad der Differenzenbildung angegeben werden. Sie können Prädiktorvariablen einschließen und Transferfunktionen für bestimmte oder alle dieser Variablen definieren und die automatische Erkennung von Ausreißern oder einer bestimmten Gruppe von Ausreißern festlegen.

Hinweis: In der Praxis bedeutet dies, dass ARIMA-Modelle besonders nützlich sind, wenn Sie Prädiktoren einschließen möchten, die zur Erklärung des Verhaltens der vorhergesagten Reihe beitragen können, wie beispielsweise die Anzahl der versendeten Kataloge oder die Anzahl der Aufrufe einer Unternehmenswebseite. Modelle für das exponentielle Glätten beschreiben das Verhalten der Zeitreihen, ohne dass versucht wird zu verstehen, warum sich die Zeitreihe so verhält. Beispielsweise ist davon auszugehen, dass eine Zeitreihe, die bisher alle 12 Monate einen Höhepunkt erreicht hat, dies auch weiterhin tun wird, auch wenn Sie nicht wissen, warum.

Außerdem ist ein **Expert Modeler** verfügbar, der automatisch das am besten angepasste ARIMA-Modell bzw. das am besten angepasste Modell mit exponentiellem Glätten für eine oder mehrere Zielvariablen ermittelt, sodass das geeignete Modell nicht mehr nach dem Prinzip von Versuch und Irrtum ermittelt werden muss. In allen Fällen wählt der Expert Modeler jeweils das beste Modell für jede der angegebenen Zielvariablen. Im Zweifelsfall sollte der Expert Modeler verwendet werden.

Bei Angabe von Prädiktorvariablen wählt der Expert Modeler diejenigen Variablen zum Einschluss in ARIMA-Modelle aus, die eine statistisch signifikante Beziehung mit der abhängigen Zeitreihe aufweisen. Modellvariablen werden gegebenenfalls durch Differenzierung und/oder Quadratwurzeltransformation bzw. Transformation mit natürlichem Logarithmus transformiert. In der Standardeinstellung berücksichtigt der Expert Modeler alle Modelle für das exponentielle Glätten sowie alle ARIMA-Modelle und wählt für jedes Zielfeld das jeweils beste Modell aus. Sie können festlegen, dass der Expert Modeler nur das beste Modell mit exponentiellem Glätten oder nur das beste ARIMA-Modell auswählen soll. Sie können auch die automatische Erkennung von Ausreißern festlegen.

Beispiel. Ein Analyst eines Breitbandproviders soll Vorhersagen zu Benutzerabonnements erstellen, um die Nutzung der Bandbreite vorherzusagen. Es werden Vorhersagen für alle lokalen Märkte benötigt, die zusammen den landesweiten Kundenstamm ergeben. Mit der Zeitreihenmodellierung können Sie Vorhersagen für die nächsten drei Monate für eine Reihe lokaler Märkte erstellen.

Voraussetzungen

Der Zeitreihenknoten weicht von anderen IBM SPSS Modeler-Knoten dahingehend ab, dass Sie ihn nicht einfach in einen Stream einfügen und den Stream ausführen können. Dem Zeitreihenknoten muss stets ein Zeitintervallknoten vorangehen, der Informationen angibt wie das zu verwendende Zeitintervall (Jahre, Quartale, Monate usw.), die für die Schätzung zu verwendenden Daten und wie weit in die Zukunft sich eine Vorhersage erstrecken soll, sofern verwendet.

Die Zeitreihendaten müssen gleichmäßige Abstände aufweisen. Bei den Methoden zur Modellierung von Zeitreihendaten ist ein einheitliches Intervall zwischen den Messungen erforderlich; fehlende Werte werden durch leere Zeilen dargestellt. Falls Ihre Daten diese Bedingung nicht bereits erfüllen, können Sie sie mithilfe des Zeitintervallknotens entsprechend transformieren.

Außerdem ist bei Zeitreihendaten zu beachten:

- Die Felder müssen numerisch sein.
- Datumsfelder können nicht als Eingaben verwendet werden.
- Partitionen werden ignoriert.

Feldoptionen

Auf der Registerkarte "Felder" geben Sie an, welche Felder bei der Erstellung des Modells verwendet werden sollen. Bevor Sie ein Modell erstellen können, müssen Sie festlegen, welche Felder als Ziele und als Eingaben verwendet werden sollen. Normalerweise verwendet der Zeitreihenknoten Feldinformationen aus einem weiter oben liegenden Typknoten. Wenn Sie einen Typknoten verwenden, um Eingabe- und Zielfelder auszuwählen, brauchen Sie auf dieser Registerkarte keine Änderungen vorzunehmen.

Typknoteneinstellungen verwenden. Diese Option weist den Knoten an, die Feldinformationen von einem weiter oben liegenden Typknoten zu verwenden. Dies ist die Standardeinstellung.

Benutzerdefinierte Einstellungen verwenden. Diese Option weist den Knoten an, die hier angegebenen Feldinformationen anstelle der in einem weiter oben liegenden Typknoten angegebenen zu verwenden. Geben Sie nach Auswahl dieser Option die unten stehenden Felder an. Beachten Sie, dass als Datumswerte gespeicherte Felder nicht als Ziel- oder Eingabefelder zulässig sind.

- **Ziele.** Wählen Sie ein oder mehrere Zielfelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Ziel* festlegen. Zielfelder für Zeitreihenmodelle müssen ein Messniveau des Typs *Stetig* aufweisen. Für jedes Zielfeld wird ein separates Modell erstellt. Für Zielfelder kommen alle angegebenen *Eingabe*-Felder mit Ausnahme des jeweiligen Zielfelds selbst als mögliche Eingaben in Betracht. Daher kann dasselbe Feld in beiden Listen vorkommen; ein solches Feld wird als mögliche Eingabe für alle Modelle verwendet, außer für das Modell, bei dem es ein Zielfeld ist.
- **Eingaben.** Wählen Sie die Eingabefelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen. Eingabefelder für Zeitreihenmodelle müssen numerisch sein.

Zeitreihenmodelle - Optionen

Modellname. Gibt den Namen des Modells an, das beim Ausführen des Knotens generiert wird.

- **Auto.** Generiert den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen oder dem Namen des Modelltyps in Fällen, in denen kein Ziel angegeben ist (z. B. Clustermodelle).
- **Benutzerdefiniert.** Hier können Sie einen benutzerdefinierten Namen für das Modellnugget angeben.

Schätzung unter Verwendung bestehender Modelle fortsetzen. Wenn Sie bereits ein Zeitreihenmodell generiert haben, wählen Sie diese Option aus, um die für dieses Modell angegebenen Kriterieneinstellungen wiederzuverwenden und einen neuen Modellknoten in der Modellpalette zu generieren, anstatt ein Modell völlig neu zu erstellen. Auf diese Weise können Sie Zeit sparen, indem Sie eine erneute Schätzung durchführen und eine neue Vorhersage auf der Grundlage derselben Modelleinstellungen erstellen wie

zuvor, nur unter Verwendung aktuellerer Daten. Wenn also beispielsweise als ursprüngliches Modell für eine bestimmte Zeitreihe der lineare Trend nach Holt verwendet wurde, wird derselbe Modelltyp für erneute Schätzungen und Vorhersagen für diese Daten verwendet; das System versucht nicht erneut, den besten Modelltyp für die neuen Daten zu ermitteln. Durch die Auswahl dieser Option werden die Steuerelemente **Methode** und **Kriterien** inaktiviert. Weitere Informationen finden Sie im Thema „Erneute Schätzung und Vorhersage" auf Seite 272.

Methode. Sie haben die Wahl zwischen Expert Modeler, exponentiellem Glätten und ARIMA. Weitere Informationen finden Sie im Thema „Zeitreihen - Modellierungsknoten" auf Seite 264. Wählen Sie **Kriterien** aus, um Optionen für die ausgewählte Methode anzugeben.

- **Expert Modeler.** Wählen Sie diese Option, um den Expert Modeler zu verwenden, der automatisch das jeweils am besten angepasste Modell für die einzelnen abhängigen Zeitreihen ermittelt.
- **Exponentielles Glätten.** Mit dieser Option können Sie ein benutzerdefiniertes Modell mit exponentiellem Glätten angeben.
- **ARIMA.** Mit dieser Option können Sie ein ARIMA-Modell angeben.

Zeitintervallinformationen

Dieser Bereich des Dialogfelds enthält Informationen zu Spezifikationen für Schätzungen und Vorhersagen, die am Zeitintervallknoten vorgenommen werden. Beachten Sie, dass dieser Abschnitt nicht angezeigt wird, wenn Sie die Option **Schätzung unter Verwendung bestehender Modelle fortsetzen** auswählen.

Die erste Zeile der Informationen gibt an, ob Datensätze aus dem Modell ausgeschlossen oder als Holdouts verwendet werden.

Die zweite Zeile bietet Informationen zu den im Zeitintervallknoten angegebenen Vorhersageperioden.

Wenn in der ersten Zeile **Kein Zeitintervall definiert** steht, bedeutet dies, dass kein Zeitintervallknoten eingebunden ist. Dies führt zu einem Fehler beim Versuch, den Stream auszuführen; Sie müssen einen Zeitintervallknoten oberhalb des Zeitreihenknotens einfügen.

Sonstige Informationen

Breite der Konfidenzgrenze (%). Konfidenzintervalle werden für die Modellvorhersagen und Residuen-Autokorrelationen berechnet. Es kann ein beliebiger positiver Wert unter 100 angegeben werden. In der Standardeinstellung wird ein Konfidenzintervall von 95 % verwendet.

Maximale Anzahl an Lags in ACF- und PACF-Ausgaben. Sie können die Höchstanzahl von Intervallen festlegen, die in Tabellen und Diagrammen für Autokorrelationen und partielle Autokorrelationen angezeigt werden.

Nur Scoring-Modelle erstellen. Markieren Sie dieses Kontrollkästchen, um die im Modell gespeicherte Datenmenge zu reduzieren. Dies kann die Leistung beim Erstellen von Modellen mit extrem vielen Zeitreihen (Zehntausende) verbessern. Wenn Sie diese Option wählen, werden die Registerkarten "Modell", "Parameter" und "Residuen" nicht im Zeitreihenmodellnugget angezeigt, aber Sie können dennoch die Daten wie üblich scoren.

Zeitreihen - Expert Modeler-Kriterien

Modelltyp. Die folgenden Optionen sind verfügbar:

- **Alle Modelle.** Der Expert Modeler berücksichtigt sowohl ARIMA-Modelle als auch Modelle mit exponentiellem Glätten.
- **Nur Modelle mit exponentiellem Glätten.** Der Expert Modeler berücksichtigt nur Modelle mit exponentiellem Glätten.

- **Nur ARIMA-Modelle.** Der Expert Modeler berücksichtigt nur ARIMA-Modelle.

Expert Modeler berücksichtigt saisonale Modelle. Diese Option ist nur verfügbar, wenn für die Arbeitsdatei eine Periodizität definiert wurde. Wenn diese Option aktiviert ist, berücksichtigt der Expert Modeler sowohl saisonale als auch nicht saisonale Modelle. Wenn diese Option inaktiviert ist, berücksichtigt der Expert Modeler nur nicht saisonale Modelle.

Ereignisse und Interventionen. Mit dieser Option können Sie bestimmte Eingabefelder als Ereignis- bzw. Interventionsfelder kennzeichnen. Dadurch wird angegeben, dass das betreffende Feld Zeitreihendaten enthält, die von Ereignissen (vorhersagbare wiederkehrende Situationen, z. B. Werbeaktionen) oder Interventionen (einmalige Vorfälle, z. B. Stromausfall, Streik) betroffen sind. Der Expert Modeler berücksichtigt nur einfache Regression und nicht frei wählbare Transferfunktionen für Eingaben, die als Ereignis- bzw. Interventionsfelder gekennzeichnet sind.

Eingabefelder müssen das Messniveau *Flag*, *Nominal* oder *Ordinal* aufweisen und müssen numerisch sein (z. B. "1"/"0" und nicht "Wahr"/"Falsch" für ein Flagfeld), um in dieser Liste angezeigt zu werden. Weitere Informationen finden Sie im Thema „Impulse und Schritte” auf Seite 259.

Ausreißer

Ausreißer automatisch erkennen. In der Standardeinstellung wird keine automatische Erkennung von Ausreißern durchgeführt. Wählen Sie diese Option aus, um die automatische Erkennung von Ausreißern durchzuführen, und wählen Sie anschließend die gewünschten Ausreißertypen aus. Weitere Informationen finden Sie im Thema „Ausreißer” auf Seite 260.

Zeitreihen - Kriterien für exponentielles Glätten

Modelltyp. Modelle für das exponentielle Glätten werden als saisonal oder nicht saisonal klassifiziert[1]. Saisonale Modelle sind nur verfügbar, wenn die unter Verwendung des Zeitintervallknotens definierte Periodizität saisonal ist. Es gibt folgende saisonale Periodizitäten: zyklische Perioden, Jahre, Quartale, Monate, Tage pro Woche, Stunden pro Tag, Minuten pro Tag und Sekunden pro Tag.

- **Einfach.** Dieses Modell eignet sich für Zeitreihen ohne Trend oder Saisonalität. Der einzige relevante Glättungsparameter für dieses Modell ist das Niveau. Einfaches exponentielles Glätten weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung eins, gleitendem Durchschnitt der Ordnung eins und ohne Konstante auf.
- **Linearer Trend nach Holt.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und ohne Saisonalität. Die dafür relevanten Glättungsparameter sind Niveau und Trend, die bei diesem Modell nicht durch die Werte des jeweils anderen Parameters eingeschränkt sind. Das Holt-Modell ist allgemeiner als das Brown-Modell, die Berechnung von Schätzungen für große Zeitreihen kann allerdings mehr Zeit in Anspruch nehmen. Das exponentielle Glätten nach Holt weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung zwei und gleitendem Durchschnitt der Ordnung zwei auf.
- **Linearer Trend nach Brown.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und ohne Saisonalität. Die dafür relevanten Glättungsparameter sind Niveau und Trend. Bei diesem Modell wird jedoch davon ausgegangen, dass diese gleich sind. Das Brown-Modell ist daher ein Spezialfall des Holt-Modells. Das exponentielle Glätten nach Brown weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung zwei und gleitendem Durchschnitt der Ordnung zwei auf, wobei der Koeffizient der zweiten Ordnung des gleitenden Durchschnitts die Hälfte des quadrierten Koeffizienten für die erste Ordnung beträgt.
- **Gedämpfter Trend.** Dieses Modell eignet sich für Zeitreihen mit auslaufendem linearem Trend und ohne Saisonalität. Die dafür relevanten Glättungsparameter sind Niveau, Trend und gedämpfter Trend.

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

Gedämpftes exponentielles Glätten weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung eins, Differenzenbildung der Ordnung eins und gleitendem Durchschnitt der Ordnung zwei auf.

- **Einfach saisonal.** Dieses Modell eignet sich für Zeitreihen ohne Trend und mit einem saisonalen Effekt, der im Zeitverlauf konstant bleibt. Die dafür relevanten Glättungsparameter sind Niveau und Saison. Saisonales exponentielles Glätten weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung eins, saisonaler Differenzenbildung der Ordnung eins und den Ordnungen 1, p und p+1 für den gleitenden Durchschnitt auf, wobei p die Anzahl der Perioden in einem saisonalen Intervall ist. Für Monatsdaten gilt: $p = 12$.
- **Additives Winters-Modell.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und mit einem saisonalen Effekt, der im Zeitverlauf konstant bleibt. Die dafür relevanten Glättungsparameter sind Niveau, Trend und Saison. Das exponentielle Glätten nach dem additiven Winters-Modell weist die größte Ähnlichkeit mit einem ARIMA-Modell mit Autoregression der Ordnung null, Differenzenbildung der Ordnung eins, saisonaler Differenzenbildung der Ordnung eins und p+1 Ordnungen für den gleitenden Durchschnitt auf, wobei p die Anzahl der Perioden in einem saisonalen Intervall ist. Für Monatsdaten gilt: $p = 12$.
- **Multiplikatives Winters-Modell.** Dieses Modell eignet sich für Zeitreihen mit linearem Trend und mit einem saisonalen Effekt, der sich mit der Größenordnung der Zeitreihe ändert. Die dafür relevanten Glättungsparameter sind Niveau, Trend und Saison. Exponentielles Glätten mit dem multiplikativen Winters-Modell weist keine Ähnlichkeit zu irgendeinem ARIMA-Modell auf.

Zieltransformation. Sie können für jede abhängige Variable eine Transformation angeben, die vor deren Modellierung durchgeführt werden soll. Weitere Informationen finden Sie im Thema „Reihentransformationen” auf Seite 263.

- **Keine.** Es wird keine Transformation durchgeführt.
- **Quadratwurzel.** Es wird eine Quadratwurzeltransformation durchgeführt.
- **Natürlicher Logarithmus.** Es wird eine Transformation mit natürlichem Logarithmus durchgeführt.

Zeitreihen - ARIMA-Kriterien

Mit dem Zeitreihenknoten können Sie benutzerdefinierte nicht saisonale oder saisonale ARIMA-Modelle - auch als Box-Jenkins-Modelle bekannt - mit oder ohne festes Set von Eingabevariablen (Prädiktorvariablen) erstellen[2]. Sie können Transferfunktionen für bestimmte oder alle Eingabevariablen definieren und die automatische Erkennung von Ausreißern oder einer bestimmten Gruppe von Ausreißern festlegen.

Alle angegebenen Eingabevariablen werden explizit in das Modell aufgenommen. Im Gegensatz dazu werden beim Expert Modeler Eingabevariablen nur aufgenommen, wenn sie eine statistisch signifikante Beziehung zu der Zielvariablen aufweisen.

Modell

Über die Registerkarte "Modelle" können Sie die Struktur eines benutzerdefinierten ARIMA-Modells festlegen.

ARIMA-Ordnungen. Geben Sie Werte für die verschiedenen ARIMA-Komponenten des Modells in die entsprechenden Zellen des Strukturrasters ein. Alle Werte müssen nicht negative Ganzzahlen sein. Bei autoregressiven Komponenten und Komponenten des gleitenden Durchschnitts stellt der Wert die höchste Ordnung dar. Alle positiven niedrigeren Ordnungen werden in das Modell eingeschlossen. Wenn Sie beispielsweise 2 angeben, enthält das Modell die Ordnungen 2 und 1. Die Zellen in der Spalte "Saisonal" sind nur verfügbar, wenn für die Arbeitsdatei eine Periodizität definiert wurde.

2. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*, 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

- **Autoregressiv (p).** Die Anzahl autoregressiver Ordnungen im Modell. Autoregressive Ordnungen geben die zurückliegenden Werte der Zeitreihe an, die für die Vorhersage der aktuellen Werte verwendet werden. Eine autoregressive Ordnung von 2 gibt beispielsweise an, dass die Werte der Zeitreihe, die zwei Zeitperioden zurückliegt, für die Vorhersage der aktuellen Werte verwendet wird.
- **Differenz (d).** Gibt die Ordnung der Differenzierung an, die vor dem Schätzen der Modelle auf die Zeitreihe angewendet wurde. Differenzierung ist erforderlich, wenn Trends vorhanden sind. (Zeitreihen mit Trends sind normalerweise nicht stationär, und bei der ARIMA-Modellierung wird Stationarität angenommen.) Mithilfe der Differenzierung werden die Effekte der Trends entfernt. Die Ordnung der Differenzierung entspricht dem Grad des Trends der Zeitreihe: Differenzierung erster Ordnung erklärt lineare Trends, Differenzierung zweiter Ordnung erklärt quadratische Trends usw.
- **Gleitender Durchschnitt (q).** Die Anzahl von Ordnungen des gleitenden Durchschnitts im Modell. Ordnungen des gleitenden Durchschnitts geben an, wie Abweichungen vom Mittelwert der Zeitreihe für zurückliegende Werte zum Vorhersagen der aktuellen Werte verwendet werden. Ordnungen des gleitenden Durchschnitts von 1 und 2 geben beispielsweise an, dass beim Vorhersagen der aktuellen Werte der Zeitreihe Abweichungen vom Mittelwert der Zeitreihe von den beiden letzten Zeitperioden berücksichtigt werden sollen.

Saisonale Ordnungen. Saisonale autoregressive Komponenten, Komponenten des gleitenden Durchschnitts und Differenzierungskomponenten entsprechen im Prinzip ihren nicht saisonalen Gegenstücken. Bei saisonalen Ordnungen werden die Werte der aktuellen Zeitreihe jedoch von Werten zurückliegender Zeitreihen beeinflusst, die um eine oder mehrere saisonalen Perioden getrennt sind. Bei monatlichen Daten (saisonale Periode von 12) beispielsweise bedeutet eine saisonale Ordnung von 1, dass der Wert der aktuellen Zeitreihe durch den Zeitreihenwert beeinflusst wird, der 12 Perioden vor dem aktuellen liegt. Eine saisonale Ordnung von 1 entspricht bei monatlichen Daten einer nicht saisonalen Ordnung von 12.

Zieltransformation. Sie können für jede Zielvariable eine Transformation angeben, die vor deren Modellierung durchgeführt werden soll. Weitere Informationen finden Sie im Thema „Reihentransformationen" auf Seite 263.

- **Keine.** Es wird keine Transformation durchgeführt.
- **Quadratwurzel.** Es wird eine Quadratwurzeltransformation durchgeführt.
- **Natürlicher Logarithmus.** Es wird eine Transformation mit natürlichem Logarithmus durchgeführt.

Konstante in Modell einschließen. Der Einschluss einer Konstanten ist das Standardverfahren, sofern Sie nicht sicher wissen, dass der Gesamtmittelwert der Zeitreihe 0 ist. Bei der Anwendung von Differenzierung empfiehlt es sich, die Konstante auszuschließen.

Transferfunktionen

Auf der Registerkarte "Transferfunktionen" können Sie Transferfunktionen für einige oder alle Eingabefelder definieren. Mithilfe von Transferfunktionen können Sie angeben, auf welche Weise frühere Werte der betreffenden Felder für die Vorhersage zukünftiger Werte der Ziel-Zeitreihe verwendet werden sollen.

Die Registerkarte wird nur angezeigt, wenn Eingabefelder (bei denen die Rolle auf *Eingabe* gesetzt ist) entweder auf dem Typknoten oder auf der Registerkarte "Felder" des Zeitreihenknotens (wählen Sie **Benutzerdefinierte Einstellungen verwenden - Eingaben** aus) angegeben sind.

In der Liste oben werden alle Eingabefelder angezeigt. Die übrigen Informationen in diesem Dialogfeld hängen davon ab, welches Eingabefeld in der Liste ausgewählt wurde.

Transferfunktionsordnungen. Geben Sie Werte für die verschiedenen Komponenten der Transferfunktion in die entsprechenden Zellen des Strukturrasters ein. Alle Werte müssen nicht negative Ganzzahlen sein. Bei Zähler- und Nennerkomponenten stellt der Wert die höchste Ordnung dar. Alle positiven niedrigeren Ordnungen werden in das Modell eingeschlossen. Darüber hinaus wird die Ordnung 0 bei Zählerkomponenten immer eingeschlossen. Wenn Sie beispielsweise 2 für den Zähler angeben, enthält das Modell die

Ordnungen 2, 1 und 0. Wenn Sie 3 für den Nenner angeben, enthält das Modell die Ordnungen 3, 2 und 1. Die Zellen in der Spalte "Saisonal" sind nur verfügbar, wenn für die Arbeitsdatei eine Periodizität definiert wurde.

Zähler. Die Zählerordnung der Transferfunktion gibt an, welche zurückliegenden Werte aus der ausgewählten unabhängigen Zeitreihe (Prädiktor-Zeitreihe) zum Vorhersagen der aktuellen Werte der abhängigen Zeitreihe verwendet werden. Ein Zähler-Term von 1 gibt beispielsweise an, dass der Wert einer unabhängigen Zeitreihe, die eine Periode zurückliegt, und der aktuelle Wert der unabhängigen Zeitreihe zum Vorhersagen des aktuellen Werts der einzelnen abhängigen Zeitreihen verwendet werden.

Nenner. Die Nennerordnung der Transferfunktion gibt an, wie Abweichungen vom Mittelwert der Zeitreihe für zurückliegende Werte der ausgewählten unabhängigen Zeitreihe (Prädiktor-Zeitreihe) zum Vorhersagen der aktuellen Werte der abhängigen Zeitreihe verwendet werden. Ein Nenner-Term von 1 gibt beispielsweise an, dass beim Vorhersagen der aktuellen Werte für die einzelnen abhängigen Zeitreihen Abweichungen vom Mittelwert einer unabhängigen Zeitreihe berücksichtigt werden sollen, die eine Zeitperiode zurückliegt.

Differenz. Gibt die Ordnung der Differenzierung an, die vor dem Schätzen der Modelle auf die ausgewählte unabhängige Zeitreihe (Prädiktoren) angewendet wurde. Wenn Trends vorhanden sind, ist die Differenzierung erforderlich, um die Effekte der Trends zu entfernen.

Saisonale Ordnungen. Saisonale Zähler-, Nenner- und Differenzierungskomponenten entsprechen im Prinzip ihren nicht saisonalen Gegenstücken. Bei saisonalen Ordnungen werden die Werte der aktuellen Zeitreihe jedoch von Werten zurückliegender Zeitreihen beeinflusst, die um eine oder mehrere saisonalen Perioden getrennt sind. Bei monatlichen Daten (saisonale Periode von 12) beispielsweise bedeutet eine saisonale Ordnung von 1, dass der Wert der aktuellen Zeitreihe durch den Zeitreihenwert beeinflusst wird, der 12 Perioden vor dem aktuellen liegt. Eine saisonale Ordnung von 1 entspricht bei monatlichen Daten einer nicht saisonalen Ordnung von 12.

Verzögerung Wenn eine Verzögerung festgelegt wird, verzögert sich der Einfluss des Eingabefelds um die Anzahl der angegebenen Intervalle. Bei einer Verzögerung mit dem Wert 5 beeinflusst der Wert des Eingabefelds zum Zeitpunkt t die Vorhersagen erst nach dem Ablauf von fünf Perioden ($t + 5$).

Transformation. Die Angabe einer Transferfunktion für ein Set von unabhängigen Variablen enthält auch eine optionale Transformation, die für diese Variablen ausgeführt werden soll.

- **Keine.** Es wird keine Transformation durchgeführt.
- **Quadratwurzel.** Es wird eine Quadratwurzeltransformation durchgeführt.
- **Natürlicher Logarithmus.** Es wird eine Transformation mit natürlichem Logarithmus durchgeführt.

Umgang mit Ausreißern

Auf der Registerkarte "Ausreißer" ist eine Reihe von Möglichkeiten für die Behandlung von Ausreißern in den Daten verfügbar[3].

Ausreißer nicht erkennen oder modellieren. In der Standardeinstellung werden Ausreißer weder erkannt noch modelliert. Wählen Sie diese Option aus, um die Erkennung und Modellierung von Ausreißern zu inaktivieren.

Ausreißer automatisch erkennen. Wählen Sie diese Option aus, um eine automatische Erkennung von Ausreißern durchzuführen, und wählen Sie mindestens einen der gezeigten Ausreißertypen aus.

Typ der zu ermittelnden Ausreißer. Wählen Sie die Ausreißertypen aus, die erkannt werden sollen. Folgende Typen werden unterstützt:

3. Pena, D., G. C. Tiao, and R. S. Tsay, eds. 2001. *A course in time series analysis*. New York: John Wiley and Sons.

- Additiv (Standard)
- Verschiebung im Niveau (Standard)
- Innovativ
- Transient
- Saisonal additiv
- Lokaler Trend
- Additiver Bereich

Weitere Informationen finden Sie im Thema „Ausreißer" auf Seite 260.

Generieren von Zeitreihenmodellen

Dieser Abschnitt bietet einige allgemeine Informationen über bestimmte Aspekte beim Generieren von Zeitreihenmodellen:

- Generieren mehrerer Modelle
- Verwenden von Zeitreihenmodellen bei der Vorhersageerstellung
- Erneute Schätzung und Vorhersage

Das generierte Modellnugget wird in einem separaten Kapitel beschrieben. Weitere Informationen finden Sie im Thema „Zeitreihenmodellnugget" auf Seite 272.

Generieren mehrerer Modelle

Bei der Zeitreihenmodellierung in IBM SPSS Modeler wird für jedes Zielfeld ein einzelnes Modell (entweder ARIMA oder exponentielle Glättung) erstellt. Wenn mehrere Zielfelder vorhanden sind, generiert IBM SPSS Modeler also mehrere Knoten in einem einzigen Vorgang. Dadurch wird Zeit gespart und Sie erhalten gleichzeitig die Möglichkeit, die Einstellungen für die einzelnen Modelle zu vergleichen.

Wenn Sie ein ARIMA-Modell und ein Modell mit exponentiellem Glätten für dasselbe Zielfeld vergleichen möchten, können Sie den Zeitreihenknoten separat ausführen und jeweils ein anderes Modell angeben.

Verwenden von Zeitreihenmodellen bei der Vorhersageerstellung

Bei der Erstellung von Zeitreihen wird eine bestimmte Reihe von geordneten Fällen, die sogenannte Schätzungsspanne, verwendet, um ein Modell zu erstellen, das zur Vorhersage der zukünftigen Werte der Zeitreihe verwendet werden kann. Dieses Modell enthält Informationen zur verwendeten Zeitspanne, einschließlich des Intervalls. Um mithilfe dieses Modells Vorhersagen zu erstellen, müssen für Zielvariablen und Prädiktorvariablen jeweils dieselbe Zeitspanne und dieselben Intervallinformationen mit derselben Zeitreihe verwendet werden.

Beispiel: Angenommen, Anfang Januar möchten Sie den monatlichen Absatz von Produkt 1 für die ersten drei Monate des Jahres vorhersagen. Sie erstellen ein Modell unter Verwendung der tatsächlichen monatlichen Absatzdaten für Produkt 1 von Januar bis Dezember des Vorjahrs ("Jahr 1"), wobei Sie als Zeitintervall "Monate" verwenden. Anschließend können Sie mit diesem Modell den Absatz von Produkt 1 für die ersten drei Monate von Jahr 2 prognostizieren.

Im Prinzip könnten Sie eine Vorhersage für beliebig viele Monate erstellen, allerdings lässt die Effektivität des Modells immer mehr nach, je weiter sich die Vorhersage in die Zukunft erstreckt. Es wäre jedoch nicht möglich, eine Vorhersage für die ersten drei Wochen von Jahr 2 zu erstellen, da für die Erstellung des Modells das Intervall "Monate" verwendet wurde. Ebenso wäre es sinnlos, dieses Modell für die Vorhersage des Verkaufs von Produkt 2 zu verwenden. Ein Zeitreihenmodell ist nur für die Daten relevant, mit denen es definiert wurde.

Erneute Schätzung und Vorhersage

Die Schätzperiode ist untrennbar mit dem generierten Modell verbunden. Daher werden bei Anwendung des aktuellen Modells auf neue Daten alle Werte außerhalb der Schätzperiode ignoriert. Zeitreihenmodelle müssen also jedes Mal, wenn neue Daten verfügbar sind, neu geschätzt werden; andere IBM SPSS Modeler-Modelle können dagegen zu Scoring-Zwecken ohne Veränderungen erneut angewendet werden.

Setzen wir das vorherige Beispiel fort: Angenommen, Anfang April von Jahr 2 liegen die tatsächlichen monatlichen Absatzdaten für Januar bis März vor. Wenn Sie jedoch das Modell, das Sie Anfang Januar erstellt haben, erneut anwenden, wird erneut eine Vorhersage für Januar bis März erstellt und die bekannten Absatzdaten für diesen Zeitraum werden ignoriert.

Die Lösung besteht darin, ein neues Modell auf der Grundlage der aktualisierten Ist-Daten zu erstellen. Sofern Sie die Vorhersageparameter nicht ändern, kann das neue Modell zur Vorhersage der nächsten drei Monate, April bis Juni, verwendet werden. Wenn Sie noch Zugriff auf den Stream haben, der zur Erzeugung des ursprünglichen Modells verwendet wurde, können Sie einfach den Verweis auf die Quellendatei in diesem Stream durch einen Verweis auf die Datei ersetzen, die die aktualisierten Daten enthält, und den Stream dann erneut ausführen, um das neue Modell zu generieren. Wenn Sie nur noch das ursprüngliche Modell als Datei besitzen, können Sie damit einen Zeitreihenknoten erstellen, den Sie anschließend einem neuen Stream hinzufügen können, der einen Verweis zur aktualisierten Quellendatei enthält. Vorausgesetzt, dieser neue Stream stellt dem Zeitreihenknoten einen Zeitintervallknoten voran, bei dem das Intervall auf "Monate" gesetzt ist, dann wird durch das Ausführen dieses neuen Streams das erforderliche neue Modell erstellt.

Zeitreihenmodellnugget

Die Zeitreihenmodellierung erstellt eine Reihe neuer Felder mit dem Präfix $TS-, wie in der folgenden Tabelle dargestellt ist.

Tabelle 23. Von der Zeitreihenmodellierungsoperation erstellte neue Felder.

Feldname	Beschreibung
$TS-*Spaltenname*	Der vom Modell für die einzelnen Ziel-Zeitreihen vorhergesagte Wert.
$TSLCI-*Spaltenname*	Die unteren Konfidenzintervalle für die einzelnen vorhergesagten Zeitreihen.*
$TSUCI-*Spaltenname*	Die oberen Konfidenzintervalle für die einzelnen vorhergesagten Zeitreihen.*
$TSNR-*Spaltenname*	Der Wert des Restrauschens für die einzelnen Spalten der Daten des generierten Modells.*
$TS-Total	Der Gesamtwert der $TS-*Spaltenname*-Werte für die betreffende Zeile.
$TSLCI-Total	Der Gesamtwert der $TSLCI-*Spaltenname*-Werte für die betreffende Zeile.*
$TSUCI-Total	Der Gesamtwert der $TSUCI-*Spaltenname*-Werte für die betreffende Zeile.*
$TSNR-Total	Der Gesamtwert der $TSNR-*Spaltenname*-Werte für die betreffende Zeile.*

* Die Sichtbarkeit dieser Felder (z. B. in der Ausgabe von einem angegliederten Tabellenknoten) hängt von Optionen auf der Registerkarte "Einstellungen" des Zeitreihenmodellnuggets ab. Weitere Informationen finden Sie im Thema „Zeitreihen - Modelleinstellungen” auf Seite 276.

Das Modellnugget vom Typ "Zeitreihe" zeigt Details der verschiedenen Modelle an, die für die einzelnen Zeitreiheneingaben im Erstellungsknoten der Zeitreihe ausgewählt wurden. Die Eingabe mehrerer Zeitreihen (z. B. Daten zu Produktlinien, Regionen oder Filialen) ist möglich und für jede Ziel-Zeitreihe wird ein separates Modell erstellt. Wenn der Umsatz in der Region Ost beispielsweise für ein ARIMA-Modell geeignet ist, die Region West sich jedoch nur für einen einfachen gleitenden Durchschnitt eignet, wird jede Region mit dem entsprechenden Modell gescort.

Die Standardausgabe zeigt für die einzelnen erstellten Modelle jeweils den Modelltyp, die Anzahl der angegebenen Prädiktoren und das Maß für die Anpassungsgüte (Standard: stationäres *R*-Quadrat) an. Wenn Sie Ausreißermethoden angegeben haben, ist eine Spalte vorhanden, in der die Anzahl der ermittelten Ausreißer angezeigt wird. Die Standardausgabe beinhaltet außerdem Spalten für Ljung-Box Q, Freiheitsgrade und Signifikanzwerte.

Außerdem können Sie die erweiterte Ausgabe auswählen; hierbei werden zusätzlich folgende Spalten angezeigt:

- R-Quadrat
- RMSE (Root Mean Square Error - Wurzel des mittleren quadratischen Fehlers)
- MAPE (Mean Absolute Percentage Error - mittlerer absoluter Fehler in Prozent)
- MAE (Mean Absolute Error - mittlerer absoluter Fehler)
- MaxAPE (Maximum Absolute Percentage Error - maximaler absoluter Fehler in Prozent)
- MaxAE (Maximum Absolute Error - maximaler absoluter Fehler)
- Norm. BIC (Normalized Bayesian Information Criterion - normalisiertes bayessches Informationskriterium)

Generieren. Ermöglicht die Erzeugung eines Zeitreihenmodellierungsknotens im Stream oder eines Modellnuggets in der Palette.

- **Modellierungsknoten generieren.** Platziert einen Zeitreihenmodellierungsknoten in einen Stream, wobei die Einstellungen übernommen werden, die zum Erstellen dieses Modellsets verwendet wurden. Dies ist beispielsweise dann sinnvoll, wenn Sie einen Stream haben, in dem Sie diese Modelleinstellungen verwenden möchten, aber nicht mehr über den Modellierungsknoten verfügen, mit dem Sie sie generiert haben.
- **Modell in Palette.** Platziert ein Modellnugget mit allen Zielen im Modell-Manager.

Modell

Abbildung 59. Schaltflächen "Alles markieren" und "Alle Markierungen aufheben"

Kontrollkästchen. Wählen Sie aus, welche Modelle Sie beim Scoren verwenden möchten. Standardmäßig sind alle Kontrollkästchen aktiviert. Mit den Schaltflächen **Alles markieren** und **Alle Markierungen aufheben** werden alle Kontrollkästchen in einem einzigen Vorgang bearbeitet.

Sortieren nach. Mit dieser Option können Sie die Ausgabezeilen in aufsteigender oder absteigender Reihenfolge einer bestimmten Spalte der Anzeige sortieren. Mit der Option "Ausgewählt" wird die Ausgabe anhand einer oder mehrerer Zeilen sortiert, die über Kontrollkästchen ausgewählt wurden. Dies ist beispielsweise sinnvoll, um Zielfelder wie "Markt_1" bis "Markt_9" vor "Markt_10" anzeigen zu lassen; bei der standardmäßigen Sortierreihenfolge wird nämlich "Markt_10" unmittelbar nach "Markt_1" angezeigt.

Ansicht. In der Standardansicht ("Einfach") wird die Grundmenge der Ausgabespalten angezeigt. Bei der Option "Erweitert" werden zusätzliche Spalten für die Maße der Anpassungsgüte angezeigt.

Anzahl der bei der Schätzung verwendeten Datensätze. Die Anzahl der Zeilen in der ursprünglichen Quellendatendatei.

Ziel. Die im Typknoten als Zielfelder (mit der Rolle *Ziel*) gekennzeichneten Felder.

Modell. Der für dieses Zielfeld verwendete Modelltyp.

Prädiktoren. Die Anzahl der für dieses Zielfeld verwendeten Prädiktoren (mit der Rolle *Eingabe*).

Ausreißer. Diese Spalte wird nur angezeigt, wenn Sie die automatische Erkennung von Ausreißern angefordert haben (bei den Expert Modeler- oder ARIMA-Kriterien). Der angezeigte Wert gibt die Anzahl der ermittelten Ausreißer an.

R-Quadrat für stationären Teil. Ein Maß, das den stationären Teil des Modells mit einem einfachen Mittelwertmodell vergleicht. Dieses Maß ist dem gewöhnlichen R-Quadrat vorzuziehen, wenn ein Trend oder ein saisonales Muster vorliegt. R-Quadrat für den stationären Teil kann auch negativ sein, es nimmt Werte zwischen minus unendlich und 1 an. Negative Werte bedeuten, dass das betrachtete Modell schlechter ist als das Basismodell. Positive Werte bedeuten, dass das betrachtete Modell besser ist als das Basismodell.

R-Quadrat. Ein Maß für die Anpassungsgüte eines linearen Modells. Wird auch als Bestimmtheitskoeffizient bezeichnet. Es gibt den Anteil der Variation der abhängigen Variablen an, der durch das Regressionsmodell erklärt wird. Er liegt zwischen 0 und 1. Kleine Werte zeigen an, dass das Modell nicht gut zu den Daten passt.

RMSE. Steht für Root Mean Square Error. Die Quadratwurzel des mittleren quadratischen Fehlers. Ein Maß dafür, wie stark eine abhängige Zeitreihe von ihrem durch das Modell vorhergesagten Niveau abweicht, und zwar ausgedrückt in derselben Maßeinheit wie die abhängige Zeitreihe.

MAPE. Mittlerer absoluter Fehler in Prozent. Ein Maß dafür, wie stark eine abhängige Zeitreihe von ihrem durch das Modell vorhergesagten Niveau abweicht. Es ist unabhängig von den verwendeten Maßeinheiten und kann daher verwendet werden, um Zeitreihen mit unterschiedlichen Einheiten zu vergleichen.

MAE. Mittlerer absoluter Fehler. Er misst, wie stark die Zeitreihe von ihrem durch das Modell vorhergesagten Niveau abweicht. MAE wird in derselben Maßeinheit angegeben wie die ursprüngliche Zeitreihe.

MaxAPE. Maximaler absoluter Fehler in Prozent (Maximum Absolute Percentage Error, also maximaler Betrag des relativen Fehlers). Dies ist der größte vorhergesagte Fehler, ausgedrückt in Prozent. Dieses Maß hilft dabei, sich ein Worst-Case-Szenario für die Vorhersagen vorzustellen.

MaxAE. Maximaler absoluter Fehler (Maximum Absolute Error, also maximaler Betrag des Fehlers). Dies ist der größte vorhergesagte Fehler, ausgedrückt in derselben Maßeinheit wie die abhängige Zeitreihe. Genau wie MaxAPE hilft er dabei, sich ein Worst-Case-Szenario für die Vorhersagen vorzustellen. Der maximale absolute Fehler und der maximale absolute Fehler in Prozent können an verschiedenen Punkten in der Zeitreihe auftreten, beispielsweise wenn der absolute Fehler für einen großen Zeitreihenwert geringfügig größer ist als der absolute Fehler für einen kleinen Zeitreihenwert. In diesem Fall tritt der maximale absolute Fehler beim größeren Zeitreihenwert und der maximale absolute Fehler in Prozent beim kleineren Zeitreihenwert auf.

Normalisiertes BIC. Normalisiertes Bayes-Informationskriterium (BIC). Ein allgemeines Maß der insgesamt erreichten Güte der Anpassung, das auch die Komplexität des Modells zu berücksichtigen versucht. Es ist ein Score, der auf dem mittleren quadratischen Fehler beruht und eine Penalisierung für die Anzahl der Modellparameter und die Länge der Zeitreihe enthält. Die Penalisierung neutralisiert die Überlegenheit von Modellen mit einer größeren Anzahl von Parametern und macht die Statistik damit gut vergleichbar für verschiedene Modelle derselben Zeitreihe.

Q. Die Ljung-Box-Q-Statistik. Ein Test der Zufälligkeit der Restfehler in diesem Modell.

df. Freiheitsgrade. Die Anzahl der Modellparameter, die bei der Schätzung eines bestimmten Ziels frei variieren können.

Sig. Signifikanzwert der Ljung-Box-Statistik. Ein Signifikanzwert von weniger als 0,05 deutet darauf hin, dass die Restfehler nicht zufällig sind.

Auswertungsstatistik. Dieser Abschnitt enthält verschiedene Auswertungsstatistiken für die verschiedenen Spalten, darunter Mittelwert, Minimum, Maximum und Perzentilwerte.

Zeitreihen - Modellparameter

Auf der Registerkarte "Parameter" sind Details verschiedener Parameter aufgeführt, die zum Erstellen eines ausgewählten Modells verwendet wurden.

Parameter für Modell anzeigen. Wählen Sie das Modell aus, für das die Parameterdetails angezeigt werden sollen.

Ziel. Der Name des von diesem Modell vorhergesagten Zielfelds (mit der Rolle *Ziel*).

Modell. Der für dieses Zielfeld verwendete Modelltyp.

Feld (nur ARIMA-Modelle). Enthält genau einen Eintrag für jede der im Modell verwendeten Variablen. Dabei wird zunächst das Ziel angegeben und dann die Prädiktoren (sofern vorhanden).

Transformation. Zeigt gegebenenfalls an, welche Art von Transformation für dieses Feld festgelegt wurde, bevor das Modell erstellt wurde.

Parameter. Der Modellparameter, für den die folgenden Details angezeigt werden:

- **Lag (nur ARIMA-Modelle).** Gibt gegebenenfalls die für diesen Parameter im Modell berücksichtigten Lags an.
- **Schätzung.** Die Parameterschätzung. Dieser Wert wird bei der Berechnung des Vorhersagewerts und der Konfidenzintervalle für das Zielfeld verwendet.
- **SE.** Der Standardfehler (Standard Error) der Parameterschätzung.
- **t.** Der Wert der Parameterschätzung dividiert durch den Standardfehler.
- **Sig.** Das Signifikanzniveau der Parameterschätzung. Werte über 0,05 werden als statistisch nicht signifikant betrachtet.

Zeitreihen - Modellresiduen

Auf der Registerkarte "Residuen" werden die Autokorrelationsfunktion (ACF) und die partielle Autokorrelationsfunktion (PACF) der Residuen (der Differenzen zwischen den erwarteten Werten und den Ist-Werten) für die einzelnen erstellten Modelle angezeigt. Weitere Informationen finden Sie im Thema „Autokorrelation und partielle Autokorrelationsfunktionen” auf Seite 262.

Diagramm für Modell anzeigen. Wählen Sie das Modell, für das Sie die die Residuen-ACF und -PACF anzeigen möchten.

Zeitreihen - Modellübersicht

Auf der Registerkarte "Übersicht" eines Modellnuggets finden Sie Informationen zum Modell selbst (*Analyse*), den im Modell verwendeten Feldern (*Felder*), den beim Erstellen des Modells verwendeten Einstellungen (*Aufbaueinstellungen*) und zum Modelltraining (*Trainingsübersicht*).

Beim ersten Durchsuchen des Knotens sind die Ergebnisse der Registerkarte "Übersicht" reduziert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie sie mithilfe des Erweiterungssteuerelements links neben dem betreffenden Element oder klicken Sie auf die Schaltfläche **Alles anzeigen**, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement die gewünschten Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche **Alles ausblenden** alle Ergebnisse ausblenden.

Analyse. Zeigt Informationen zum jeweiligen Modell an.

Felder. Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden.

Erstellungseinstellungen. Enthält Informationen zu den beim Aufbau des Modells verwendeten Einstellungen.

Trainingsübersicht. In diesem Abschnitt werden folgende Informationen angezeigt: der Typ des Modells, der für seine Erstellung verwendete Stream, der Benutzer, der ihn erstellt hat, der Erstellungszeitpunkt und die für den Aufbau des Modells benötigte Zeit.

Zeitreihen - Modelleinstellungen

Auf der Registerkarte "Einstellungen" können Sie angeben, welche zusätzlichen Felder durch den Modellierungsvorgang erstellt werden sollen.

Neue Felder für jedes zu scorende Modell erstellen. Mit dieser Option können Sie die neuen Felder angeben, die für jedes zu scorende Modell erstellt werden sollen.

- **Obere und untere Konfidenzgrenzen berechnen.** Wenn diese Option aktiviert ist, werden neue Felder (mit den Standardpräfixen $TSLCI- und $TSUCI-) für die obere bzw. untere Grenze des Konfidenzintervalls für die einzelnen Zielfelder erstellt, zusammen mit den Gesamtsummen dieser Werte.
- **Restrauschen berechnen.** Wenn diese Option aktiviert ist, wird ein neues Feld (mit dem Standardpräfix $TSNR-) für die Modellresiduen der einzelnen Zielfelder erstellt, zusammen mit einer Gesamtsumme dieser Werte.

Kapitel 14. Lernfähige Antwortknotenmodelle

SLRM-Knoten

Mit dem **Knoten für das lernfähige Antwortmodell** (SLRM - Self-Learning Response Model) können Sie ein Modell erstellen, das Sie während der Erweiterung des Datasets ständig aktualisieren bzw. neu schätzen können, ohne dass Sie das Modell jedes Mal anhand des vollständigen Datasets neu erstellen müssen. Dies ist beispielsweise dann nützlich, wenn Sie mehrere Produkte führen und ermitteln möchten, welches Produkt ein Kunde mit der größten Wahrscheinlichkeit kauft, wenn Sie es ihm anbieten. Mit diesem Modell können Sie prognostizieren, welche Angebote für die Kunden am geeignetsten sind und mit welcher Wahrscheinlichkeit die Angebote angenommen werden.

Das Modell kann zunächst mit einem kleinen Dataset mit zufällig ausgewählten Angeboten und den Reaktionen auf diese Angebote erstellt werden. Wenn das Dataset größer wird, kann das Modell aktualisiert werden, wodurch es besser in der Lage ist, die geeignetsten Angebote für Kunden und die Wahrscheinlichkeit, dass diese angenommen werden, auf der Grundlage anderer Eingabefelder, wie Alter, Geschlecht, Beruf und Einkommen, zu prognostizieren. Die verfügbaren Angebote können geändert werden, indem sie im Knotendialogfeld hinzugefügt bzw. entfernt werden. Es ist also nicht erforderlich, das Zielfeld des Datasets zu ändern.

Bei einer Kombination mit IBM SPSS Collaboration and Deployment Services können Sie automatische regelmäßige Aktualisierungen des Modells einrichten. Dieser Vorgang, der keine Überwachung oder Eingriffe von menschlicher Seite erfordert, stellt eine flexible und kostengünstige Lösung für Organisationen und Anwendungen dar, bei denen individuelle Eingriffe durch einen Data-Mining-Experten nicht möglich oder erforderlich sind.

Beispiel. Ein Finanzinstitut möchte profitablere Ergebnisse erzielen, indem jedem Kunden das Angebot zugeordnet wird, das mit der größten Wahrscheinlichkeit angenommen wird. Mit einem lernfähigen Modell können Sie basierend auf früheren Werbeaktionen Merkmale von Kunden ermitteln, die mit hoher Wahrscheinlichkeit positiv antworten werden. Sie können damit das Modell auch basierend auf den aktuellsten Kundenantworten in Echtzeit aktualisieren.

Feldoptionen für den SLRM-Knoten

Vor der Ausführung eines SLRM-Knotens müssen Sie die Ziel- und Zielantwortfelder auf der Registerkarte "Felder" des Knotens angeben.

Zielfeld. Wählen Sie das Zielfeld aus der Liste, z. B. ein nominales Feld (Setfeld) , das die verschiedenen Produkte enthält, die Sie den Kunden anbieten möchten.

Hinweis: Das Zielfeld muss den Speichertyp "Zeichenfolge", nicht "Numerisch" aufweisen.

Zielantwortfeld. Wählen Sie das Zielantwortfeld aus der Liste aus. Beispiele: "Angenommen" oder "Abgelehnt".

Hinweis: Dieses Feld muss ein Flagfeld sein. Der Wahr-Wert des Flag zeigt die Annahme, der Falsch-Wert die Ablehnung des Angebots an.

Die restlichen Felder in diesem Dialogfeld sind Standardfelder, die überall in IBM SPSS Modeler verwendet werden. Weitere Informationen finden Sie im Thema „Feldoptionen der Modellierungsknoten” auf Seite 31.

Hinweis: Wenn die Quellendaten Bereiche enthalten, die als stetige (numerischer Bereich) Eingabefelder verwendet werden sollen, müssen Sie sicherstellen, dass die Metadaten die Details für Maximum und Minimum für jeden Bereich enthalten.

Modelloptionen für den SLRM-Knoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Training des bestehenden Modells fortsetzen. In der Standardeinstellung wird bei jeder Ausführung eines Modellierungsknotens ein völlig neuees Modell erstellt. Bei Auswahl dieser Option wird das Training mit dem letzten, vom Knoten erfolgreich aufgebauten Modell fortgesetzt. Dies ermöglicht die Aktualisierung eines bestehenden Modells, ohne dass ein Zugriff auf die ursprünglichen Daten erforderlich ist, und kann zu einer wesentlich schnelleren Leistung führen, da *ausschließlich* die neuen bzw. aktualisierten Datensätze in den Stream eingespeist werden. Details zum vorherigen Modell werden zusammen mit dem Modellierungsknoten gespeichert, wodurch diese Option auch dann verwendet werden kann, wenn das vorherige Modellnugget nicht mehr im Stream oder in der Modellpalette verfügbar ist.

Werte des Zielfelds Standardmäßig ist dieses Feld auf **Alle verwenden** gesetzt, was bedeutet, dass ein Modell erstellt wird, das alle Angebote enthält, die dem ausgewählten Zielfeldwert zugeordnet sind. Wenn Sie ein Modell erstellen möchten, das nur einige der Angebote des Zielfelds enthält, klicken Sie auf **Angeben** und verwenden Sie die Schaltflächen **Hinzufügen**, **Bearbeiten** und **Löschen**, um die Namen der Angebote, für die ein Modell erstellt werden soll, einzugeben bzw. abzuändern. Wenn Sie beispielsweise ein Ziel ausgewählt haben, das alle von Ihnen gelieferten Produkte auflistet, können Sie die angebotenen Produkte mit diesem Feld auf einige wenige einschränken, die Sie hier eingeben.

Modellauswertung. Die Felder in diesem Fenster sind dahingehend unabhängig vom Modell, dass sie das Scoring nicht beeinflussen. Stattdessen ermöglichen Sie Ihnen, eine visuelle Darstellung davon zu erstellen, wie gut das Modell Ergebnisse vorhersagt.

Hinweis: Um die Ergebnisse der Modellauswertung im Modellnugget anzuzeigen, müssen Sie auch das Feld **Modellevaluation anzeigen** aktivieren.

- **Modellauswertung einschließen.** Wählen Sie dieses Feld aus, um Diagramme zu erstellen, die die prognostizierte Genauigkeit des Modells für jedes ausgewählte Angebot zeigen.
- **Startwert für Zufallsgenerator festlegen.** Bei der Schätzung der Genauigkeit eines Modells auf der Grundlage eines Zufallsprozentsatzes können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Gewünschten Startwert eingeben. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.
- **Simulierte Stichprobengröße.** Geben Sie die Anzahl der Datensätze an, die bei der Bewertung des Modells in der Stichprobe verwendet werden sollen. Der Standardwert ist 100.
- **Anzahl der Iterationen.** Mit dieser Option können Sie die Erstellung der Modellauswertung nach der angegebenen Anzahl an Iterationen beenden. Dient zur Angabe der maximalen Anzahl der Iterationen; der Standardwert lautet "20".

  Hinweis: Beachten Sie, dass große Stichproben und eine hohe Anzahl von Iterationen den Zeitbedarf für die Modellerstellung erhöhen.

Modellevaluation anzeigen. Wählen Sie diese Option, um eine grafische Darstellung der Ergebnisse im Modellnugget anzuzeigen.

Einstellungsoptionen für den SLRM-Knoten

Mit den Einstellungsoptionen für den Knoten können Sie die Feinabstimmung der Modellerstellung vornehmen.

Maximale Anzahl an Vorhersagen pro Datensatz. Mit dieser Option können Sie die Anzahl der Vorhersagen für die einzelnen Datensätze im Dataset einschränken. Der Standardwert ist 3.

Sie könnten beispielsweise sechs Angebote haben (z. B. Sparbuch, Hypothek, Autokredit, Rentensparplan, Kreditkarte und Versicherung), möchten jedoch nur die beiden empfehlenswertesten Angebote ermitteln. In diesem Fall würden Sie das Feld auf 2 setzen. Wenn Sie das Modell erstellen und mit einer Tabelle verknüpfen, sehen Sie zwei Vorhersagespalten (und die zugehörige Konfidenz für die Wahrscheinlichkeit, dass das Angebot angenommen wird) pro Datensatz. Die Vorhersagen können jedes der sechs möglichen Angebote enthalten.

Randomisierungsgrad. Um Verzerrungen zu vermeiden, beispielsweise in einem kleinen oder unvollständigen Dataset, und alle potenziellen Angebote gleich zu behandeln, können Sie einen Randomisierungsgrad zur Angebotsauswahl und die Wahrscheinlichkeit, dass sie als empfohlene Angebote aufgenommen werden, hinzufügen. Die Randomisierung wird als Prozentsatz ausgedrückt, der als Dezimalwert zwischen 0,0 (keine Randomisierung) und 1,0 (vollständig zufällig) angezeigt wird. Der Standardwert lautet 0,0.

Startwert für Zufallsgenerator festlegen. Wenn Sie einen Randomisierungsgrad für die Auswahl eines Angebots angeben, können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Gewünschten Startwert eingeben. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.

Hinweis: Bei Verwendung der Option **Startwert für Zufallsgenerator festlegen** mit Datensätzen, die aus einer Datenbank eingelesen wurden, ist möglicherweise vor der Stichprobenziehung ein Sortierknoten erforderlich, um zu gewährleisten, dass bei jeder Ausführung des Knotens dasselbe Ergebnis erzielt wird. Dies liegt daran, dass der Startwert für den Zufallsgenerator von der Reihenfolge der Datensätze abhängt, die in relationalen Datenbanken nicht unbedingt gleich bleibt.

Sortierreihenfolge. Wählen Sie die Reihenfolge aus, in der die Angebote im erstellten Modell angezeigt werden sollen:

- **Absteigend.** Das Modell zeigt die Angebote mit den höchsten Scores zuerst an. Hierbei handelt es sich um die Angebote, die mit der größten Wahrscheinlichkeit angenommen werden.
- **Aufsteigend.** Das Modell zeigt die Angebote mit den niedrigsten Scores zuerst an. Hierbei handelt es sich um die Angebote, die mit der größten Wahrscheinlichkeit abgelehnt werden. Dies kann beispielsweise nützlich sein, wenn Sie ermitteln möchten, welche Kunden aus einer Marketingkampagne für ein bestimmtes Angebot gestrichen werden sollen.

Vorgaben für Zielfelder. Beim Erstellen eines Modells kann es bestimmte Datenaspekte geben, die Sie aktiv begünstigen bzw. entfernen möchten. Wenn Sie beispielsweise ein Modell erstellen, das das beste Finanzangebot auswählt, für das beim Kunden geworben werden soll, möchten Sie möglicherweise sicherstellen, dass ein bestimmtes Angebot immer aufgenommen wird, unabhängig davon, wie gut sein Score bei den einzelnen Kunden ist.

Um ein Angebot in diesem Fenster einzuschließen und seinen Präferenzgrad zu bearbeiten, klicken Sie auf **Hinzufügen**, geben Sie den Namen des Angebots ein (z. B. "Sparbuch" oder "Hypothek" und klicken Sie auf **OK**.

- **Wert.** Hier wird der Name des hinzugefügten Angebots angezeigt.

- **Vorgabe.** Gibt den Präferenzgrad an, der auf das Angebot angewendet werden soll. Die Präferenz wird als Prozentsatz ausgedrückt, der als Dezimalwert zwischen 0,0 (keine Präferenz) und 1,0 (höchste Präferenz) angezeigt wird. Der Standardwert lautet 0,0.
- **Immer einschließen.** Aktivieren Sie dieses Kontrollkästchen, um sicherzustellen, dass ein bestimmtes Angebot immer in die Vorhersagen eingeschlossen wird.

  Hinweis: Wenn die **Vorgabe** auf 0,0 gesetzt ist, wird die Einstellung **Immer einschließen** ignoriert.

Reliabilität des Modells berücksichtigen. Ein gut strukturiertes Modell, das reich an Daten ist und durch mehrere erneute Generierungen eine Feinabstimmung erfahren hat, sollte stets genauere Ergebnisse liefern als ein ganz neues Modell mit wenigen Daten. Wenn Sie die höhere Reliabilität des ausgereifteren Modells nutzen möchten, aktivieren Sie dieses Kontrollkästchen.

SLRM-Modellnuggets

Hinweis: Ergebnisse werden in dieser Registerkarte nur angezeigt, wenn Sie auf der Registerkarte "Modelloptionen" sowohl **Modellauswertung einschließen** als auch **Modellevaluation anzeigen** auswählen.

Wenn Sie einen Stream ausführen, der ein SLRM-Modell enthält, berechnet der Knoten Schätzungen für die Genauigkeit der Vorhersagen für jeden Wert des Zielfelds (Angebot) und die Wichtigkeit der einzelnen verwendeten Prädiktoren.

Hinweis: Wenn Sie auf der Registerkarte "Modell" des Modellierungsknotens die Option **Training des bestehenden Modells fortsetzen**ausgewählt haben, werden die im Modellnugget angezeigten Informationen bei jeder erneuten Erstellung des Modells aktualisiert.

Bei mit IBM SPSS Modeler 12.0 oder höher erstellten Modellen ist die Registerkarte "Modell" des Modellnuggets in zwei Spalten unterteilt:

Linke Spalte.

- **Ansicht.** Wenn Sie mehrere Angebote haben, wählen Sie das Angebot aus, für das Ergebnisse angezeigt werden sollen.
- **Leistungsfähigkeit des Modells.** Zeigt die geschätzte Modellgenauigkeit für jedes Angebot an. Das Testset wird durch Simulation generiert.

Rechte Spalte.

- **Ansicht.** Wählen Sie aus, ob Details zu **Assoziation mit Antwort** oder zu **Bedeutsamkeit der Variablen** angezeigt werden sollen.
- **Assoziation mit Antwort.** Zeigt die Assoziation (Korrelation) der einzelnen Prädiktoren mit der Zielvariablen an.
- **Prädiktoreinfluss.** Gibt die relative Wichtigkeit der einzelnen Prädiktoren für die Schätzung des Modells an. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und diejenigen Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Dieses Diagramm kann auf dieselbe Weise interpretiert werden wie für andere Modelle, die den Prädiktoreinfluss anzeigen. Allerdings wird bei SLRM das Diagramm vom SLRM-Algorithmus durch Simulation erzeugt. Dies geschieht, indem jeder Prädiktor nacheinander aus dem Modell entfernt und angezeigt wird, wie dies die Modellgenauigkeit beeinflusst. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors” auf Seite 43.

SLRM-Modell - Einstellungen

Auf der Registerkarte "Einstellungen" für ein SLRM-Modellnugget werden Optionen zum Ändern des erstellten Modells angegeben. Beispielsweise können Sie mit dem SLRM-Knoten unter Verwendung derselben Daten und Einstellungen mehrere verschiedene Modelle erstellen und anschließend diese Registerkarte bei jedem Modell verwenden, um die Einstellungen leicht abzuändern und zu ermitteln, welche Auswirkungen dies auf die Ergebnisse hat.

Hinweis: Diese Registerkarte ist erst verfügbar, nachdem das Modellnugget einem Stream hinzugefügt wurde.

Maximale Anzahl an Vorhersagen pro Datensatz. Mit dieser Option können Sie die Anzahl der Vorhersagen für die einzelnen Datensätze im Dataset einschränken. Der Standardwert ist 3.

Sie könnten beispielsweise sechs Angebote haben (z. B. Sparbuch, Hypothek, Autokredit, Rentensparplan, Kreditkarte und Versicherung), möchten jedoch nur die beiden empfehlenswertesten Angebote ermitteln. In diesem Fall würden Sie das Feld auf 2 setzen. Wenn Sie das Modell erstellen und mit einer Tabelle verknüpfen, sehen Sie zwei Vorhersagespalten (und die zugehörige Konfidenz für die Wahrscheinlichkeit, dass das Angebot angenommen wird) pro Datensatz. Die Vorhersagen können jedes der sechs möglichen Angebote enthalten.

Randomisierungsgrad. Um Verzerrungen zu vermeiden, beispielsweise in einem kleinen oder unvollständigen Dataset, und alle potenziellen Angebote gleich zu behandeln, können Sie einen Randomisierungsgrad zur Angebotsauswahl und die Wahrscheinlichkeit, dass sie als empfohlene Angebote aufgenommen werden, hinzufügen. Die Randomisierung wird als Prozentsatz ausgedrückt, der als Dezimalwert zwischen 0,0 (keine Randomisierung) und 1,0 (vollständig zufällig) angezeigt wird. Der Standardwert lautet 0,0.

Startwert für Zufallsgenerator festlegen. Wenn Sie einen Randomisierungsgrad für die Auswahl eines Angebots angeben, können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Gewünschten Startwert eingeben. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.

Hinweis: Bei Verwendung der Option **Startwert für Zufallsgenerator festlegen** mit Datensätzen, die aus einer Datenbank eingelesen wurden, ist möglicherweise vor der Stichprobenziehung ein Sortierknoten erforderlich, um zu gewährleisten, dass bei jeder Ausführung des Knotens dasselbe Ergebnis erzielt wird. Dies liegt daran, dass der Startwert für den Zufallsgenerator von der Reihenfolge der Datensätze abhängt, die in relationalen Datenbanken nicht unbedingt gleich bleibt.

Sortierreihenfolge. Wählen Sie die Reihenfolge aus, in der die Angebote im erstellten Modell angezeigt werden sollen:

- **Absteigend.** Das Modell zeigt die Angebote mit den höchsten Scores zuerst an. Hierbei handelt es sich um die Angebote, die mit der größten Wahrscheinlichkeit angenommen werden.
- **Aufsteigend.** Das Modell zeigt die Angebote mit den niedrigsten Scores zuerst an. Hierbei handelt es sich um die Angebote, die mit der größten Wahrscheinlichkeit abgelehnt werden. Dies kann beispielsweise nützlich sein, wenn Sie ermitteln möchten, welche Kunden aus einer Marketingkampagne für ein bestimmtes Angebot gestrichen werden sollen.

Vorgaben für Zielfelder. Beim Erstellen eines Modells kann es bestimmte Datenaspekte geben, die Sie aktiv begünstigen bzw. entfernen möchten. Wenn Sie beispielsweise ein Modell erstellen, das das beste Finanzangebot auswählt, für das beim Kunden geworben werden soll, möchten Sie möglicherweise sicherstellen, dass ein bestimmtes Angebot immer aufgenommen wird, unabhängig davon, wie gut sein Score bei den einzelnen Kunden ist.

Um ein Angebot in diesem Fenster einzuschließen und seinen Präferenzgrad zu bearbeiten, klicken Sie auf **Hinzufügen**, geben Sie den Namen des Angebots ein (z. B. "Sparbuch" oder "Hypothek" und klicken Sie auf **OK**.

- **Wert.** Hier wird der Name des hinzugefügten Angebots angezeigt.
- **Vorgabe.** Gibt den Präferenzgrad an, der auf das Angebot angewendet werden soll. Die Präferenz wird als Prozentsatz ausgedrückt, der als Dezimalwert zwischen 0,0 (keine Präferenz) und 1,0 (höchste Präferenz) angezeigt wird. Der Standardwert lautet 0,0.

- **Immer einschließen.** Aktivieren Sie dieses Kontrollkästchen, um sicherzustellen, dass ein bestimmtes Angebot immer in die Vorhersagen eingeschlossen wird.

  Hinweis: Wenn die **Vorgabe** auf 0,0 gesetzt ist, wird die Einstellung **Immer einschließen** ignoriert.

Reliabilität des Modells berücksichtigen. Ein gut strukturiertes Modell, das reich an Daten ist und durch mehrere erneute Generierungen eine Feinabstimmung erfahren hat, sollte stets genauere Ergebnisse liefern als ein ganz neues Modell mit wenigen Daten. Wenn Sie die höhere Reliabilität des ausgereifteren Modells nutzen möchten, aktivieren Sie dieses Kontrollkästchen.

Kapitel 15. SVM-Modelle

Informationen zu SVM

SVM (Support Vector Machine) ist ein robustes Klassifikations- und Regressionsverfahren, das die Vorhersagegenauigkeit von Modellen maximiert, ohne dass es zu einer Überanpassung an die Trainingsdaten kommt. SVM ist insbesondere für die Analyse von Daten mit einer sehr großen Anzahl (z. B. mehrere Tausend) an Prädiktorfeldern geeignet.

Für SVM gibt es Anwendungsbereiche in zahlreichen Fachgebieten, wie Customer Relationship Management (CRM), Gesichts- und sonstige Bilderkennung, Bioinformatik, Konzeptextraktion beim Textmining, Intrusion Detection, Proteinstrukturvorhersage und Stimm- und Spracherkennung.

Funktionsweise von SVM

SVM (Support Vector Machine) funktioniert durch Zuordnung von Daten zu einem hochdimensionalen Merkmalbereich, sodass Datenpunkte kategorisiert erden können, selbst wenn die Daten anderweitig nicht linear getrennt werden können. Eine Trennlinie zwischen den Kategorien wird ermittelt. Anschließend werden die Daten derart transformiert, dass die Trennlinie als Hyperebene gezeichnet werden könnte. Danach kann anhand der Eigenschaften neuer Daten die Gruppe vorhergesagt werden, zu der ein neuer Datensatz gehören sollte.

Betrachten Sie beispielsweise die folgende Abbildung, in der die Datenpunkte in zwei verschiedene Kategorien fallen.

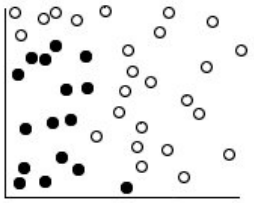


Abbildung 60. Ursprüngliches Dataset

Die zwei Kategorien können durch eine Kurve getrennt werden, wie in der folgenden Abbildung dargestellt ist.

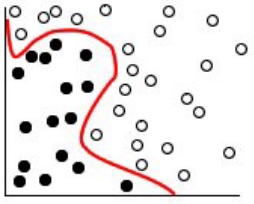


Abbildung 61. Daten nach Hinzufügen der Trennlinie

Nach der Transformation kann die Grenze zwischen den zwei Kategorien durch eine Hyperebene, wie in der folgenden Abbildung dargestellt, definiert werden.

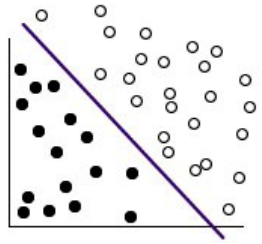


Abbildung 62. Transformierte Daten

Die für die Transformation verwendete mathematische Funktion wird als **Kern**-Funktion bezeichnet. SVM in IBM SPSS Modeler unterstützt folgende Kerntypen:

- Linear
- Polynomial
- Radiale Basisfunktion (RBF)
- Sigmoid

Eine lineare Kernfunktion wird empfohlen, wenn die lineare Trennung der Daten unproblematisch ist. In anderen Fällen sollte eine der folgenden Funktionen verwendet werden. Sie sollten die verschiedenen Funktionen ausprobieren, um in jedem Fall das beste Modell zu erzielen, da jede davon andere Algorithmen und Parameter verwendet.

Feinabstimmung von SVM-Modellen

Neben der Trennlinie zwischen den Kategorien, ermittelt ein SVM-Modell zur Klassifizierung auch Randlinien, die den Raum zwischen zwei Kategorien definieren.

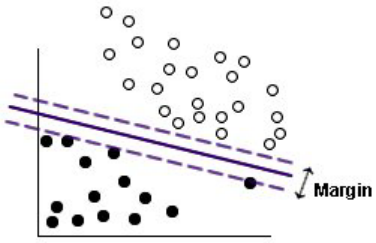


Abbildung 63. Daten mit einem vorläufigen Modell

Die in den Rändern liegenden Datenpunkte werden als **Supportvektoren** bezeichnet.

Je breiter der Rand zwischen den beiden Kategorien ist, desto besser ist das Modell bei der Vorhersage der Kategorie für neue Datensätze. Im vorherigen Beispiel ist der Rand nicht besonders breit und das Modell gilt als **überangepasst**. Ein geringes Maß an Fehlklassifizierung kann akzeptiert werden, um den Rand zu verbreitern. In der folgenden Abbildung finden Sie ein Beispiel hierfür.

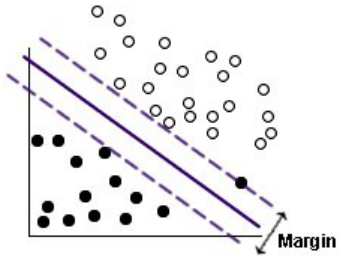


Abbildung 64. Daten mit einem verbesserten Modell

In manchen Fällen ist eine lineare Trennung schwieriger. In der folgenden Abbildung finden Sie ein Beispiel hierfür.

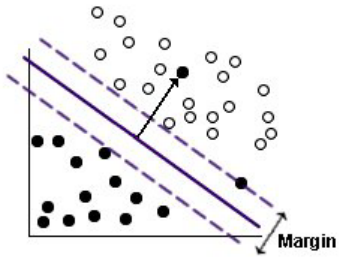


Abbildung 65. Problem bei linearer Trennung

In solchen Fällen besteht das Ziel darin, die optimale Balance zwischen einem möglichst breiten Rand und einer möglichst kleinen Zahl fehlklassifizierter Datenpunkte zu finden. Die Kernfunktion weist einen **Regularisierungsparameter** (als "C" bekannt) auf, der den Ausgleich zwischen diesen beiden Werten steuert. Es wird vermutlich erforderlich sein, verschiedene Werte für diesen und andere Kernparameter auszuprobieren, um das beste Modell zu finden.

SVM-Knoten

Mit dem SCM-Knoten können Sie eine Support Vector Machine zum Klassifizieren von Daten verwenden. SCM eignet sich insbesondere für umfangreiche Datasets, also solche mit einer großen Anzahl an Prädiktorfeldern. Mit den Standardeinstellungen im Knoten können Sie in relativ kurzer Zeit ein Grundmodell erstellen. Alternativ können Sie mithilfe der Experteneinstellungen verschiedene Typen von SVM-Modellen ausprobieren.

Nach der Erstellung des Modells haben Sie folgende Möglichkeiten:

- Durchsuchen des Modellnuggets zur Anzeige der relativen Bedeutsamkeit der Eingabefelder bei der Erstellung des Modells.
- Anfügen eines Tabellenknotens zum Modellnugget zur Anzeige der Modellausgabe.

Beispiel. Ein medizinischer Forscher hat ein Dataset mit den Eigenschaften einer Reihe von Stichproben menschlicher Zellen erstellt, die von Patienten stammen, bei denen ein Krebsrisiko angenommen wurde. Die Analyse der ursprünglichen Daten ergab, dass bei vielen der Eigenschaften deutliche Unterschiede zwischen den gutartigen und den bösartigen Proben bestehen. Der Forscher möchte ein SVM-Modell entwickeln, das die Werte ähnlicher Zellenmerkmale in Proben von anderen Patienten verwenden kann, um eine Frühindikation dafür abzugeben, ob die Proben gutartig oder bösartig sind.

Modelloptionen für SVM-Knoten

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen” auf Seite 28.

Expertenoptionen für SVM-Knoten

Wenn Sie über umfassende Kenntnisse im Bereich Support Vector Machines verfügen, können Sie mithilfe der Expertenoptionen eine Feinabstimmung des Trainingsvorgangs vornehmen. Für den Zugriff auf die Expertenoptionen setzen Sie den Modus auf der Registerkarte "Experten" auf **Experten**.

Alle Wahrscheinlichkeiten anhängen (nur gültig für kategoriale Ziele). Wenn diese Option ausgewählt ist, gibt sie an, dass Wahrscheinlichkeiten für jeden möglichen Wert eines Zielfelds vom Typ "nominal" oder "Flag" für jeden vom Knoten verarbeiteten Datensatz angezeigt werden. Wenn diese Option nicht ausgewählt ist, wird für Zielfelder vom Typ "Flag" oder "Nominal" nur die Wahrscheinlichkeit des vorhergesagten Werts angezeigt. Die für dieses Kontrollkästchen festgelegte Einstellung bestimmt die Standardeinstellung des entsprechenden Kontrollkästchens in der Anzeige für das Modellnugget.

Stoppkriterien. Bestimmt, wann der Optimierungsalgorithmus gestoppt werden soll. Die Werte liegen im Bereich von 1.0E-1 bis 1.0E-6. Der Standardwert ist 1.0E-3. Wenn Sie den Wert verringern, erhalten Sie ein genaueres Modell. Jedoch benötigen Sie mehr Zeit zum Trainieren des Modells.

Regularisierungsparameter (C). Steuert den Ausgleich zwischen der Maximierung des Rands und der Minimierung des Fehlerterms für das Training. Die Werte sollten normalerweise zwischen 1 und 10 (einschließlich) liegen. Der Standardwert ist 10. Bei einer Erhöhung des Werts wird die Klassifizierungsgenauigkeit für die Trainingsdaten verbessert (bzw. der Regressionsfehler verringert), dies kann jedoch auch zu einer Überanpassung führen.

Regressionsgenauigkeit (Epsilon). Nur verwendet, wenn es sich beim Messniveau des Zielfelds um *Continuous* (Stetig) handelt. Führt dazu, dass Fehler hingenommen werden, vorausgesetzt, dass sie unter dem hier angegebenen Wert liegen. Eine Erhöhung des Werts kann zu schnellerer Modellierung führen, jedoch geht dies auf Kosten der Genauigkeit.

Kerntyp. Bestimmt den Typ der für die Transformation verwendeten Kernfunktion. Unterschiedliche Kerntypen führen dazu, dass die Trennlinie auf unterschiedliche Weise berechnet wird, weshalb es ratsam ist, mit den verschiedenen Optionen zu experimentieren. Der Standardwert lautet **RBF** (Radiale Basisfunktion).

RBF-Gamma. Nur aktiviert, wenn der Kerntyp auf **RBF** gesetzt ist. Der Wert sollte normalerweise zwischen $3/k$ und $6/k$ liegen, wobei k die Anzahl der Eingabefelder ist. Bei 12 Eingabefeldern beispielsweise wären somit Werte im Bereich von 0,25 bis 0,5 einen Versuch wert. Bei einer Erhöhung des Werts wird die Klassifizierungsgenauigkeit für die Trainingsdaten verbessert (bzw. der Regressionsfehler verringert), dies kann jedoch auch zu einer Überanpassung führen.

Gamma. Nur aktiviert, wenn der Kerntyp auf **Polynomial** oder **Sigmoid**. gesetzt ist. Bei einer Erhöhung des Werts wird die Klassifizierungsgenauigkeit für die Trainingsdaten verbessert (bzw. der Regressionsfehler verringert), dies kann jedoch auch zu einer Überanpassung führen.

Verzerrung. Nur aktiviert, wenn der Kerntyp auf **Polynomial** oder **Sigmoid**. gesetzt ist. Legt den Wert `coef0` in der Kernfunktion fest. Der Standardwert 0 ist in den meisten Fällen geeignet.

Grad. Nur aktiviert, wenn der Kerntyp auf **Polynomial** gesetzt ist. Steuert die Komplexität (Dimension) des Zuordnungsraums. Normalerweise werden nur Werte bis maximal 10 verwendet.

SVM-Modellnugget

Das SVM-Modell erstellt eine Reihe neuer Felder. Das wichtigste dieser Felder ist das Feld **$S-fieldname**, das den vom Modell vorhergesagten Wert für das Zielfeld anzeigt.

Anzahl und Namen der vom Modell erstellten neuen Felder hängen vom Messniveau des Zielfelds ab (in den folgenden Tabellen durch *Feldname* angegeben).

Um diese Felder und die zugehörigen Werte anzuzeigen, müssen Sie einen Tabellenknoten zum SVM-Modellnugget hinzufügen und den Tabellenknoten ausführen.

Tabelle 24. Messniveau des Zielfelds ist 'Nominal' oder 'Flag'

Neuer Feldname	Beschreibung
$S-*Feldname*	Vorhergesagter Wert des Zielfelds.
$SP-*Feldname*	Wahrscheinlichkeit des vorhergesagten Werts.
$SP-*Wert*	Wahrscheinlichkeit jedes möglichen Werts von "Nominal" oder "Flag" (nur angezeigt, wenn auf der Registerkarte "Einstellungen" des Modellnuggets die Option **Alle Wahrscheinlichkeiten anhängen** aktiviert ist).
$SRP-*Wert*	(Nur bei Flagzielen) Raw-Propensity-Scores (SRP) und Adjusted-Propensity-Scores (SAP), die die Likelihood eines Ergebnisses vom Typ "wahr" für das Zielfeld angeben. Diese Scores werden nur angezeigt, wenn vor der Generierung des Modells die entsprechenden Kontrollkästchen auf der Registerkarte "Analysieren" des SVM-Modellierungsknotens ausgewählt wurden. Weitere Informationen finden Sie im Thema „Analyseoptionen bei Modellierungsknoten" auf Seite 35.
$SAP-*Wert*	

Tabelle 25. Messniveau des Zielfelds ist 'Continuous' (Stetig)

Neuer Feldname	Beschreibung
$S-*Feldname*	Vorhergesagter Wert des Zielfelds.

Bedeutung des Prädiktors

Optional kann auf der Registerkarte "Modell" auch ein Diagramm, das den relative Einfluss der einzelnen Prädiktoren für die Schätzung des Modells angibt, angezeigt werden. Normalerweise ist es sinnvoll, die Modellierungsbemühungen auf die wichtigsten Prädiktoren zu konzentrieren und die Prädiktoren zu verwerfen bzw. zu ignorieren, die die geringste Bedeutung haben. Beachten Sie, dass dieses Diagramm nur verfügbar ist, wenn vor dem Generieren des Modells **Prädiktoreinfluss berechnen** auf der Registerkarte "Analysieren" ausgewählt wurde. Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Hinweis: Die Berechnung des Prädiktoreinflusses dauert bei einem SVM-Modell möglicherweise länger als bei anderen Modelltypen und ist nicht standardmäßig auf der Registerkarte "Analysieren" ausgewählt. Die Auswahl dieser Option kann die Leistung verlangsamen, insbesondere bei großen Datasets.

Einstellungen beim SVM-Modell

Auf der Registerkarte "Einstellungen" können Sie zusätzliche Felder angeben, die bei der Anzeige der Ergebnisse verwendet werden sollen (z. B. durch Ausführen eines Tabellenknotens, der mit dem Nugget verknüpft ist). Sie können den Effekt jeder dieser Optionen sehen, indem Sie sie auswählen und auf die Schaltfläche "Vorschau" klicken. Führen Sie in der Vorschau-Ausgabe einen Bildlauf nach rechts durch, um die zusätzlichen Felder zu sehen.

Alle Wahrscheinlichkeiten anhängen (nur gültig für kategoriale Ziele). Wenn diese Option ausgewählt ist, werden für jeden Datensatz, der vom Knoten verarbeitet wird, die Wahrscheinlichkeiten für jeden möglichen Wert eines Zielfelds vom Typ "Nominal" oder "Flag" angezeigt. Wenn diese Option nicht ausgewählt ist, werden für Zielfelder vom Typ "Flag" oder "Nominal" nur der vorhergesagte Wert und seine Wahrscheinlichkeit angezeigt.

Die Standardeinstellung dieses Kontrollkästchens richtet sich nach dem entsprechenden Kontrollkästchen des Modellierungsknotens.

Raw-Propensity-Scores berechnen. Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angegebenen wahren Ergebnisses angeben. Diese Werte werden zusätzlich zu den anderen Vorhersage- und Konfidenzwerten angegeben, die gegebenenfalls während des Scorings erstellt werden.

Adjusted-Propensity-Scores berechnen. Raw-Propensity-Scores beruhen ausschließlich auf den Trainingsdaten und sind aufgrund der Neigung vieler Modelle zur übermäßigen Anpassung an die Trainingsdaten möglicherweise zu optimistisch. Bei Adjusted Propensitys wird versucht, dies durch Evaluation der Modellleistung anhand einer Test- bzw. Validierungspartition zu kompensieren. Bei dieser Option muss im Stream ein Partitionsfeld definiert sein und die Adjusted-Propensity-Scores müssen im Modellierungsknoten aktiviert werden, bevor das Modell generiert wird.

Kapitel 16. Nächste-Nachbarn-Modelle

KNN-Knoten

Die Nächste-Nachbarn-Analyse ist eine Methode zum Klassifizieren von Fällen auf der Grundlage ihrer Ähnlichkeit mit anderen Fällen. Beim maschinellen Lernen wurde sie entwickelt, um Datenmuster zu erkennen, ohne dass eine exakte Übereinstimmung mit gespeicherten Mustern oder Fällen benötigt wird. Ähnliche Fälle liegen nah beieinander und Fälle mit geringer Ähnlichkeit sind weit voneinander entfernt. Somit gilt die Distanz zwischen zwei Fällen als Maß für ihre Unähnlichkeit.

Befinden sich Fälle nahe beieinander, werden sie als "Nachbarn" bezeichnet. Wenn ein neuer Fall (Holdout) angegeben wird, wird seine Distanz zu jedem der Fälle im Modell berechnet. Die Klassifizierungen der ähnlichsten Fälle - die nächsten Nachbarn - werden gezählt und der neue Fall wird einer Kategorie zugeordnet, die die größte Anzahl der nächsten Nachbarn enthält.

Sie können die Anzahl der zu untersuchenden nächsten Nachbarn angeben; dieser Wert wird *k* genannt. Die Abbildungen zeigen, wie ein neuer Fall mit zwei verschiedenen Werten für *k* klassifiziert würde. Wenn *k* = 5 ist, wird der neue Fall in Kategorie *1* gesetzt, da die Mehrheit der nächsten Nachbarn zur Kategorie *1* gehört. Wenn jedoch *k* = 9 ist, wird der neue Fall in Kategorie *0* gesetzt, da die Mehrheit der nächsten Nachbarn zur Kategorie *0* gehört.

Die Nächste-Nachbarn-Analyse kann auch zur Berechnung von Werten für ein stetiges Ziel verwendet werden. Dabei wird der durchschnittliche oder Median-Zielwert der nächsten Nachbarn verwendet, um den vorhergesagten Wert für den neuen Fall zu beziehen.

Zieloptionen für KNN-Knoten

Auf der Registerkarte "Ziele" können Sie wählen, ob Sie ein Modell erstellen möchten, das den Wert eines Zielfelds in Ihren Eingabedaten auf der Basis der Werte seiner nächsten Nachbarn vorhersagt, oder ob Sie einfach die nächsten Nachbarn für einen bestimmten Fall von Interesse herausfinden möchten.

Welche Art der Analyse möchten Sie ausführen?

Ein Zielfeld vorhersagen. Wählen Sie diese Option, wenn Sie den Wert eines Zielfelds auf der Grundlage der Werte seiner nächsten Nachbarn vorhersagen möchten.

Nur nächstgelegene Nachbarn ermitteln. Wählen Sie diese Option, wenn Sie nur die nächsten Nachbarn für ein bestimmtes Eingabefeld sehen möchten.

Wenn Sie sich entscheiden, nur einen der nächsten Nachbarn zu identifizieren, werden die übrigen Optionen in dieser Registerkarte (für Genauigkeit und Geschwindigkeit) inaktiviert, da sie nur für die Vorhersage von Zielen relevant sind.

Wie lautet Ihr Ziel?

Mit dieser Optionsgruppe können Sie beim Vorhersagen eines Zielfelds entscheiden, ob Geschwindigkeit, Genauigkeit und/oder beide die wichtigsten Faktoren beim Vorhersagen eines Zielfelds sind. Alternativ können Sie die Einstellungen selbst anpassen.

Wenn Sie die Option "Balancieren", "Geschwindigkeit" oder "Genauigkeit" wählen, trifft der Algorithmus die am besten geeignete Einstellungskombination für diese Option als Voreinstellung. Erfahrene Benutzer möchten diese Voreinstellungen eventuell überschreiben; dies ist in den verschiedenen Bereichen der Registerkarte "Einstellungen" möglich.

Geschwindigkeit und Genauigkeit ausbalancieren. Wählt die beste Anzahl an Nachbarn in einem kleinen Bereich aus.

Geschwindigkeit. Findet eine feste Anzahl an Nachbarn.

Genauigkeit. Wählt die beste Anzahl an Nachbarn in einem größeren Bereich aus und berechnet Entfernungen anhand der Prädiktorenbedeutung.

Benutzerdefinierte Analyse. Wählen Sie diese Option, um den Algorithmus auf der Registerkarte "Einstellungen" genauer einzustellen.

Hinweis: Anders als bei den meisten anderen Modellen nimmt die Größe des daraus hervorgehenden KNN-Modells linear mit der Menge der Trainingsdaten zu. Wird beim Versuch, ein KNN-Modell zu erstellen, eine Fehlernachricht angezeigt, dass nicht genügend Speicher vorhanden ist, versuchen Sie, den maximalen von IBM SPSS Modeler verwendeten Systemspeicher zu erhöhen. Wählen Sie dazu

Tools > **Optionen** > **Systemoptionen**

und geben Sie die neue Größe im Feld **Maximale Speichergröße** ein. Im Dialogfeld "Systemoptionen" vorgenommene Änderungen werden erst nach einem Neustart von IBM SPSS Modeler wirksam.

KNN-Knoten - Einstellungen

Auf der Registerkarte "Einstellungen" geben Sie die spezifischen Optionen für die Nächste-Nachbarn-Analyse an. Die linke Randleiste am Bildschirm listet die Bereiche auf, die Sie zur Festlegung der Optionen verwenden.

Modell

Das Modellfenster bietet Optionen, die steuern, wie das Modell erstellt werden soll, z. B. ob Partitionierung oder Aufteilungsmodelle verwendet werden, ob numerische Eingabefelder so transformiert werden, dass sie alle in demselben Bereich liegen, und wie bestimmte Fälle verwaltet werden. Sie können auch einen benutzerdefinierten Namen für das Modell angeben.

Anmerkung: Partitionierte Daten verwenden und **Fallbeschriftungen verwenden** können nicht dasselbe Feld verwenden.

Modellname. Sie können den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen (oder dem Modelltyp in Fällen, in denen kein solches Feld angegeben ist) generieren oder einen benutzerdefinierten Namen eingeben.

Partitionierte Daten verwenden. Wenn ein Partitionsfeld definiert ist, gewährleistet diese Option, dass nur Daten aus der Trainingspartition für die Modellerstellung verwendet werden.

Aufteilungsmodelle erstellen. Erstellt ein separates Modell für jeden möglichen Wert von Eingabefeldern, die als Aufteilungsfelder angegeben werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen” auf Seite 28.

Felder manuell auswählen... Der Knoten verwendet standardmäßig die Partitions- und Aufteilungsfeldeinstellungen (sofern vorhanden) des Typknotens. Sie können diese Einstellungen jedoch hier überschreiben. Um die **Partitions**- und **Aufteilungs**felder zu aktivieren, wählen Sie die Registerkarte **Felder** und dann die Option **Benutzerdefinierte Einstellungen verwenden**. Kehren Sie anschließend hierher zurück.

- **Partition.** In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer anderen Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln. Wenn mehrere Partitionsfelder mithilfe von Typ- oder Partitionsknoten erstellt wur-

den, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.) Beachten Sie außerdem, dass die Partitionierung auch auf der Registerkarte "Modelloptionen" des Knotens aktiviert sein muss, wenn die ausgewählte Partitionierung in Ihrer Analyse angewendet werden soll. (Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung ohne Änderung der Feldeinstellungen inaktiviert werden.)

- **Aufteilungen.** Wählen Sie für Aufteilungsmodelle das Aufteilungsfeld bzw. die Aufteilungsfelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Aufteilung* festlegen. Sie können nur Felder vom Typ **Flag**, **Nominal** oder **Ordinal** als Aufteilungsfelder festlegen. Als Aufteilungsfelder gewählte Felder können nicht als Ziel-, Prädiktor-, Partitions-, Häufigkeits oder Gewichtungsfelder verwendet werden. Weitere Informationen finden Sie im Thema „Erstellen von aufgeteilten Modellen” auf Seite 28.

Bereichseingaben normalisieren. Markieren Sie dieses Kontrollkästchen, um die Werte für stetige Eingabefelder zu normalisieren. Normalisierte Funktionen umfassen denselben Wertebereich, was die Leistung des Schätzalgorithmus verbessern kann. Die angepasste Normalisierung, [2*(x–min)/(max–min)]–1, wird verwendet. Angepasste normalisierte Werte liegen zwischen –1 und 1.

Fallbeschriftungen verwenden. Wählen Sie dieses Kontrollkästchen aus, um die Dropdown-Liste zu aktivieren, über die Sie ein Feld auswählen können, dessen Werte als Beschriftungen verwendet werden, um die Fälle, die von Interesse sind, im Prädiktorbereichsdiagramm, Peerdiagramm und in der Quadrantenkarte im Modellviewer zu kennzeichnen. Sie können ein beliebiges Feld mit einem Messniveau von *Nominal*, *Ordinal* oder *Flag* als Beschriftungsfeld wählen. Wenn Sie hier kein Feld wählen, werden Datensätze in den Modellviewer-Diagrammen mit den nächstgelegenen Nachbarn angezeigt, wobei diese anhand der Zeilennummer in den Quellendaten identifiziert werden. Wenn Sie die Daten überhaupt nach der Erstellung des Modells ändern, verwenden Sie Beschriftungen, damit Sie nicht jedesmal auf die Quellendaten zurückverweisen müssen, um die Fälle in der Anzeige zu identifizieren.

Fokusdatensatz identifizieren. Markieren Sie dieses Kontrollkästchen, um die Dropdown-Liste zu aktivieren, in der Sie ein Eingabefeld von besonderem Interesse kennzeichnen können (nur für Flagfelder). Wenn Sie hier ein Feld festlegen, sind die Punkte, die dieses Feld repräsentieren, anfangs im Modellviewer ausgewählt, während das Modell erstellt wird. Die Auswahl eines Fokusdatensatzes an dieser Stelle ist optional. Jeder Punkt kann temporär zu einem Fokusdatensatz werden, wenn er manuell im Modellviewer ausgewählt wird.

Nachbarn

Das Fenster "Nachbarn" enthält ein Set an Optionen, die steuern, wie die Anzahl der nächsten Nachbarn berechnet wird.

Anzahl der nächstgelegenen Nachbarn (k). Geben Sie die Anzahl der nächsten Nachbarn für einen bestimmten Fall ein. Beachten Sie dabei, dass eine höhere Anzahl an Nachbarn nicht unbedingt ein präziseres Modell hervorbringt.

Wenn geplant ist, dass ein Ziel vorhergesagt werden soll, stehen zwei Optionen zur Auswahl:

- **Festen Wert für k angeben.** Verwenden Sie diese Option, wenn Sie eine feste Anzahl von nächsten Nachbarn angeben möchten, die gesucht werden sollen.
- **k automatisch auswählen.** Alternativ können Sie über die Felder **Minimum** und **Maximum** einen Wertebereich angeben und für die Prozedur das Auswählen der "besten" Anzahl von Nachbarn innerhalb dieses Bereichs zulassen. Die Methode zur Bestimmung der Anzahl der nächsten Nachbarn hängt davon ab, ob die Merkmalauswahl im Merkmalauswahlfenster verlangt wird.

  Wenn die Merkmalauswahl aktiviert wurde, wird für jeden Wert von *k* im angegebenen Bereich eine Merkmalauswahl durchgeführt und *k* und die zugehörige Funktionsgruppe mit der niedrigsten Fehlerrate (oder dem geringsten Quadratsummen-Fehler, falls das Ziel stetig ist) werden ausgewählt.

Wenn eine Merkmalauswahl nicht aktiviert ist, wird eine *V*-fache Kreuzvalidierung verwendet, um die "beste" Anzahl an Nachbarn auszuwählen. Siehe den Bereich "Kreuzvalidierung" für die Steuerung der Zuweisung von Aufteilungen.

Distanzberechnung. Mit diesem Wert wird das Längenmaßsystem für die Messung der Ähnlichkeit von Fällen festgelegt.

- **Euklidische Metrik.** Der Abstand zwischen zwei Fällen, x und y, ergibt sich aus der Quadratwurzel der Summe, über alle Dimensionen, der quadrierten Differenzen zwischen den Werten für die Fälle.
- **City-Block-Metrik.** Die Distanz zwischen zwei Fällen ergibt sich aus der Summe, über alle Dimensionen, der absoluten Differenzen zwischen den Werten der Fälle. Dies wird auch als "Manhattan-Distanz" bezeichnet.

Optional: Wenn geplant ist, ein Ziel vorherzusagen, können Sie beim Berechnen von Distanzen die Merkmale nach ihrer normalisierten Bedeutung gewichten. Die Merkmalswichtigkeit für einen Prädiktor wird durch das Verhältnis der Fehlerquote oder des Quadratsummenfehlers des Modells (wobei der Prädiktor vom Modell entfernt wird) zur Fehlerquote bzw. zum Quadratsummenfehler für das vollständige Modell berechnet. Die normalisierte Wichtigkeit wird durch die Neugewichtung der Werte der Funktionswichtigkeit berechnet, sodass deren Summe 1 ergibt.

Merkmale bei Berechnung von Abständen nach Wichtigkeit gewichten. (Wird nur angezeigt, wenn ein Ziel vorhergesagt werden soll.) Markieren Sie dieses Kontrollkästchen, damit der Prädiktoreinfluss bei der Distanzberechnung zwischen Nachbarn verwendet wird. Der Prädiktoreinfluss wird dann im Modellnugget angezeigt und in Vorhersagen verwendet (und beeinflusst damit das Scoring). Weitere Informationen finden Sie im Thema „Bedeutung des Prädiktors" auf Seite 43.

Vorhersagen für Bereichsziel. (Wird nur angezeigt, wenn ein Ziel vorhergesagt werden soll.) Wenn ein stetiges Ziel (numerischer Bereich) angegeben ist, definiert dies, ob der vorhergesagte Wert auf der Basis des Mittel- oder Medianwerts der nächsten Nachbarn berechnet wird.

Merkmalauswahl

Dieser Bereich wird nur aktiviert, wenn ein Ziel vorhergesagt werden soll. Hier können Sie Optionen zur Merkmalauswahl anfordern und angeben. Standardmäßig werden bei der Merkmalauswahl alle Funktionen berücksichtigt, Sie können optional aber auch eine Untergruppe von Funktionen auswählen, die in das Modell aufgenommen werden sollen.

Merkmalauswahl durchführen. Markieren Sie dieses Kontrollkästchen, um die Optionen zur Merkmalauswahl zu aktivieren.

- **Erzwungene Eingabe.** Klicken Sie auf die Feldauswahlschaltfläche neben diesem Feld und wählen Sie ein oder mehrere Merkmale, deren Aufnahme in das Modell erzwungen werden soll.

Stoppkriterium. Bei jedem Schritt wird die Funktion, deren Integration in das Modell den geringsten Fehler hervorruft (für kategoriale Ziele als Fehlerrate und für stetige Ziele als Quadratsummenfehler berechnet), für die Integration in das Modell in Betracht gezogen. Die Vorwärtsselektion wird fortgesetzt, bis die angegebene Bedingung erfüllt wird.

- **Stoppen, wenn die angegebene Anzahl an Merkmalen ausgewählt wurde.** Der Algorithmus fügt neben den erzwungenen Funktionen eine feste Anzahl an Funktionen in das Modell ein. Geben Sie eine positive Ganzzahl ein. Eine geringere Anzahl an Werten führt zu einem sparsameren Modell. Dabei läuft man allerdings Gefahr, wichtige Funktionen zu vernachlässigen. Bei einer höheren Anzahl an Werten werden alle wichtigen Funktionen erfasst, dafür läuft man aber Gefahr, Funktionen einzufügen, die den Modellfehler erhöhen.
- **Stoppen, wenn die Änderung im absoluten Fehlerquotienten kleiner oder gleich dem Minimum ist.** Der Algorithmus wird beendet, wenn die Änderung im absoluten Fehlerquotienten vermuten lässt, dass das Modell durch Hinzufügen weiterer Funktionen nicht mehr weiter optimiert werden kann. Geben Sie eine positive Zahl an. Geringere Werte für die Mindeständerung berücksichtigen mehr Merkmale und bergen das Risiko, dass Merkmale aufgenommen werden, die dem Modell keinen zusätzli-

chen Wert bescheren. Bei einem höheren Wert für die minimale Änderung werden mehr Funktionen ausgeschlossen, was dazu führen kann, dass Funktionen ausgeschlossen werden, die wichtig für das Modell wären. Der "optimale" Wert für die minimale Änderung hängt von Ihren Daten und Ihrer Anwendung ab. Informationen dazu, wie Sie beurteilen, welche Funktionen am wichtigsten sind, finden Sie im Protokoll über die Merkmalauswahlfehler in der Ausgabe. Weitere Informationen finden Sie im Thema „Prädiktor-Auswahlfehler-Protokoll" auf Seite 297.

Kreuzvalidierung

Dieser Bereich wird nur aktiviert, wenn ein Ziel vorhergesagt werden soll. Die Optionen in diesem Bereich steuern, ob beim Berechnen der nächsten Nachbarn Kreuzvalidierung verwendet werden soll.

Bei der Kreuzvalidierung wird die Stichprobe in mehrere Teilstichproben oder **Aufteilungen** gegliedert. Anschließend werden Nächste-Nachbarn-Modelle erzeugt; dabei werden nacheinander die Daten der einzelnen Stichproben ausgeschlossen. Das erste Modell beruht auf allen Fällen mit Ausnahme der Fälle in der ersten Stichprobenaufteilung, das zweite Modell auf allen Fällen mit Ausnahme der Fälle in der zweiten Stichprobenaufteilung usw. Bei jedem Modell wird jeweils der Fehler geschätzt. Hierzu wird das Modell auf die Teilstichprobe angewendet, die beim Erstellen des Modells ausgeschlossen war. Die "beste" Anzahl an nächstgelegenen Nachbarn ist die Anzahl, die die wenigsten Fehler für alle Aufteilungen erzeugt.

Kreuzvalidierungsaufteilungen. Um die "beste" Anzahl an Nachbarn zu ermitteln wird eine *V*-fache Kreuzvalidierung durchgeführt. Bei Merkmalauswahl ist sie aus Leistungsgründen nicht verfügbar.

- **Fälle willkürlich Aufteilungen zuweisen.** Geben Sie die Anzahl an Aufteilungen an, die für die Kreuzvalidierung herangezogen werden sollen. Die Prozedur weist Fälle willkürlich Aufteilungen zu und nummeriert sie von 1 bis *V*, die Anzahl an Aufteilungen.
- **Startwert für Zufallsgenerator festlegen.** Bei der Schätzung der Genauigkeit eines Modells auf der Grundlage eines Zufallsprozentsatzes können Sie mit dieser Option dieselben Ergebnisse in einer anderen Sitzung replizieren. Wenn Sie den vom Zufallszahlengenerator verwendeten Startwert angeben, stellen Sie sicher, dass bei jeder Ausführung des Knotens dieselben Datensätze zugewiesen werden. Gewünschten Startwert eingeben. Wenn diese Option nicht ausgewählt ist, wird bei jeder Ausführung des Knotens eine andere Stichprobe generiert.
- **Feld für die Zuweisung von Fällen verwenden.** Geben Sie ein numerisches Feld an, das jeden Fall im aktiven Dataset einer Unterteilung zuweist. Das Feld muss numerisch sein und Werte von 1 bis *V* annehmen. Wenn Werte in diesem Bereich fehlen und wenn für Aufteilungsfelder etwaige Aufteilungsmodelle wirksam sind, wird ein Fehler ausgelöst.

Analysieren

Der Bereich "Analyse" wird nur aktiviert, wenn ein Ziel vorhergesagt werden soll. Sie können diesen Bereich verwenden, um anzugeben, ob das Modell zusätzliche Variablen enthalten soll, die Folgendes enthalten:

- Wahrscheinlichkeiten für jeden möglichen Wert eines Zielfelds
- Distanzen zwischen einem Fall und seinen nächsten Nachbarn
- Scores für Raw Propensity und Adjusted Propensity (nur für Flagziele)

Alle Wahrscheinlichkeiten anhängen. Wenn diese Option ausgewählt ist, werden für jeden Datensatz, der vom Knoten verarbeitet wird, die Wahrscheinlichkeiten für jeden möglichen Wert eines Zielfelds vom Typ "Nominal" oder "Flag" angezeigt. Wenn diese Option nicht ausgewählt ist, werden für Zielfelder vom Typ "Flag" oder "Nominal" nur der vorhergesagte Wert und seine Wahrscheinlichkeit angezeigt.

Abstände zwischen Fällen und k nächstgelegenen Nachbarn speichern. Für jeden Fokusdatensatz wird eine separate Variable für jeden der *k* nächsten Nachbarn des Fokusdatensatzes (aus der Trainingsstichprobe) und die entsprechenden *k* nächsten Distanzen erstellt.

Propensity-Scores

Propensity-Scores können im Modellierungsknoten oder auf der Registerkarte "Einstellungen" im Modellnugget aktiviert werden. Diese Funktionalität ist nur verfügbar, wenn das ausgewählte Ziel ein Flagfeld ist. Weitere Informationen finden Sie im Thema „Propensity-Scores" auf Seite 36.

Raw-Propensity-Scores berechnen. Raw-Propensity-Scores werden ausschließlich auf der Grundlage der Trainingsdaten aus dem Modell abgeleitet. Wenn das Modell den Wert *wahr* (wird antworten) vorhersagt, ist die Neigung mit P identisch. Dabei ist P die Wahrscheinlichkeit der Vorhersage. Wenn das Modell den Wert "falsch" vorhersagt, wird die Neigung als (1 - P) berechnet.

- Wenn Sie bei der Modellerstellung diese Option auswählen, werden standardmäßig Propensity-Scores im Modellnugget aktiviert. Sie können jedoch immer festlegen, dass Raw-Propensity-Scores im Modellnugget aktiviert werden sollen, unabhängig davon, ob Sie sie im Modellierungsknoten auswählen.
- Beim Scoren des Modells werden Raw-Propensity-Scores in einem Feld hinzugefügt, bei dem die Buchstaben *RP* an das Standardpräfix angehängt sind. Wenn sich die Vorhersagen beispielsweise im Feld *$R-Abwanderung* befinden, lautet der Name des Felds für den Propensity-Score *$RRP-Abwanderung*.

Adjusted-Propensity-Scores berechnen. Raw Propensitys basieren ausschließlich auf vom Modell angegebenen Schätzungen. Beim Modell kann jedoch eine Überanpassung vorliegen, was zu übermäßig optimistischen Schätzungen für die Neigung führt. Adjusted Propensitys versuchen, dies zu kompensieren, indem untersucht wird, wie leistungsfähig das Modell bei den Test- bzw. Validierungspartitionen ist, und die Neigungen entsprechend angepasst werden, um eine bessere Schätzung zu erzielen.

- Diese Einstellung ist nur möglich, wenn ein gültiges Partitionsfeld im Stream vorhanden ist.
- Anders als rohe Konfidenzscores müssen Adjusted-Propensity-Scores bei der Erstellung des Modells berechnet werden; andernfalls sind sie beim Scoren des Modellnuggets nicht verfügbar.
- Beim Scoren des Modells werden Adjusted-Propensity-Scores in einem Feld hinzugefügt, bei dem die Buchstaben *AP* an das Standardpräfix angehängt sind. Wenn sich die Vorhersagen beispielsweise im Feld *$R-Abwanderung* befinden, lautet der Name des Felds für den Propensity-Score *$RAP-Abwanderung*. Adjusted-Propensity-Scores sind bei logistischen Regressionsmodellen nicht verfügbar.
- Bei der Berechnung der Adjusted-Propensity-Scores darf die für die Berechnung verwendete Test- bzw. Validierungspartition nicht ausbalanciert worden sein. Um dies zu vermeiden, müssen Sie darauf achten, dass in etwaigen weiter oben im Stream befindlichen Balancierungsknoten die Option **Balancierung nur für Trainingsdaten durchführen** ausgewählt wurde. Zusätzlich gilt: Wenn eine komplexe Stichprobe gezogen wurde, werden dadurch die Adjusted-Propensity-Scores ungültig.
- Adjusted-Propensity-Scores sind bei verstärkten Baum- und Regelsetmodellen nicht verfügbar. Weitere Informationen finden Sie im Thema „Verbesserte C5.0-Modelle" auf Seite 113.

KNN-Modellnugget

Das KNN-Modell erstellt eine Reihe neuer Felder wie in der folgenden Tabelle gezeigt. Um diese Felder und die zugehörigen Werte anzuzeigen, müssen Sie einen Tabellenknoten zum KNN-Modellnugget hinzufügen und den Tabellenknoten ausführen oder auf die Schaltfläche "Vorschau" am Nugget klicken.

Tabelle 26. KNN-Modellfelder

Neuer Feldname	Beschreibung
$KNN-*Feldname*	Vorhergesagter Wert des Zielfelds.
$KNNP-*Feldname*	Wahrscheinlichkeit des vorhergesagten Werts.
$KNNP-*Wert*	Wahrscheinlichkeit jedes möglichen Werts von nominalen oder Flagfelds. Wird nur angezeigt, wenn auf der Registerkarte "Einstellungen" des Modellnuggets die Option **Alle Wahrscheinlichkeiten anhängen** aktiviert ist.
$KNN-Nachbar-*n*	Der Name des *n*. nächsten Nachbarn zum Fokusdatensatz. Wird nur eingeschlossen, wenn für **Anzeigen von Nearest** auf der Registerkarte "Einstellungen" des Modellnuggets ein Wert ungleich null angegeben ist.

Tabelle 26. KNN-Modellfelder (Forts.)

Neuer Feldname	Beschreibung
$KNN-Distanz-*n*	Die relative Distanz des Fokusdatensatzes vom *n*. nächsten Nachbarn. Wird nur eingeschlossen, wenn für **Anzeigen von Nearest** auf der Registerkarte "Einstellungen" des Modellnuggets ein Wert ungleich null angegeben ist.

Modellansicht "Nächste Nachbarn"

Modellansicht

Das Fenster der Modellansicht setzt sich aus zwei Bereichen zusammen:

- Im ersten Bereich wird eine Übersicht des Modells, die sogenannte Hauptansicht, angezeigt.
- Im zweiten Bereich wird eine der beiden folgenden Ansichten angezeigt:

  Die Hilfsmodellansicht enthält mehr Informationen zum Modell, ist dafür aber weniger stark auf das Modell an sich konzentriert.

  Die verknüpfte Ansicht zeigt Details zu einer bestimmten Funktion des Modells an, wenn der Benutzer einen Teil der Hauptansicht ansteuert.

Standardmäßig wird im ersten Bereich der Prädiktorbereich und im zweiten Bereich das Diagramm für den Prädiktoreinfluss angezeigt. Wenn das Diagramm für den Prädiktoreinfluss nicht verfügbar ist (d. h. wenn **Merkmale nach Bedeutung gewichten** nicht im Bereich "Nachbarn" der Registerkarte "Einstellungen" ausgewählt wurde), wird die erste verfügbare Ansicht aus dem Dropdown-Menü "Ansicht" angezeigt.

Wenn für eine Ansicht keine Informationen verfügbar sind, wird sie im Dropdown-Menü "Ansicht" nicht angezeigt.

Prädiktorbereich: Das Prädiktorbereichsdiagramm ist ein interaktives Diagramm für den Prädiktorbereich (bzw. -unterbereich, bei mehr als drei Prädiktoren). Jede Achse stellt einen Prädiktor im Modell dar und die Position der Punkte in der Tabelle gibt die Werte dieser Prädiktoren für Fälle in den Trainings- und Holdout-Partitionen an.

Schlüssel. Neben den Prädiktorwerten liefern die Punkte im Diagramm weitere Informationen.

- Die Form gibt die Partition an, zu der ein Punkt gehört (Training oder Holdout).
- Die Farbe/Schattierung eines Punkts gibt den Wert des Ziels für diesen Fall an. Dabei entsprechen eindeutige Farbwerte den Kategorien eines kategorialen Ziels und Schattierungen dem Wertebereich eines stetigen Ziels. Für Trainings-Partitionen ist der angegebene Wert der festgestellte Wert. Für Holdout-Partitionen handelt es sich um den vorhergesagten Wert. Wenn kein Ziel angegeben ist, wird diese Erläuterung nicht angezeigt.
- Kräftigere Umrisse weisen auf Fokusfälle hin. Fokusdatensätze werden im Zusammenhang mit ihren *k* nächstgelegenen Nachbarn angezeigt.

Steuerelemente und Interaktivität. Sie können den Prädiktorbereich mit einer Reihe an Steuerelementen im Diagramm untersuchen.

- Sie können festlegen, welches Subset an Prädiktoren im Diagramm angezeigt werden soll, und ändern, welche Prädiktoren in den Dimensionen dargestellt werden.
- "Fokusdatensätze" sind Punkte, die im Diagramm "Prädiktorraum" ausgewählt werden. Wenn Sie eine Fokusdatensatzvariable angegeben haben, werden zuerst die Punkte ausgewählt, die die Fokusdatensätze darstellen. Es kann jedoch jeder Punkt vorübergehend ein Fokusdatensatz werden, wenn Sie ihn auswählen. Die "gängigen" Steuerelemente für die Punktauswahl gelten. Wenn Sie auf einen Punkt klicken, wird dieser Punkt ausgewählt und alle anderen Punkte abgewählt. Wenn Sie die Steuertaste drü-

cken und auf einen Punkt klicken, wird der Punkt zu dem Set der ausgewählten Punkte hinzugefügt. Verknüpfte Ansichten wie das Peers-Diagramm werden automatisch mit den Fällen aktualisiert, die im Prädiktorbereich ausgewählt werden.

- Sie können die Anzahl an für Fokusdatensätze anzuzeigenden nächstgelegenen Nachbarn (*k*) ändern.
- Wenn Sie die Maus über einen Punkt im Diagramm bewegen, wird eine QuickInfo mit dem Wert der Fallbeschriftung oder, wenn keine Fallbeschriftungen definiert sind, der Fallnummer und dem festgestellten und vorhergesagten Zielwert angezeigt.
- Mit der Schaltfläche "Zurücksetzen" können Sie den Prädiktorraum in seinen ursprünglichen Zustand zurückversetzen.

Ändern der Achsen im Prädiktorbereichsdiagramm: Sie können steuern, welche Funktionen an den Achsen im Prädiktorbereichsdiagramm angezeigt werden.

So ändern Sie die Achseneinstellungen:

1. Klicken Sie auf die Schaltfläche für den Bearbeitungsmodus (Pinselsymbol) im linken Bereich, um den Modus "Bearbeiten" für den Prädiktorbereich zu wählen.
2. Ändern Sie die Ansicht (beliebig) im rechten Bereich. Der Bereich **Zonen anzeigen** wird zwischen den beiden Hauptbereichen angezeigt.
3. Klicken Sie auf das Kontrollkästchen **Zonen anzeigen**.
4. Klicken Sie auf einen beliebigen Datenpunkt im Prädiktorbereich.
5. So ersetzen Sie einen Prädiktor durch eine Funktion desselben Datentyps:
   - Ziehen Sie den neuen Prädiktor über die Zonenbeschriftung (mit der kleinen Schaltfläche "X"), das Sie ersetzen möchten.
6. So ersetzen Sie eine Achse durch einen Prädiktor eines anderen Datentyps:
   - Klicken Sie bei der Zonenbeschriftung des zu ersetzenden Prädiktors auf die kleine Schaltfläche "X". Der Prädiktorbereich ändert sich in eine zweidimensionale Ansicht.
   - Ziehen Sie den neuen Prädiktor auf die Zonenbeschriftung **Dimension hinzufügen**.
7. Klicken Sie auf die Schaltfläche für den Interaktionsmodus (Pfeilspitzensymbol) im linken Bereich, um den Modus "Bearbeiten" zu verlassen.

Prädiktoreinfluss: In der Regel konzentriert man sich bei der Modellerstellung auf die Prädiktorfelder, die am wichtigsten sind, und vernachlässigt jene, die weniger wichtig sind. Dabei unterstützt Sie das Wichtigkeitsdiagramm für die Prädiktoren, da es die relative Wichtigkeit der einzelnen Prädiktoren für das Modell angibt. Da die Werte relativ sind, beträgt die Summe der Werte aller Prädiktoren im Diagramm 1,0. Die Wichtigkeit der Prädiktoren steht in keinem Bezug zur Genauigkeit des Modells. Sie bezieht sich lediglich auf die Wichtigkeit der einzelnen Prädiktoren für eine Vorhersage und nicht auf die Genauigkeit der Vorhersage.

Abstände zwischen nächstgelegenen Nachbarn: Diese Tabelle zeigt nur die *k* nächstgelegenen Nachbarn und Abstände für Fokusdatensätze an. Sie ist verfügbar, wenn eine Fokusdatensatz-ID im Modellierungsknoten angegeben wurde, und zeigt nur Fokusdatensätze, die durch diese Variable identifiziert werden.

Jede Zeile der:

- Die Spalte **Fokusdatensatz** enthält den Wert der Fallbeschriftungsvariablen für den Fokusdatensatz. Wenn keine Fallbeschriftungen angegeben wurden, enthält diese Spalte die Fallnummer des Fokusdatensatzes.
- Die *i*-te Spalte unter der Gruppe **Nächste Nachbarn** enthält den Wert der Fallbeschriftungsvariablen für den *i*-ten nächsten Nachbarn des Fokusdatensatzes. Wenn keine Fallbeschriftungen definiert sind, enthält diese Spalte die Fallnummer des *i*-ten nächsten Nachbarn des Fokusdatensatzes.
- Die *i*-te Spalte unter der Gruppe **Nächste Distanzen** enthält die Distanz des *i*-ten nächsten Nachbarn zum Fokusdatensatz.

Peers: Dieses Diagramm enthält die Fokusfälle und ihre *k* nächstgelegenen Nachbarn für jeden Prädiktor im Ziel. Es ist verfügbar, wenn ein Fokusfall im Prädiktorbereich ausgewählt ist.

Das Vergleichsdiagramm ist auf zwei Arten mit dem Prädiktorbereich verknüpft.

- Im Peers-Diagramm werden die im Prädiktorbereich gewählten Fokusfälle sowie ihre *k* nächstgelegenen Nachbarn angezeigt.
- Der Wert *k* wird im Prädiktorbereich gewählt und im Peers-Diagramm herangezogen.

Prädiktoren auswählen. Ermöglicht Ihnen die Auswahl der Prädiktoren für die Anzeige im Vergleichsdiagramm.

Quadrantenkarte: Dieses Diagramm zeigt die Fokusfälle und ihre *k* nächstgelegenen Nachbarn als Streudiagramm (oder Punktdiagramm, je nach Messniveau des Ziels) mit dem Ziel auf der *y*-Achse und einem metrischen Prädiktor auf der *x*-Achse nach Prädiktoren in einzelne Felder unterteilt an. Es ist verfügbar, wenn ein Ziel vorhanden und ein Fokusfall im Prädiktorbereich ausgewählt ist.

- Für stetige Variablen werden bei den Mittelwerten der Variablen in der Trainingspartition Referenzlinien gezogen.

Prädiktoren auswählen. Ermöglicht Ihnen, die Prädiktoren für die Anzeige in der Quadrantenkarte auszuwählen.

Prädiktor-Auswahlfehler-Protokoll: Punkte im Diagramm zeigen den Fehler (je nach Messniveau des Ziels entweder die Fehlerrate oder den Quadratsummenfehler) auf der *y*-Achse für das Modell mit dem Prädiktor auf der *x*-Achse an (plus allen Prädiktoren weiter links auf der *x*-Achse). Dieses Diagramm ist verfügbar, wenn ein Ziel und eine Merkmalauswahl aktiviert sind.

Klassifikationstabelle: Diese Tabelle enthält die Kreuzklassifikation der festgestellten Werte im Vergleich zu den vorhergesagten Werten des Ziels nach Partitionen. Verfügbar, wenn ein Ziel vorhanden ist und es kategorial ist (Flag, nominal oder ordinal).

- Die Zeile **(Fehlend)** in der Holdout-Partition enthält Holdout-Fälle mit fehlenden Werten im Ziel. Diese Fälle tragen zu den "Prozent insgesamt"-Werten, aber nicht den "Gesamtprozent korrekt"-Werten der Holdout-Stichprobe bei.

Fehlerzusammenfassung: Diese Tabelle ist verfügbar, wenn eine Zielvariable vorhanden ist. Sie enthält den Fehler, der dem Modell zugeordnet ist, Quadratsummen für ein stetiges Ziel und die Fehlerrate (100 % – Gesamtprozent korrekt) für ein kategoriales Ziel.

KNN-Modell-Einstellungen

Auf der Registerkarte "Einstellungen" können Sie zusätzliche Felder angeben, die bei der Anzeige der Ergebnisse verwendet werden sollen (z. B. durch Ausführen eines Tabellenknotens, der mit dem Nugget verknüpft ist). Sie können den Effekt jeder dieser Optionen sehen, indem Sie sie auswählen und auf die Schaltfläche "Vorschau" klicken. Führen Sie in der Vorschau-Ausgabe einen Bildlauf nach rechts durch, um die zusätzlichen Felder zu sehen.

Alle Wahrscheinlichkeiten anhängen (nur gültig für kategoriale Ziele). Wenn diese Option ausgewählt ist, werden für jeden Datensatz, der vom Knoten verarbeitet wird, die Wahrscheinlichkeiten für jeden möglichen Wert eines Zielfelds vom Typ "Nominal" oder "Flag" angezeigt. Wenn diese Option nicht ausgewählt ist, werden für Zielfelder vom Typ "Flag" oder "Nominal" nur der vorhergesagte Wert und seine Wahrscheinlichkeit angezeigt.

Die Standardeinstellung dieses Kontrollkästchens richtet sich nach dem entsprechenden Kontrollkästchen des Modellierungsknotens.

Raw-Propensity-Scores berechnen. Bei Modellen mit einem Flagziel (das als Vorhersage "Ja" bzw. "Nein" ausgibt) können Sie Propensity-Scores anfordern, die die Wahrscheinlichkeit des für das Zielfeld angege-

benen wahren Ergebnisses angeben. Diese Werte werden zusätzlich zu den anderen Vorhersage- und Konfidenzwerten angegeben, die gegebenenfalls während des Scorings erstellt werden.

Adjusted-Propensity-Scores berechnen. Raw-Propensity-Scores beruhen ausschließlich auf den Trainingsdaten und sind aufgrund der Neigung vieler Modelle zur übermäßigen Anpassung an die Trainingsdaten möglicherweise zu optimistisch. Bei Adjusted Propensitys wird versucht, dies durch Evaluation der Modellleistung anhand einer Test- bzw. Validierungspartition zu kompensieren. Bei dieser Option muss im Stream ein Partitionsfeld definiert sein und die Adjusted-Propensity-Scores müssen im Modellierungsknoten aktiviert werden, bevor das Modell generiert wird.

Nächsten anzeigen. Wenn Sie diesen Wert auf *n* einstellen (mit *n* als positive Ganzzahl ungleich null), werden die *n* nächsten Nachbarn des Fokusdatensatzes zusammen mit ihren Distanzen vom Fokusdatensatz in das Modell aufgenommen.

Bemerkungen

Diese Informationen wurden für weltweit angebotene Produkte und Dienstleistungen erarbeitet.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite „Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Markten oder eingetragene Marken der Intel Corporation oder der Tochtergesellschaften des Unternehmens in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA, anderen Ländern oder beidem.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein.

Glossar

A

AICC . Ein Maß für die Auswahl und den Vergleich von gemischten Modellen, das auf -2 (Restricted) Log-Likelihood beruht. Kleinere Werte stehen für bessere Modelle. Das AICC "korrigiert" das AIC für kleine Stichprobenumfänge. Wenn die Stichprobengröße zunimmt, konvergiert das AICC zu dem AIC.

B

Bayes-Informationskriterium (BIC) . Ein Maß für die Auswahl und den Vergleich von Modellen, das auf -2 Log-Likelihood beruht. Kleinere Werte stehen für bessere Modelle. Das BIC bestraft ebenfalls überparametrisierte Modelle, und zwar stärker als das AIC.

Box' M-Test . Ein Test auf Gleichheit der Kovarianzmatrizen der Gruppen. Bei hinreichend großen Stichproben bedeutet ein nicht signifikanter p-Wert, dass die Anhaltspunkte für unterschiedliche Matrizen nicht ausreichend sind. Der Test ist empfindlich gegenüber Abweichungen von der multivariaten Normalverteilung.

F

Fälle . Für jeden Fall werden Codes für die tatsächliche Gruppe, die vorhergesagte Gruppe, A-posteriori-Wahrscheinlichkeiten und Diskriminanzscores angezeigt.

Fisher . Zeigt die Koeffizienten der Klassifizierungsfunktion nach Fisher an, die direkt für die Klassifizierung verwendet werden können. Es wird ein eigenes Set von Koeffizienten der Klassifizierungsfunktion für jede Gruppe ermittelt. Ein Fall wird der Gruppe zugewiesen, für die er den größten Diskriminanzscore (Klassifizierungsfunktionswert) aufweist.

H

Hazarddiagramm . Zeigt die kumulative Hazardfunktion auf einer linearen Skala an.

K

Klassifikationsergebnisse . Die Anzahl der Fälle, die auf Grundlage der Diskriminanzanalyse jeder der Gruppen richtig oder falsch zugeordnet werden. Zuweilen auch als Konfusionsmatrix bezeichnet.

Kombiniertes Streudiagramm aller Gruppen . Erzeugt ein alle Gruppen umfassendes Streudiagramm der Werte für die ersten beiden Diskriminanzfunktionen. Wenn nur eine Funktion vorliegt, wird stattdessen ein Histogramm angezeigt.

Kovarianz . Ein nicht standardisiertes Maß für den Zusammenhang zwischen zwei Variablen. Es ist gleich der Kreuzproduktabweichung geteilt durch N-1.

Korrelationsmatrix innerhalb der Gruppen . Zeigt eine in Pools zusammengefasste Korrelationsmatrix innerhalb der Gruppen an, die als Durchschnitt der separaten Kovarianzmatrizen für alle Gruppen vor der Berechnung der Korrelationen bestimmt wird.

Kovarianz innerhalb der Gruppen . Zeigt eine Pools zusammengefasste Kovarianzmatrix innerhalb der Gruppen an, die sich von der Gesamtkovarianzmatrix unterscheiden kann. Die Matrix wird als Mittel der einzelnen Kovarianzmatrizen für alle Gruppen berechnet.

Kurtosis . Ein Maß dafür, wie sehr die Beobachtungen um einen zentralen Punkt gruppiert sind. Bei einer Normalverteilung ist der Wert der Kurtosis gleich 0. Bei positiver Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung enger um das Zentrum der Verteilung gruppiert und haben dünnere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der leptokurtischen Verteilung im Vergleich zu einer Normalverteilung dicker. Bei negativer Kurtosis sind die Beobachtungen im Vergleich zu einer Normalverteilung weniger eng

gruppiert und haben dickere Flanken bis hin zu den Extremwerten der Verteilung. Ab dort sind die Flanken der platykurtischen Verteilung im Vergleich zu einer Normalverteilung dünner.

L

Klassifikation mit Fallauslassung . Jeder Fall der Analyse wird durch Funktionen aus allen anderen Fällen unter Auslassung dieses Falls klassifiziert. Diese Klassifikation wird auch als "U-Methode" bezeichnet.

M

MAE . Mittlerer absoluter Fehler. Er misst, wie stark die Zeitreihe von ihrem durch das Modell vorhergesagten Niveau abweicht. MAE wird in derselben Maßeinheit angegeben wie die ursprüngliche Zeitreihe.

Mahalanobis-Distanz . Dieses Maß gibt an, wie weit die Werte der unabhängigen Variablen eines Falls vom Mittelwert aller Fälle abweichen. Eine große Mahalanobis-Distanz charakterisiert einen Fall, der bei einer oder mehreren unabhängigen Variablen Extremwerte besitzt.

MAPE . Mittlerer absoluter Fehler in Prozent. Ein Maß dafür, wie stark eine abhängige Zeitreihe von ihrem durch das Modell vorhergesagten Niveau abweicht. Es ist unabhängig von den verwendeten Maßeinheiten und kann daher verwendet werden, um Zeitreihen mit unterschiedlichen Einheiten zu vergleichen.

MaxAE . Maximaler absoluter Fehler (Maximum Absolute Error, also maximaler Betrag des Fehlers). Dies ist der größte vorhergesagte Fehler, ausgedrückt in derselben Maßeinheit wie die abhängige Zeitreihe. Genau wie MaxAPE hilft er dabei, sich ein Worst-Case-Szenario für die Vorhersagen vorzustellen. Der maximale absolute Fehler und der maximale absolute Fehler in Prozent können an verschiedenen Punkten in der Zeitreihe auftreten, beispielsweise wenn der absolute Fehler für einen großen Zeitreihenwert geringfügig größer ist als der absolute Fehler für einen kleinen Zeitreihenwert. In diesem Fall tritt der maximale absolute Fehler beim größeren Zeitreihenwert und der maximale absolute Fehler in Prozent beim kleineren Zeitreihenwert auf.

MaxAPE . Maximaler absoluter Fehler in Prozent (Maximum Absolute Percentage Error, also maximaler Betrag des relativen Fehlers). Dies ist der größte vorhergesagte Fehler, ausgedrückt in Prozent. Dieses Maß hilft dabei, sich ein Worst-Case-Szenario für die Vorhersagen vorzustellen.

Variablenaufnahme durch Maximieren des kleinsten F-Quotienten . Eine Methode für die Variablenauswahl in einer schrittweisen Analyse. Sie beruht auf der Maximierung eines F-Quotienten, der aus der Mahalanobis-Distanz zwischen den Gruppen errechnet wird.

Maximum . Der größte Wert einer numerischen Variablen.

Mittelwert . Ein Lagemaß (zentrale Tendenz). Die Summe der Ränge, geteilt durch die Zahl der Fälle.

Mittelwerte . Zeigt Gesamt- und Gruppenmittelwerte sowie Standardabweichungen für die unabhängigen Variablen an.

Median . Wert, über und unter dem jeweils die Hälfte der Fälle liegt; 50. Perzentil. Bei einer geraden Anzahl von Fällen ist der Median der Mittelwert der beiden mittleren Fälle, wenn diese auf- oder absteigend sortiert sind. Der Median ist ein Lagemaß, das gegenüber Ausreißern unempfindlich ist (im Gegensatz zum Mittelwert, der durch wenige extrem niedrige oder hohe Werte beeinflusst werden kann).

Wilks-Lambda minimieren . Eine Auswahlmethode für Variablen bei der schrittweisen Diskriminanzanalyse. Die Aufnahme von Variablen in die Gleichung erfolgt anhand der jeweiligen Verringerung von Wilks-Lambda. Bei jedem Schritt wird diejenige Variable aufgenommen, die den Gesamtwert von Wilks-Lambda am meisten vermindert.

Minimum . Der kleinste Wert einer numerischen Variablen.

Modalwert . Der am häufigsten auftretende Wert. Wenn mehrere Werte gleichermaßen die größte Häufigkeit aufweisen, ist jeder von ihnen ein Modalwert.

N

Normalisiertes BIC . Normalisiertes Bayes-Informationskriterium (BIC). Ein allgemeines Maß der insgesamt erreichten Güte der Anpassung, das auch die Komplexität des Modells zu berücksichtigen versucht. Es ist ein Score, der auf dem mittleren quadratischen Fehler beruht und eine Penalisierung für die Anzahl der Modellparameter und die Länge der Zeitreihe enthält. Die Penalisierung neutralisiert die Überlegenheit von Modellen mit einer größeren Anzahl von Parametern und macht die Statistik damit gut vergleichbar für verschiedene Modelle derselben Zeitreihe.

O

Eins-minus-Überleben . Erzeugt ein Diagramm der Werte "1 - Überlebensfunktion" auf einer linearen Skala.

R

Bereich . Die Differenz zwischen den größten und kleinsten Werten einer numerischen Variablen; Maximalwert minus Minimalwert.

Rao-V (Diskriminanzanalyse) . Ein Maß für die Unterschiede zwischen Gruppenmittelwerten. Auch Lawley-Hotelling-Spur genannt. Bei jedem Schritt wird die Variable aufgenommen, die den Anstieg des Rao-V maximiert. Wenn Sie diese Option ausgewählt haben, geben Sie den Minimalwert ein, den eine Variable für die Aufnahme in die Analyse aufweisen muss.

RMSE . Steht für Root Mean Square Error. Die Quadratwurzel des mittleren quadratischen Fehlers. Ein Maß dafür, wie stark eine abhängige Zeitreihe von ihrem durch das Modell vorhergesagten Niveau abweicht, und zwar ausgedrückt in derselben Maßeinheit wie die abhängige Zeitreihe.

R-Quadrat . Ein Maß für die Anpassungsgüte eines linearen Modells. Wird auch als Bestimmtheitskoeffizient bezeichnet. Es gibt den Anteil der Variation der abhängigen Variablen an, der durch das Regressionsmodell erklärt wird. Er liegt zwischen 0 und 1. Kleine Werte zeigen an, dass das Modell nicht gut zu den Daten passt.

S

Gruppenspezifisch . Für die Klassifizierung werden gruppenspezifische Kovarianzmatrizen verwendet. Da die Klassifizierung auf Diskriminanzfunktionen und nicht auf ursprünglichen Variablen basiert, entspricht diese Option nicht immer der Verwendung einer quadratischen Diskriminanzfunktion.

Kovarianz der einzelnen Gruppen . Zeigt separate Kovarianzmatrizen für jede Gruppe an.

Gruppenspezifische Diagramme . Erzeugt gruppenspezifische Streudiagramme der Werte für die ersten beiden Diskriminanzfunktionen. Wenn nur eine Funktion vorliegt, werden stattdessen Histogramme angezeigt.

Bonferroni sequenziell . Hierbei handelt es sich um ein sequentielles schrittweises Bonferroni-Verfahren, das deutlich weniger konservativ ist was die Ablehnung einzelner Hypothesen anbelangt, aber dennoch dasselbe allgemeine Signifikanzniveau beibehält.

Sidak (sequenziell) . Hierbei handelt es sich um ein sequentielles schrittweises Sidak-Verfahren, das deutlich weniger konservativ ist was die Ablehnung einzelner Hypothesen anbelangt, aber dennoch dasselbe allgemeine Signifikanzniveau beibehält.

Schiefe . Ein Maß der Asymmetrie der Verteilung. Die Normalverteilung ist symmetrisch, ihre Schiefe hat den Wert 0. Eine Verteilung mit einer deutlichen positiven Schiefe läuft nach rechts lang aus (lange rechte Flanke). Eine Verteilung mit einer deutlichen negativen Schiefe läuft nach links lang aus (lange linke Flanke). Als Faustregel kann man verwenden, dass ein Schiefewert, der mehr als doppelt so groß ist wie sein Standardfehler, als Abweichung von der Symmetrie gilt.

Standardabweichung . Ein Maß für die Streuung um den Mittelwert, definiert als Quadratwurzel aus der Varianz. Die Standardabweichung wird in denselben Einheiten gemessen wie die ursprüngliche Variable.

Standardabweichung . Ein Maß für die Streuung um den Mittelwert. In einer Normalverteilung liegen 68 % der Fälle innerhalb von einer Standardabweichung des Mittelwerts und 95 % der Fälle innerhalb von zwei Standardabwei-

chungen. Wenn beispielsweise für das Alter der Mittelwert 45 und die Standardabweichung 10 beträgt, liegen bei einer Normalverteilung 95 % der Fälle im Bereich zwischen 25 und 65.

Standardfehler . Ein Maß für die Abweichung des Werts einer Teststatistik zwischen Stichproben. Dies ist die Standardabweichung der Stichprobenverteilung einer Statistik. So ist z. B. der Standardfehler des Mittelwerts die Standardabweichung des Stichprobenmittelwerts.

Standardfehler der Kurtosis . Der Quotient aus der Kurtosis und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) Ein großer positiver Wert für die Kurtosis deutet darauf hin, dass die Flanken der Verteilung länger sind als bei einer Normalverteilung; ein negativer Wert bedeutet, dass sie kürzer sind (etwa wie bei einer kastenförmigen, gleichförmigen Verteilung).

Standardfehler des Mittelwerts . Ein Maß dafür, wie stark der Mittelwert von Stichprobe zu Stichprobe in derselben Verteilung variieren kann. Dieser Wert kann für einen ungefähren Vergleich des beobachteten Mittelwerts mit einem hypothetischen Wert verwendet werden. (Es kann geschlossen werden, dass die beiden Werte unterschiedlich sind, wenn das Verhältnis der Differenz zum Standardfehler kleiner als -2 oder größer als +2 ist.)

Standardfehler der Schiefe . Der Quotient aus der Schiefe und deren Standardfehler kann als Test auf Normalverteilung verwendet werden. (Sie können die Normalverteilung ausschließen, wenn der Quotient unter -2 oder über +2 liegt.) Ein großer positiver Wert für die Schiefe bedeutet, dass die Verteilung eine lange rechte Flanke hat; ein extremer negativer Wert bedeutet, dass sie eine lange linke Flanke hat.

R-Quadrat für stationären Teil . Ein Maß, das den stationären Teil des Modells mit einem einfachen Mittelwertmodell vergleicht. Dieses Maß ist dem gewöhnlichen R-Quadrat vorzuziehen, wenn ein Trend oder ein saisonales Muster vorliegt. R-Quadrat für den stationären Teil kann auch negativ sein, es nimmt Werte zwischen minus unendlich und 1 an. Negative Werte bedeuten, dass das betrachtete Modell schlechter ist als das Basismodell. Positive Werte bedeuten, dass das betrachtete Modell besser ist als das Basismodell.

Summe . Die Summe der Werte über alle Fälle mit nicht fehlenden Werten.

Überlebensdiagramm . Zeigt die kumulative Überlebensfunktion auf einer linearen Skala an.

T

Territorien . Ein Diagramm der Grenzen, mit denen Fälle auf der Grundlage von Funktionswerten in Gruppen klassifiziert werden. Die Zahlen entsprechen den Gruppen, in die die Fälle klassifiziert wurden. Der Mittelwert jeder Gruppe wird durch einen darin liegenden Stern (*) angezeigt. Dieses Diagramm wird nicht angezeigt, wenn nur eine einzige Diskriminanzfunktion vorliegt.

Gesamte Kovarianz . Zeigt die Kovarianzmatrix für alle Fälle an, so als wären sie aus einer einzigen Stichprobe.

U

Nicht erklärte Varianz . Bei jedem Schritt wird die Variable aufgenommen, welche die Summe der nicht erklärten Streuung zwischen den Gruppen minimiert.

Eindeutig . Bewertet alle Effekte gleichzeitig; damit werden alle Effekte an alle sonstigen Effekte jedweden Typs angepasst.

Univariate ANOVA . Führt für jede unabhängige Variable eine einfaktorielle Varianzanalyse durch, d. h. einen Test auf Gleichheit der Gruppenmittelwerte.

Nicht standardisiert . Zeigt die nicht standardisierten Koeffizienten der Diskriminanzfunktion an.

F-Wert verwenden . Eine Variable wird in ein Modell aufgenommen, wenn ihr F-Wert größer als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn der F-Wert kleiner als der Ausschlusswert ist. Der Aufnahmewert muss größer sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen in das Modell aufzunehmen, senken Sie den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, erhöhen Sie den Ausschlusswert.

F-Wahrscheinlichkeit verwenden . Eine Variable wird in das Modell aufgenommen, wenn das Signifikanzniveau ihres F-Werts kleiner als der Aufnahmewert ist. Sie wird ausgeschlossen, wenn das Signifikanzniveau größer als der Ausschlusswert ist. Der Aufnahmewert muss kleiner sein als der Ausschlusswert und beide Werte müssen positiv sein. Um mehr Variablen in das Modell aufzunehmen, erhöhen Sie den Aufnahmewert. Um mehr Variablen aus dem Modell auszuschließen, senken Sie den Ausschlusswert.

V

Gültig . Gültige Fälle, d. h. solche, die weder den systemdefiniert fehlenden Wert noch einen benutzerdefiniert fehlenden Wert aufweisen.

Varianz . Ein Maß der Streuung um den Mittelwert, gleich der Summe der quadrierten Abweichungen vom Mittelwert geteilt durch eins weniger als die Anzahl der Fälle. Die Maßeinheit der Varianz ist das Quadrat der Maßeinheiten der Variablen.

W

Innerhalb der Gruppen . Zur Klassifizierung von Fällen wird die in Pools zusammengefasste Kovarianzmatrix innerhalb der Gruppen verwendet.

Index

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z