IBM SPSS Modeler 14.2-Anwendungshandbuch

Hinweis: Lesen Sie vor der Verwendung dieser Informationen und des zugehörigen Produkts die allgemeinen Informationen unter Hinweise auf S. .

Vorwort

IBM® SPSS® Modeler ist die auf Unternehmensebene einsetzbare Data-Mining-Workbench von IBM Corp.. 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 benutzen 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 Benutzeroberfläche von SPSS Modeler erleichtert die Anwendung des spezifischen Geschäftswissens der Benutzer, was zu leistungsstärkeren Vorhersagemodellen führt und die Zeit bis zur Lösungserstellung verkürzt. SPSS Modeler bietet zahlreiche Modellierungsverfahren, beispielsweise Algorithmen für Vorhersage, Klassifizierung, Segmentierung und Assoziationserkennung. Nach der Modellerstellung ermöglicht IBM® SPSS® Modeler Solution Publisher die unternehmensweite Bereitstellung für Entscheidungsträger oder in einer Datenbank.

Über IBM Business Analytics

IBM Business Analytics-Software bietet vollständige, einheitliche und genaue Informationen, auf die Entscheidungsträger vertrauen, um die Unternehmensleistung zu steigern. Ein umfassendes Portfolio von Anwendungen für Unternehmensinformationen, Vorhersageanalysen, Verwaltung der Finanzleistung und Strategie sowie Analysen bietet sofort klare und umsetzbare Einblicke in die aktuelle Leistung und ermöglicht die Vorhersage zukünftiger Ergebnisse. In Kombination mit umfassenden Branchenlösungen, bewährten Vorgehensweisen und professionellen Dienstleistungen können Unternehmen jeder Größe optimale Produktivität erreichen, die Entscheidungsfindung zuverlässig automatisieren und bessere Ergebnisse erzielen.

Als Teil dieses Portfolios unterstützt die IBM SPSS Predictive Analytics-Software Unternehmen dabei, zukünftige Ereignisse vorherzusagen und aktiv auf diese Erkenntnisse zu reagieren, um bessere Geschäftsergebnisse zu erzielen. Kunden aus den Bereichen Wirtschaft, Behörden und Bildung aus aller Welt verlassen sich auf die IBM SPSS-Technologie. Sie bringt Ihnen beim Gewinnen, Halten und Ausbauen neuer Kundenbeziehungen einen Wettbewerbsvorteil und verringert gleichzeitig das Betrugs- sowie andere Risiken. Durch Integration der IBM SPSS-Software in den täglichen Betrieb können diese Unternehmen qualifizierte Vorhersagen treffen und dadurch die Entscheidungsfindung so ausrichten und automatisieren, dass Geschäftsziele erreicht werden und ein messbarer Wettbewerbsvorteil entsteht. Wenn Sie weitere Informationen wünschen oder einen Mitarbeiter kontaktieren möchten, ist dies unter *http://www.ibm.com/spss* möglich.

Technischer Support

Kunden mit Wartungsvertrag können den technischen Support in Anspruch nehmen. Kunden können sich an den technischen Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Corp.-Produkten oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Die Kontaktdaten des Technischen Supports finden Sie auf der IBM Corp.-Website

unter *http://www.ibm.com/support*. Sie müssen bei der Kontaktaufnahme Ihren Namen, Ihre Organisation und Ihre Supportvereinbarung angeben.

Inhalt

Kapitel 1

Informationen zu IBM SPSS Modeler

IBM® SPSS® Modeler ist ein Set von Data-Mining-Tools, mit dem Sie auf der Grundlage Ihres Geschäftswissens schnell und einfach Vorhersagemodelle erstellen und zur Erleichterung der Entscheidungsfindung in die Betriebsabläufe einbinden können. 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.

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.

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

IBM SPSS Modeler Server

SPSS Modeler verwendet eine Client/Server-Architektur zur Verteilung von Anforderungen für ressourcenintensive Vorgänge an leistungsstarke Serversoftware, wodurch bei größeren Daten-Sets eine schnellere Leistung erzielt werden kann. Neben den hier aufgeführten Produkten können auch weitere Produkte bzw. Aktualisierungen verfügbar sein. Weitere Informationen finden Sie unter *http://www.ibm.com/software/analytics/spss/products/modeler/*.

SPSS Modeler. SPSS Modeler Clementine Client ist eine im Funktionsumfang vollständige Version des installierten Produkts und kann auf dem Desktop-Computer des Benutzers ausgeführt werden. Es kann im lokalen Modus als Standalone-Produkt oder im verteilten Modus zusammen mit IBM® SPSS® Modeler Server verwendet werden, um im Falle von großen Daten-Sets die Leistung zu verbessern.

SPSS Modeler Server. SPSS Modeler Server wird ständig im verteilten Analysemodus zusammen mit einer oder mehreren IBM® SPSS® Modeler-Installationen ausgeführt, wodurch eine herausragende Leistung bei großen Daten-Sets erzielt werden kann, da speicherintensive Operationen auf dem Server durchgeführt werden können, ohne Daten auf den Client-Computer herunterzuladen. SPSS Modeler Server bietet außerdem Unterstützung für SQL-Optimierung sowie Funktionen zur Modellierung innerhalb der Datenbank, wodurch Leistungsfähigkeit und Automatisierung weiter verbessert werden. Es muss mindestens eine SPSS Modeler-Installation vorhanden sein, um eine Analyse durchzuführen.

IBM SPSS Modeler-Optionen

Die folgenden Komponenten und Funktionen können separat erworben und für die Verwendung mit SPSS Modeler lizenziert werden. Beachten Sie, dass zu einem späteren Zeitpunkt möglicherweise noch weitere Produkte und Updates erhältlich sind. Weitere Informationen finden Sie unter *http://www.ibm.com/software/analytics/spss/products/modeler/*.

- SPSS Modeler Server-Zugriff, der eine bessere Skalierbarkeit und bessere Leistungsfähigkeit bei großen Daten-Sets bietet sowie Unterstützung für SQL-Optimierung und Funktionen zur Modellierung innerhalb der Datenbank.
- SPSS Modeler Solution Publisher, für Scoring in Echtzeit oder automatisiertes Scoring außerhalb der SPSS Modeler-Umgebung. Für weitere Informationen siehe Thema IBM SPSS Modeler Solution Publisher in Kapitel 2 in *IBM SPSS Modeler 14.2 Solution Publisher*.
- Adapter zum Deployment für IBM SPSS Collaboration and Deployment Services oder die Thin-Client-Anwendung IBM SPSS Modeler Advantage. Für weitere Informationen siehe Thema Speichern und Bereitstellen von IBM SPSS Collaboration and Deployment Services Repository-Objekten in Kapitel 9 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

IBM SPSS Text Analytics

IBM® SPSS® Text Analytics ist ein vollständig integriertes Zusatzprodukt für SPSS Modeler, das hoch entwickelte linguistische Technologien und die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) benutzt, 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.

- Der Text-Mining-Knoten bietet die Modellierung von Konzepten und Kategorien sowie eine interaktive Workbench, in der Sie eine erweiterte Untersuchung von Textlinks und Clustern durchführen, Ihre eigenen Kategorien erstellen und die Vorlagen für linguistische Ressourcen verfeinern können.
- Eine Reihe von Importformaten wird unterstützt, darunter Blogs und andere webbasierte Quellen.
- Benutzerdefinierte Vorlagen, Bibliotheken und Wörterbücher für bestimmte Domänen, wie CRM und Genomforschung, sind ebenfalls eingeschlossen.

Anmerkung: Für den Zugriff auf diese Komponente ist eine separate Lizenz erforderlich. Weitere Informationen finden Sie unter *http://www.ibm.com/software/analytics/spss/products/modeler/*.

IBM SPSS Modeler-Dokumentation

Die vollständige Dokumentation im Online-Hilfe-Format finden Sie im Hilfe-Menü von SPSS Modeler. Dazu gehören 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 im PDF-Format finden Sie im Ordner *\Documentation* auf der jeweiligen Produkt-DVD.

- **IBM SPSS Modeler-Benutzerhandbuch.** Allgemeine Einführung in die Verwendung von SPSS Modeler, in der u. a. die Erstellung von Daten-Streams, der Umgang mit fehlenden Werten, die Erstellung von CLEM-Ausdrücken, die Arbeit mit Projekten und Berichten sowie das Packen von Streams für das Deployment in IBM SPSS Collaboration and Deployment Services, Predictive Applications (Prognoseanwendungen) oder IBM SPSS Modeler Advantage beschrieben werden.
- **Quellen-, Prozess- und Ausgabeknoten in IBM SPSS Modeler.** 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. Für weitere Informationen siehe Thema Überblick über Modellierungsknoten in Kapitel 3 in *IBM SPSS Modeler 14.2-Modellierungsknoten*.
- **IBM SPSS Modeler-Algorithmushandbuch.** Beschreibung der mathematischen Grundlagen der in SPSS Modeler verwendeten Modellierungsmethoden.
- **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 Hilfe-Menü aufgerufen werden. Für weitere Informationen siehe Thema Anwendungsbeispiele in *IBM SPSS Modeler 14.2-Benutzerhandbuch*.
- **Skripterstellung und Automatisierung in IBM SPSS Modeler.** Informationen zur Automatisierung des Systems über Skripterstellung, einschließlich der Eigenschaften, die zur Bearbeitung von Knoten und Streams verwendet werden können.
- **IBM SPSS Modeler Deployment-Handbuch.** Informationen zum Ausführen von SPSS Modeler-Streams und -Szenarien 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 SPSS Modeler zu integrieren.
- **In-Database Mining-Handbuch für IBM SPSS Modeler.** 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- und -Leistungshandbuch.** Informationen zur Konfiguration und Verwaltung von IBM® SPSS® Modeler Server.

- **IBM SPSS Modeler Administration Console – Benutzerhandbuch.** Informationen zur Installation und Nutzung der Konsolen-Benutzeroberfläche zur Überwachung und Konfiguration von SPSS Modeler Server. Die Konsole ist als Plugin für die Deployment Manager-Anwendung implementiert.
- **IBM SPSS Modeler Solution Publisher-Handbuch.** SPSS Modeler Solution Publisher ist eine Zusatzkomponente, mit der Unternehmen Streams zur Verwendung außerhalb der SPSS Modeler-Standardumgebung veröffentlichen können.
- **IBM SPSS Modeler-Handbuch zu CRISP-DM.** Schritt-für-Schritt-Anleitung für das Data-Mining mit SPSS Modeler unter Verwendung der CRISP-DM-Methode.

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 Daten-Sets sind viel kleiner als die riesigen 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 Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Für weitere Informationen siehe Thema Ordner “Demos” in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Beispiele für die Datenbank-Modellierung. Die Beispiele finden Sie im *IBM SPSS Modeler In-Database Mining-Handbuch*.

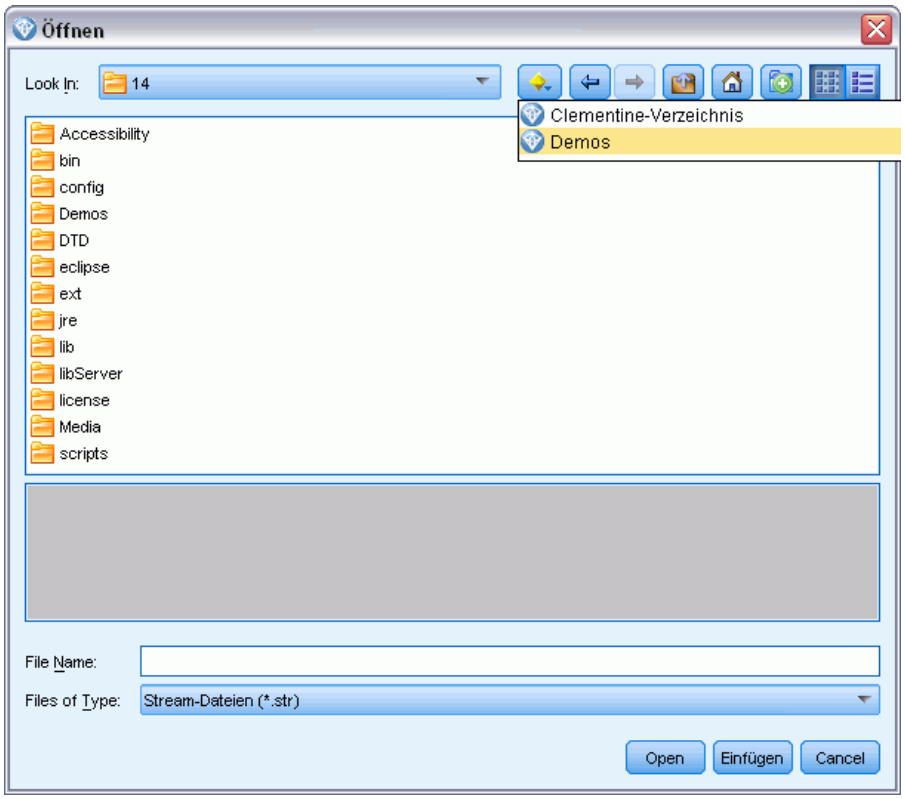
Skriptbeispiele. Die Beispiele finden Sie im *IBM SPSS Modeler Handbuch für die Skripterstellung und Automatisierung*.

Ordner “Demos”

Die in den Anwendungsbeispielen verwendeten Datendateien und Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Auf diesen Ordner können Sie auch über die Programmgruppe IBM SPSS Modeler 14.2 im

Windows-Startmenü oder durch Klicken auf *Demos* in der Liste der zuletzt angezeigten Verzeichnisse im Dialogfeld "Datei öffnen" zugreifen.

Abbildung 1-1
Auswahl des Ordners "Demos" in der Liste der zuletzt angezeigten Verzeichnisse



Teil I:
Einführung und erste Schritte

Kapitel
2

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 Daten-Sets sind viel kleiner als die riesigen 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 Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Für weitere Informationen siehe Thema Ordner "Demos" in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Beispiele für die Datenbank-Modellierung. Die Beispiele finden Sie im *IBM SPSS Modeler In-Database Mining-Handbuch*.

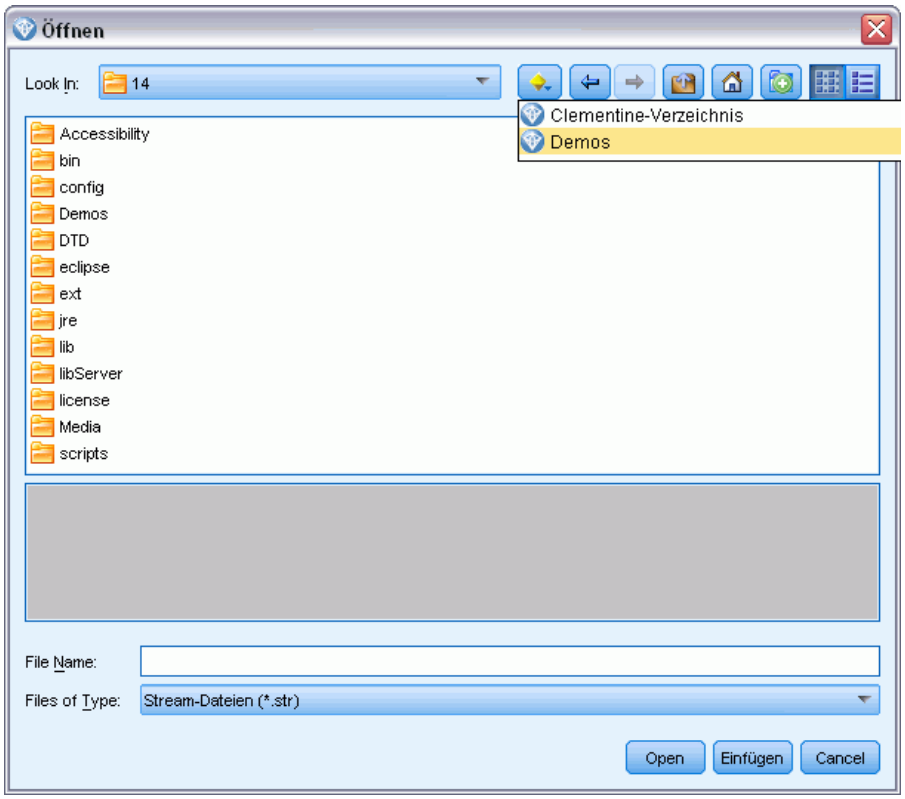
Skriptbeispiele. Die Beispiele finden Sie im *IBM SPSS Modeler Handbuch für die Skripterstellung und Automatisierung*.

Ordner "Demos"

Die in den Anwendungsbeispielen verwendeten Datendateien und Beispiel-Streams wurden im Ordner *Demos*, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Auf diesen Ordner können Sie auch über die Programmgruppe IBM SPSS Modeler 14.2 im

Windows-Startmenü oder durch Klicken auf *Demos* in der Liste der zuletzt angezeigten Verzeichnisse im Dialogfeld "Datei öffnen" zugreifen.

Abbildung 2-1
Auswahl des Ordners "Demos" in der Liste der zuletzt angezeigten Verzeichnisse



Kapitel
3

IBM SPSS Modeler-Überblick

Erste Schritte

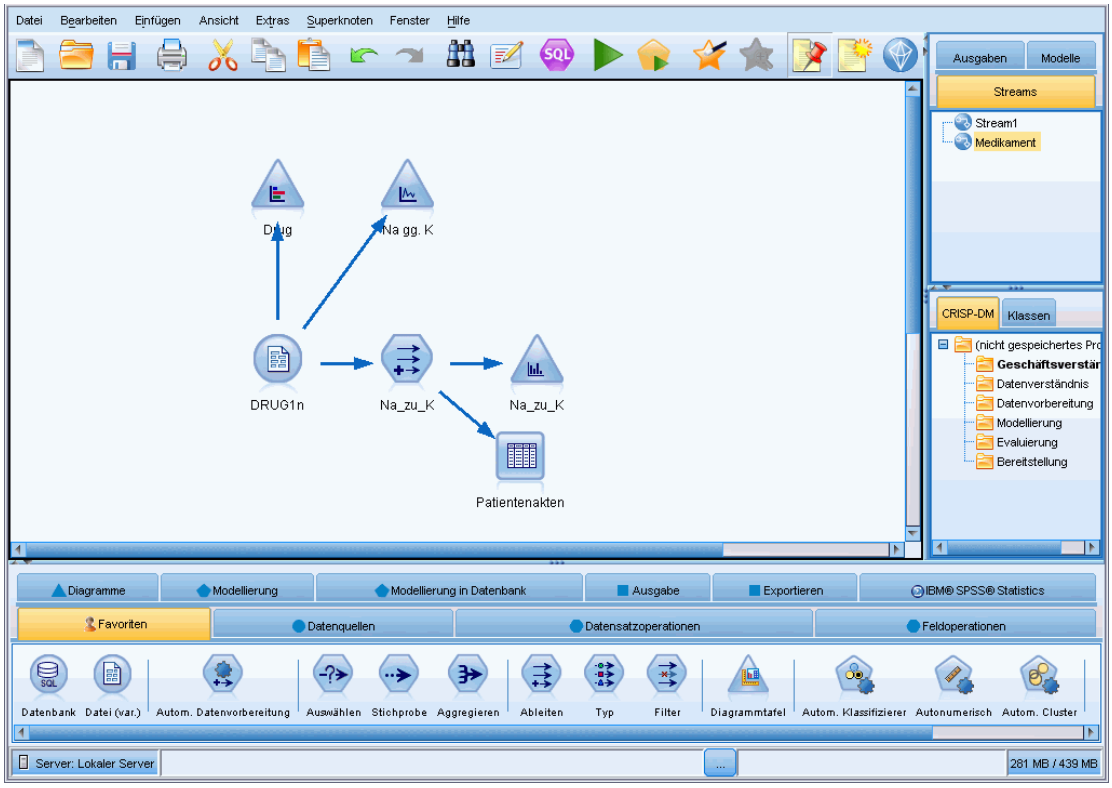
Als Data-Mining-Anwendung bietet IBM® SPSS® Modeler eine strategische Methode zum Auffinden nützlicher Beziehungen in großen Datenbeständen. Im Unterschied zu herkömmlicheren statistischen Methoden müssen Sie zu Beginn nicht unbedingt wissen, wonach Sie suchen. Sie können Ihre Daten durch Anpassen verschiedener Modelle und Überprüfen unterschiedlicher Beziehungen so lange untersuchen, bis Sie auf nützliche Informationen stoßen.

Starten von IBM SPSS Modeler

Klicken Sie zum Starten der Anwendung auf:
Start > [Alle] Programme > IBM SPSS Modeler14.2 > IBM SPSS Modeler14.2

Nach einigen Sekunden wird das Hauptfenster angezeigt.

Abbildung 3-1
IBM SPSS Modeler-Hauptfenster



Starten über die Befehlszeile

Sie können die Befehlszeile Ihres Betriebssystems wie folgt verwenden, um IBM® SPSS® Modeler zu starten:

- Öffnen Sie auf einem Computer, auf dem IBM® SPSS® Modeler installiert ist, ein DOS- oder Befehlszeilenfenster.
- Um die SPSS Modeler-Schnittstelle im interaktiven Modus zu starten, geben Sie den Befehl modelerclient und dann die gewünschten Argumente ein, z. B.:

```
modelerclient -stream report.str -execute
```

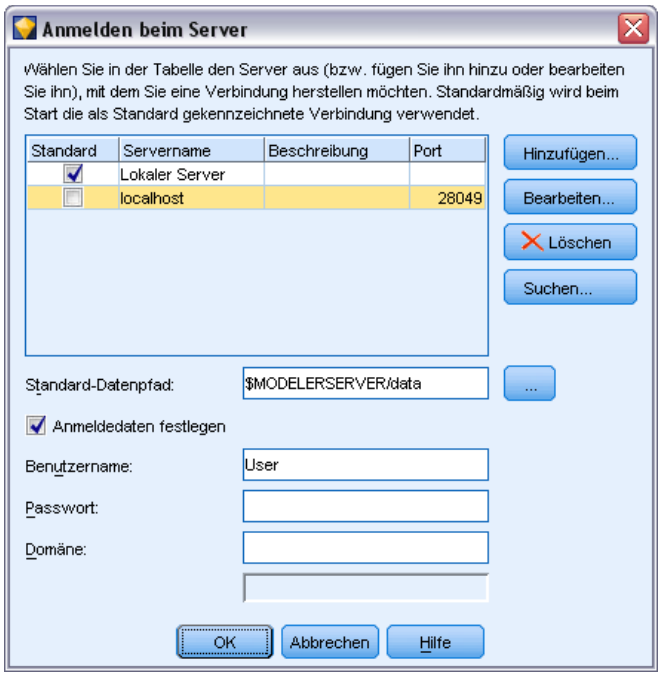
Mithilfe der verfügbaren Argumente (Flags) können Sie eine Verbindung zu einem Server herstellen, Streams laden, Skripts ausführen oder je nach Bedarf weitere Parameter angeben.

Verbindung mit IBM SPSS Modeler Server

IBM® SPSS® Modeler kann als eigenständige Anwendung oder als Client ausgeführt werden, der direkt mit IBM® SPSS® Modeler Server oder über das Plugin Coordinator of Processes von IBM® SPSS® Collaboration and Deployment Services mit einem SPSS Modeler Server oder Server-Cluster verbunden ist. Der aktuelle Verbindungsstatus wird unten links im SPSS Modeler-Fenster angezeigt.

Wenn Sie eine Verbindung zu einem Server herstellen möchten, können Sie den Namen des Servers, mit dem eine Verbindung hergestellt werden soll, manuell eingeben oder einen zuvor definierten Namen auswählen. Wenn Sie IBM SPSS Collaboration and Deployment Services verwenden, können Sie im Dialogfeld für die Anmeldung beim Server eine Liste von Servern bzw. Server-Clustern durchsuchen. Die Möglichkeit, die auf einem Netzwerk ausgeführten Statistics-Dienste zu durchsuchen, wird über den Coordinator of Processes bereitgestellt. Für weitere Informationen siehe Thema Lastenausgleich mit Server-Clustern in Anhang D in *IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch*.

Abbildung 3-2
Dialogfeld "Anmelden beim Server"



So stellen Sie eine Verbindung mit einem Server her:

- Klicken Sie im Menü "Extras" auf die Option Anmelden beim Server. Das Dialogfeld "Anmelden beim Server" wird geöffnet. Alternativ können Sie auf den Bereich des Verbindungsstatus im SPSS Modeler-Fenster doppelklicken.

- Legen Sie in diesem Dialogfeld die Optionen zum Verbinden mit dem lokalen Servercomputer fest oder wählen Sie eine Verbindung in der Tabelle aus.
  - Klicken Sie auf Hinzufügen bzw. Bearbeiten, um eine Verbindung hinzuzufügen bzw. zu bearbeiten. Für weitere Informationen siehe Thema Hinzufügen und Bearbeiten der IBM SPSS Modeler Server-Verbindung in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.
  - Klicken Sie auf Suche, um auf einen Server bzw. Server-Cluster in Coordinator of Processes zuzugreifen. Für weitere Informationen siehe Thema Suchen nach Servern in IBM SPSS Collaboration and Deployment Services in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Servertabelle. Diese Tabelle enthält die Menge der definierten Serververbindungen. In der Tabelle werden Standardverbindung, Servername, Beschreibung und Portnummer angegeben. Sie können manuell eine neue Verbindung hinzufügen sowie eine bestehende Verbindung auswählen bzw. danach suchen. Um einen bestimmten Server als Standardverbindung einzurichten, aktivieren Sie in der Tabelle für die Verbindung das Kontrollkästchen in der Spalte "Standard".

Standard-Datenpfad. Geben Sie einen Pfad an, der für Daten auf dem Server-Rechner verwendet wird. Mit der Auslassungsschaltfläche (...) wechseln Sie zum gewünschten Verzeichnis.

Anmeldedaten einstellen. Belassen Sie dieses Kontrollkästchen nicht aktiviert, um die Funktion für die **Einzelanmeldung** zu aktivieren. Diese versucht, Sie mithilfe Ihres lokalen Benutzernamens und Passworts beim Server anzumelden. Falls die Einzelanmeldung nicht möglich ist oder Sie das Kontrollkästchen zur Deaktivierung der Einzelanmeldung aktivieren (z. B. zur Anmeldung in ein Administratorkonto), wird ein weiteres Fenster angezeigt, in dem Sie aufgefordert werden, Ihre Anmeldedaten einzugeben.

Benutzer-ID. Geben Sie den Benutzernamen ein, mit dem die Anmeldung beim Server erfolgen soll.

Paßwort. Geben Sie das Passwort ein, das zum angegebenen Benutzernamen gehört.

Domäne. Geben Sie die Domäne an, mit der die Anmeldung beim Server erfolgen soll. Ein Domänenname ist nur dann erforderlich, wenn sich der Server-Computer in einer anderen Windows-Domäne befindet als der Client-Computer.

▶ Klicken Sie auf OK, um die Verbindung herzustellen.

So trennen Sie eine Verbindung mit einem Server:

▶ Klicken Sie im Menü “Extras” auf die Option Anmelden beim Server. Das Dialogfeld “Anmelden beim Server” wird geöffnet. Alternativ können Sie auf den Bereich des Verbindungsstatus im SPSS Modeler-Fenster doppelklicken.

▶ Wählen Sie im Dialogfeld den lokalen Server aus und klicken Sie auf OK.

Hinzufügen und Bearbeiten der IBM SPSS Modeler Server-Verbindung

Serververbindungen können manuell im Kontrollkästchen “Anmelden beim Server” bearbeitet bzw. hinzugefügt werden. Durch Klicken auf “Hinzufügen” können Sie auf ein leeres Dialogfeld vom Typ “Server hinzufügen/bearbeiten” zugreifen, in dem Sie Details zur Serververbindung eingeben können. Durch Auswahl einer bestehenden Verbindung und Klicken auf “Bearbeiten” im Dialogfeld “Anmelden beim Server” wird das Dialogfeld “Server hinzufügen/bearbeiten” mit den Details für die betreffende Verbindung geöffnet, sodass Sie etwaige Änderungen vornehmen können.

Hinweis: Serververbindungen, die aus IBM® SPSS® Collaboration and Deployment Services hinzugefügt wurden, können nicht bearbeitet werden, da Name, Port und sonstige Details in IBM SPSS Collaboration and Deployment Services definiert sind.

Abbildung 3-3
Dialogfeld “Anmeldung beim Server: Server hinzufügen/bearbeiten”

So fügen Sie Serververbindungen hinzu:

- Klicken Sie im Menü “Extras” auf die Option Anmelden beim Server. Das Dialogfeld “Anmelden beim Server” wird geöffnet.
- Klicken Sie in diesem Dialogfeld auf Hinzufügen. Das Dialogfeld “Anmeldung beim Server: Server hinzufügen/bearbeiten” wird angezeigt.
- Geben Sie die Details für die Serververbindung ein und klicken Sie auf OK, um die Verbindung zu speichern und zum Dialogfeld “Anmeldung beim Server” zurückzukehren.
  - **Server.** Geben Sie einen verfügbaren Server an oder wählen Sie einen aus der Liste aus. Der Server-Computer lässt sich anhand eines alphanumerischen Namens (z. B. *meinserver*) oder der dem Server-Computer zugewiesenen IP-Adresse (z. B. 202.123.456.78) identifizieren.
  - **Port.** Geben Sie die Portnummer an, die der Server überwacht. Wenn der Standardwert nicht funktioniert, fragen Sie Ihren Systemadministrator nach der richtigen Portnummer.
  - **Beschreibung.** Geben Sie eine optionale Beschreibung für diese Serververbindung ein.
  - **Verbindung verschlüsseln (mit SSL).** Legt fest, ob eine SSL-Verbindung (**Secure Sockets Layer**) verwendet werden soll. SSL ist ein weit verbreitetes Protokoll zum Schutz der über ein Netzwerk versendeten Daten. Um diese Funktion verwenden zu können, muss SSL auf dem Server, auf dem sich IBM® SPSS® Modeler Server befindet, aktiviert sein. Wenden Sie sich gegebenenfalls an den lokalen Administrator, wenn Sie weitere Informationen benötigen.

So bearbeiten Sie Serververbindungen:

- Klicken Sie im Menü “Extras” auf die Option Anmelden beim Server. Das Dialogfeld “Anmelden beim Server” wird geöffnet.
- Wählen Sie in diesem Dialogfeld die zu bearbeitende Verbindung aus und klicken Sie dann auf Bearbeiten. Das Dialogfeld “Anmeldung beim Server: Server hinzufügen/bearbeiten” wird angezeigt.
- Ändern Sie die Details für die Serververbindung und klicken Sie auf OK, um die Änderungen zu speichern und zum Dialogfeld “Anmeldung beim Server” zurückzukehren.

Suchen nach Servern in IBM SPSS Collaboration and Deployment Services

Anstatt eine Serververbindung manuell einzugeben, können Sie einen im Netzwerk verfügbaren Server oder Server-Cluster über Coordinator of Processes auswählen. Diese Funktion ist in IBM® SPSS® Collaboration and Deployment Services verfügbar. Ein Server-Cluster ist eine Gruppe von Servern, aus denen Coordinator of Processes den Server ermittelt, der am besten für die Beantwortung einer Verarbeitungsanforderung geeignet ist. Für weitere Informationen siehe Thema Lastenausgleich mit Server-Clustern in Anhang D in *IBM SPSS Modeler Server 14.2-Verwaltungs- und Leistungshandbuch*.

Sie können zwar auch manuell Server im Dialogfeld “Anmelden beim Server” hinzufügen, durch die Suche nach verfügbaren Servern können Sie jedoch eine Verbindung zu Servern herstellen, ohne den richtigen Servernamen und die Portnummer kennen zu müssen. Diese Informationen werden automatisch bereitgestellt. Allerdings benötigen Sie auch bei dieser Variante die richtigen Anmeldeinformationen, wie Benutzername, Domäne und Passwort.

Hinweis: Wenn Sie keinen Zugriff auf die Funktion Coordinator of Processes haben, können Sie dennoch den Namen des Servers, mit dem eine Verbindung hergestellt werden soll, manuell eingeben oder einen zuvor definierten Namen auswählen. Für weitere Informationen siehe Thema Hinzufügen und Bearbeiten der IBM SPSS Modeler Server-Verbindung in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Abbildung 3-4
Dialogfeld "Nach Servern suchen"

So suchen Sie nach Servern und Clustern:

- Klicken Sie im Menü "Extras" auf die Option Anmelden beim Server. Das Dialogfeld "Anmelden beim Server" wird geöffnet.
- Klicken Sie in diesem Dialogfeld auf Suche, um das Dialogfeld "Nach Servern suchen" zu öffnen. Wenn Sie versuchen, Coordinator of Processes zu durchsuchen, ohne bei IBM SPSS Collaboration and Deployment Services angemeldet zu sein, werden Sie zur Anmeldung aufgefordert. Für weitere Informationen siehe Thema Verbindung mit dem IBM SPSS Collaboration and Deployment Services Repository in Kapitel 9 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.
- Wählen Sie den Server bzw. Server-Cluster in der Liste aus.
- Klicken Sie auf OK, um das Dialogfeld zu schließen und diese Verbindung zu der Tabelle im Dialogfeld "Anmelden beim Server" hinzuzufügen.

Ändern des temporären Verzeichnisses

Für einige in IBM® SPSS® Modeler Server durchgeführten Operationen müssen u. U. temporäre Dateien erstellt werden. IBM® SPSS® Modeler verwendet standardmäßig das temporäre Systemverzeichnis zum Erstellen von temporären Dateien. Sie können das temporäre Verzeichnis wie folgt ändern:

- Erstellen Sie ein neues Verzeichnis namens *spss* und ein Unterverzeichnis namens *servertemp*.
- Bearbeiten Sie die Datei *options.cfg*, die sich im Unterverzeichnis */config* Ihres SPSS Modeler-Installationsverzeichnisses befindet. Bearbeiten Sie den Parameter temp_directory in diesem Feld, damit er wie folgt lautet: temp_directory, "C:/spss/servertemp".

- Anschließend müssen Sie den SPSS Modeler Server-Dienst neu starten. Dies ist möglich, wenn Sie auf die Registerkarte Dienste Ihrer Windows-Systemsteuerung klicken. Halten Sie den Dienst einfach an und starten Sie ihn erneut, um die von Ihnen durchgeführten Änderungen zu aktivieren. Durch das Neustarten des Rechners wird auch der Dienst neu gestartet.

  Alle temporären Dateien werden nun in dieses neue Verzeichnis geschrieben.

  Hinweis: Der häufigste Fehler bei diesem Vorgang besteht darin, dass der falsche Schrägstrich verwendet wird. Aufgrund der UNIX-Vergangenheit von SPSS Modeler werden normale Schrägstriche (/) verwendet.

Starten mehrerer IBM SPSS Modeler-Sitzungen

Wenn Sie mehrere IBM® SPSS® Modeler-Sitzungen gleichzeitig starten möchten, müssen Sie einige Änderungen an den IBM® SPSS® Modeler- und Windows-Einstellungen vornehmen. Dies kann beispielsweise erforderlich sein, wenn Sie zwei separate Serverlizenzen besitzen und zwei Streams vom selben Client-Computer aus mit zwei verschiedenen Servern ausführen möchten.

So aktivieren Sie mehrere SPSS Modeler-Sitzungen:

- Klicken Sie auf:
  Start > [Alle] Programme > IBM SPSS Modeler14.2
- Klicken Sie mit der rechten Maustaste auf die Verknüpfung “IBM SPSS Modeler14.2” (die mit dem Symbol) und wählen Sie die Option Eigenschaften.
- Fügen Sie im Textfeld Ziel die Option -noshare am Ende der Zeichenfolge hinzu.
- Wählen Sie im Windows Explorer folgende Optionen:
  Werkzeuge > Ordneroptionen...
- Wählen Sie auf der Registerkarte “Dateitypen” die Option für SPSS Modeler-Streams und klicken Sie auf Erweitert.
- Wählen Sie im Dialogfeld “Dateityp bearbeiten” die Option Öffnen mit SPSS Modeler und klicken Sie auf Bearbeiten.
- Fügen Sie im Textfeld Anwendung für diesen Vorgang vor dem Argument -stream das Argument -noshare ein.

IBM SPSS Modeler-Oberfläche auf einen Blick

An jedem Punkt des Data-Mining-Prozesses können Sie auf der benutzerfreundlichen Benutzeroberfläche von IBM® SPSS® Modeler Ihr spezielles Fachwissen einbringen. Modellierungsalgorithmen, wie Vorhersage, Klassifizierung, Segmentierung und Assoziationserkennung, gewährleisten leistungsstarke und genaue Modelle. Die Modellergebnisse können problemlos angewendet und in Datenbanken, in IBM® SPSS® Statistics und in einer Vielzahl anderer Anwendungen eingelesen werden.

Die Datenarbeit mit SPSS Modeler besteht aus drei Schritten.

- Zunächst lesen Sie Daten in SPSS Modeler ein.
- Anschließend lassen Sie die Daten eine Reihe von Bearbeitungen durchlaufen.
- Schließlich senden Sie die Daten an ein Ziel.

Diese Reihenfolge wird als **Daten-Stream** bezeichnet, da die Daten Datensatz für Datensatz von der Quelle durch jeden Bearbeitungsschritt zum Ziel fließen, was entweder zu einer Datenausgabe vom Typ “Modell” oder “Typ” führt.

Abbildung 3-5
Ein einfacher Stream

Stream-Zeichenbereich von IBM SPSS Modeler

Der Stream-Zeichenbereich ist der größte Bereich des IBM® SPSS® Modeler-Fensters. Hier erstellen und bearbeiten Sie Daten-Streams.

Die Erstellung von Streams erfolgt durch Zeichnen von Diagrammen von Datenoperationen, die für Ihren Geschäftsbetrieb relevant sind, auf den Hauptzeichenbereich in der Benutzeroberfläche. Jede Operation wird durch ein Symbol oder einen **Knoten** dargestellt, und die Knoten sind in einem **Stream** miteinander verbunden, der den Datenfluss durch jede Operation darstellt.

Sie können in SPSS Modeler mit mehreren Streams gleichzeitig arbeiten, entweder im selben Stream-Zeichenbereich oder durch Öffnen eines neuen Stream-Zeichenbereichs. Während einer Sitzung werden die Streams im Stream-Manager, rechts oben im SPSS Modeler-Fenster gespeichert.

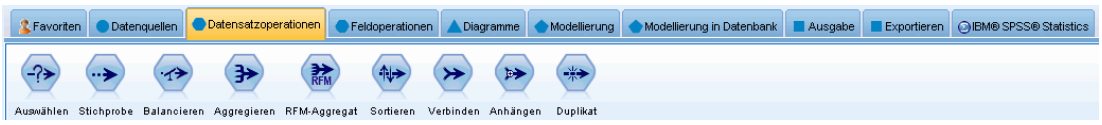
Knotenpalette

Die meisten Daten und Modellierungstools in IBM® SPSS® Modeler befinden sich in der **Knotenpalette**, unten im Fenster, unterhalb des Stream-Zeichenbereichs.

Die Palettenregisterkarte “Datensatzoperationen” beispielsweise enthält Knoten, mit denen Sie Operationen auf die **Datensätze** anwenden können, wie beispielsweise Auswählen, Zusammenführen (Mergen) und Anhängen.

Um Knoten zum Zeichenbereich hinzuzufügen, doppelklicken Sie in den Knoten-Paletten auf die entsprechenden Symbole oder ziehen Sie sie auf den Zeichenbereich. Anschließend verbinden Sie sie, um einen **Stream** zu erstellen, der den Datenfluss darstellt.

Abbildung 3-6
Die Registerkarte “Datensatzoperationen” in der Knotenpalette



Jede Palettenregisterkarte enthält eine Sammlung verwandter Knoten, die für verschiedene Phasen der Stream-Operationen verwendet werden, wie:

- **Datenquellen.** Diese Knoten lesen Daten in SPSS Modeler ein.
- **Datensatzoperationen.** Diese Knoten führen Operationen an **Datensätzen** durch, wie beispielsweise Auswählen, Zusammenführen (Mergen) und Anhängen.
- **Feldoperationen.** Diese Knoten führen Operationen an **Datenfeldern** durch, wie beispielsweise Filtern, Ableiten neuer Felder und Festlegen des Messniveaus für bestimmte Felder.
- **Diagramme.** Diese Knoten bieten eine grafische Darstellung der Daten vor und nach der Modellierung. Diagramme umfassen Plots, Histogramme, Netzdiagrammknoten und Evaluierungsdiagramme.
- **Modellierung.** Diese Knoten verwenden die in SPSS Modeler verfügbaren Modellierungsalgorithmen, wie neuronale Netze, Entscheidungsbäume, Clusteralgorithmen und Datensequenzierung.
- **Datenbank-Modellierung.** Knoten verwenden die Modellierungsalgorithmen, die in Microsoft SQL Server-, IBM DB2- und Oracle-Datenbanken verfügbar sind.
- **Ausgabe.** Knoten erzeugen verschiedenste Ausgabetypen für Daten, Diagramme und Modellergebnisse, die in SPSS Modeler angezeigt werden können.
- **Exportieren.** Knoten erzeugen verschiedenste Ausgabetypen, die in externen Anwendungen angezeigt werden können, z. B. in IBM® SPSS® Data Collection oder Excel.
- **SPSS Statistics.** Knoten importieren Daten aus IBM® SPSS® Statistics oder exportieren sie nach SPSS Statistics und führen SPSS Statistics-Prozeduren aus.

Je vertrauter Sie im Umgang mit SPSS Modeler werden, desto besser können Sie den Paletteninhalt für Ihre eigene Verwendung anpassen. Für weitere Informationen siehe Thema Anpassen der Knotenpalette in Kapitel 12 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

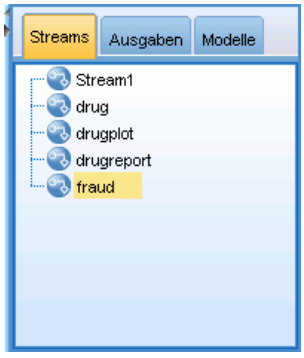
Der Berichtsbereich, der sich unter der Knotenpalette befindet, bietet Feedback zum Fortschritt verschiedener Operationen, z. B. wann die Daten in den Daten-Stream eingelesen werden. Außerdem bietet der Statusbereich, der sich ebenfalls unterhalb der Knotenpalette befindet, Informationen über die aktuelle Aktivität der Anwendung sowie Anweisungen, wann ein Benutzerfeedback erforderlich ist.

IBM SPSS Modeler-Manager.

Oben rechts im Fenster befindet sich der Manager-Bereich. Dieser enthält drei Registerkarten, die zum Verwalten von Streams, Ausgaben und Modellen verwendet werden.

Sie können die Registerkarte “Streams” verwenden, um die in einer Sitzung erstellten Streams zu öffnen, umzubenennen, zu speichern und zu löschen.

Abbildung 3-7
Registerkarte “Streams”

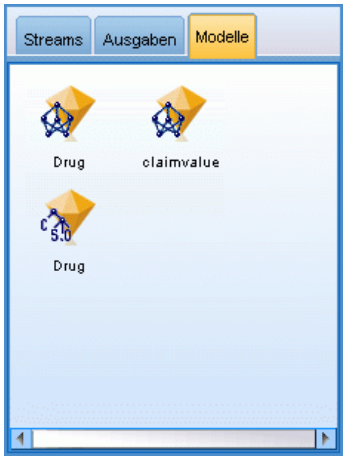


Die Registerkarte “Ausgaben” enthält eine Reihe von Dateien, wie beispielsweise Diagramme und Tabellen, die von den Stream-Operationen in IBM® SPSS® Modeler produziert wurden. Sie können die auf dieser Registerkarte aufgeführten Tabellen, Diagramme und Berichte anzeigen, speichern, umbenennen und schließen.

Abbildung 3-8
Registerkarte “Ausgaben”

Die Registerkarte “Modelle” ist die leistungsstärkste der Manager-Registerkarten. Diese Registerkarte enthält sämtliche Modell-**Nuggets**, die die in SPSS Modeler erstellten Modelle enthalten, für die aktuelle Sitzung. Auf der Registerkarte “Modelle” können diese Modelle direkt durchsucht oder dem Stream im Zeichenbereich hinzugefügt werden.

Abbildung 3-9
Registerkarte "Modelle" mit Modell-Nuggets



IBM SPSS Modeler-Projekte

Unten rechts im Fenster befindet sich der Projektbereich, der zum Erstellen und Verwalten von Data-Mining-**Projekten** verwendet wird (Gruppen von Dateien, die in Bezug zu einer Data-Mining-Aufgabe stehen). Es gibt zwei Methoden zur Ansicht von in IBM® SPSS® Modeler erstellten Projekten—die Ansicht "Klassen" und die Ansicht "CRISP-DM".

Die Registerkarte "CRISP-DM" bietet ein Verfahren zur Organisation von Projekten gemäß dem Cross-Industry Standard Process for Data Mining, einer in der Branche bewährten, nicht eigentumsrechtlich geschützten Methode. Sowohl erfahrene Data-Mining-Experten als auch Neulinge können mit dem CRISP-DM-Tool ihre Arbeit besser organisieren und an andere weitergeben.

Abbildung 3-10
CRISP-DM, Ansicht

Die Registerkarte “Klassen” bietet eine Methode, mit der Sie Ihre Arbeit in SPSS Modeler (nach den erstellten Objekttypen) in Kategorien organisieren können. Diese Ansicht ist hilfreich für das Inventarisieren von Daten, Streams und Modellen.

Abbildung 3-11
Ansicht “Klassen”

IBM SPSS Modeler-Symbolleiste

Oben im IBM® SPSS® Modeler-Fenster sehen Sie eine Symbolleiste, die eine Vielzahl nützlicher Funktionen bietet. Im Folgenden werden die Schaltflächen der Symbolleiste und ihre Funktionen beschrieben.


Neuen Stream erstellen


Stream öffnen


Stream speichern


Aktuellen Stream drucken


Ausschneiden & in die Zwischenablage verschieben


In Zwischenablage kopieren


Einfügen der Auswahl


Letzte Aktion rückgängig


Wiederholen


Knoten suchen


Stream-Eigenschaften bearbeiten

Vorschau für SQL-Erzeugung

	Aktuellen Stream ausführen		Stream-Auswahl ausführen
	Stream anhalten (wird nur während der Stream-Ausführung aktiv)		Superknoten hinzufügen
	Vergrößern (nur Superknoten)		Verkleinern (nur Superknoten)
	Kein Markup in Stream		Kommentar einfügen
	Stream-Markup ausblenden (falls vorhanden)		Ausgeblendeten Stream-Markup einblenden
	Stream in IBM® SPSS® Modeler Advantage öffnen		

Stream-Markup umfasst Stream-Kommentare, Modellverknüpfungen und die Anzeige von Scoring-Verzweigungen.

Weitere Informationen zu Stream-Kommentaren erhalten Sie unter Hinzufügen von Kommentaren und Anmerkungen zu Knoten und Streams auf S. .

Weitere Informationen zur Markierung von Scoring-Verzweigungen erhalten Sie unter Scoring-Verzweigung auf S. .

Eine Beschreibung zu Modellverknüpfungen finden Sie im Handbuch zu den *IBM SPSS-Modellierungsknoten*.

Anpassen der Symbolleiste

Sie können zahlreiche Aspekte der Symbolleiste ändern, z. B.:

- ob sie angezeigt wird,
- ob für die Symbole eine QuickInfo verfügbar ist,
- ob große oder kleine Symbole angezeigt werden.

So schalten Sie die Anzeige der Symbolleiste ein bzw. aus:

- Klicken Sie im Hauptmenü auf Folgendes:
  Ansicht > Symbolleiste > Anzeigen

So ändern Sie die Einstellungen für QuickInfo oder Symbolgröße:

- Klicken Sie im Hauptmenü auf Folgendes:
  Ansicht > Symbolleiste > Anpassen

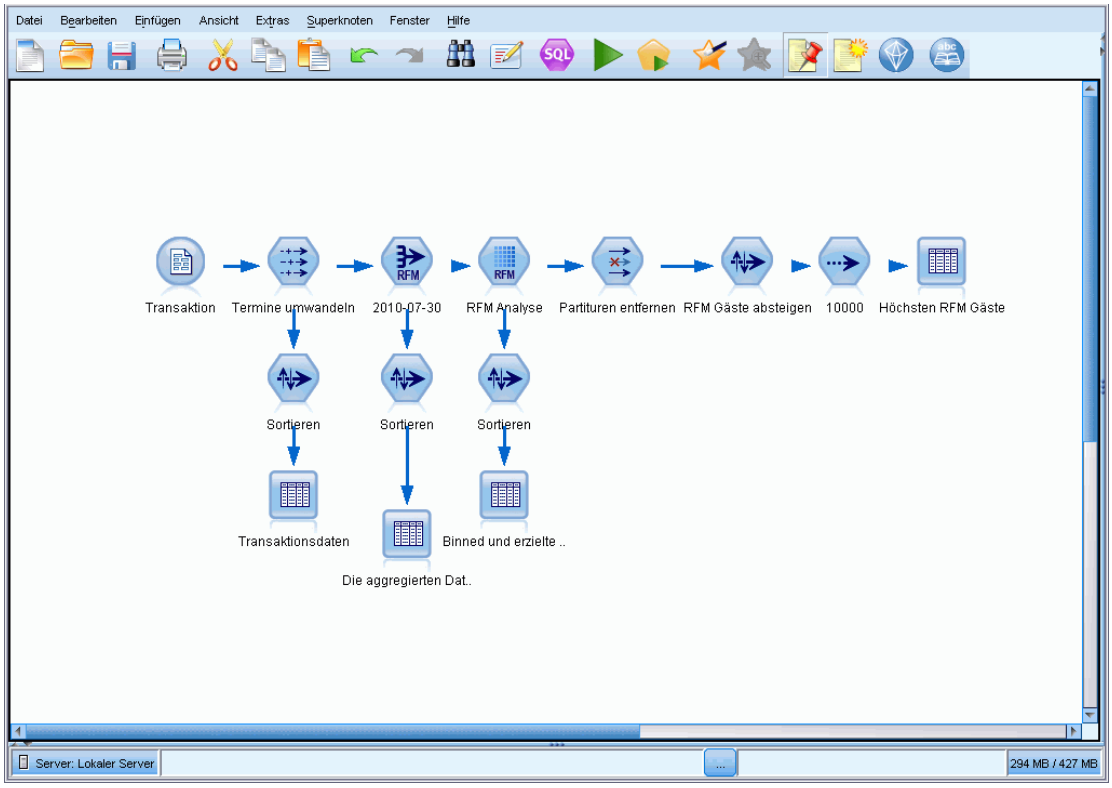
Klicken Sie wie gewünscht auf QuickInfos einblenden oder Große Symbole.

Anpassen des IBM SPSS Modeler-Fensters

Mit den Teilern zwischen den verschiedenen Bereichen der IBM® SPSS® Modeler-Benutzeroberfläche können Sie je nach Bedarf die Größe von Tools ändern oder Tools schließen. Wenn Sie z. B. mit einem großen Stream arbeiten, können Sie die kleinen Pfeile, die sich auf den Teilern befinden, zum Schließen der Knotenpalette, des Manager-Bereichs und des Projektbereichs verwenden. So wird die Größe des Stream-Zeichenbereichs vergrößert und bietet somit genügend Arbeitsfläche für große Streams oder mehrere Streams.

Alternativ können Sie auch im Menü “Ansicht” auf Knotenpalette, Manager oder Projekt klicken, um die Anzeige dieser Elemente ein- oder auszuschalten.

Abbildung 3-12
Maximierter Stream-Zeichenbereich



Als Alternative zum Schließen der Knotenpalette und des Manager- und Projektbereichs können Sie den Stream-Zeichenbereich als scrollbare Seite verwenden, indem Sie die Bildlaufleisten an der Seite und unten im SPSS Modeler-Fenster vertikal und horizontal bewegen.

Sie können auch die Anzeige von Bildschirm-Markup steuern, die Stream-Kommentare, Modellverknüpfungen und die Anzeige von Scoring-Verzweigungen umfasst. Um diese Anzeige ein- oder auszublenden, klicken Sie auf:

Ansicht > Stream-Markup

Verwenden der Maus in IBM SPSS Modeler

Am häufigsten wird die Maus in IBM® SPSS® Modeler wie folgt verwendet:

- **Einfaches Klicken.** Verwenden Sie entweder die rechte oder die linke Maustaste, um Optionen aus den Menüs auszuwählen, Popup-Menüs zu öffnen und verschiedene andere Standardsteuerelemente und Optionen zu verwenden. Klicken Sie und halten Sie die Maustaste gedrückt, um Knoten zu verschieben und zu ziehen.
- **Doppelklicken.** Doppelklicken Sie mit der linken Maustaste, um Knoten auf dem Stream-Zeichenbereich abzulegen und die bereits vorhandenen Knoten zu bearbeiten.
- **Klicken mit der mittleren Maustaste.** Klicken Sie auf die mittlere Maustaste und ziehen Sie den Cursor, um Knoten im Stream-Zeichenbereich miteinander zu verbinden. Doppelklicken Sie auf die mittlere Maustaste, um die Verbindung eines Knotens zu lösen. Wenn Sie nicht über eine Maus mit drei Tasten verfügen, können Sie diese Funktion simulieren, indem Sie auf die ALT-Taste drücken und gleichzeitig auf die Maus klicken und ziehen.

Verwenden von Direktzugriffstasten

Für viele visuelle Programmieroperationen in IBM® SPSS® Modeler gibt es Direktzugriffstasten. Sie können z. B. einen Knoten löschen, indem Sie darauf klicken und die Entf-Taste auf der Tastatur drücken. Auf ähnliche Weise können Sie einen Stream speichern, indem Sie auf die Taste “S” und gleichzeitig die Strg-Taste drücken. Steuerbefehle wie dieser werden durch eine Kombination aus Strg und einer anderen Taste, z. B. Strg+S, angezeigt.

Es gibt eine Reihe von Direktzugriffstasten, die für Windows-Standardoperationen, z. B. Strg+X zum Ausschneiden, verwendet werden. Diese Direktzugriffstasten werden in SPSS Modeler zusammen mit den folgenden anwendungsspezifischen Direktzugriffstasten unterstützt.

Hinweis: In manchen Fällen stehen alte, in SPSS Modeler verwendete Zugriffstasten mit den Windows-Standardzugriffstasten in Konflikt. Diese alten Zugriffstasten werden durch Hinzufügen der ALT-Taste unterstützt. So kann z. B. Strg+Alt+C zum Aktivieren und Deaktivieren des Caches verwendet werden.

Tabelle 3-1
Unterstützte Direktzugriffstasten

Tastenkürze	Funktion
Strg+A	Alles markieren
Strg+X	Ausschneiden
Strg+N	Neuer Stream

Tastenkürzel	Funktion
Strg+O	Stream öffnen
Strg+P	Drucken
Strg+C	Kopieren
Strg+V	Einfügen
Strg+Z	Rückg��ngig
Strg+Q	Alle Knoten unterhalb des ausgewählten Knotens auswählen
Strg+W	Auswahl aller abwärts liegenden Knoten aufheben (wechselt mit Strg+Q)
Strg+E	Vom ausgewählten Knoten ausführen
Strg+S	Aktuellen Stream speichern
Alt+Pfeiltasten	Ausgewählte Knoten im Stream-Zeichenbereich in die Richtung, in die die verwendete Pfeiltaste zeigt, verschieben.
Umschalt+F10	Popup-Menü für den ausgewählten Knoten öffnen

Tabelle 3-2
Unterstützte Direktzugriffstasten für alte Tastenkombinationen

Tastenkürzel	Funktion
Strg+Alt+D	Knoten duplizieren
Strg+Alt+L	Knoten laden
Strg+Alt+R	Knoten umbenennen
Strg+Alt+U	Benutzereingabeknoten erstellen
Strg+Alt+C	Cache aktivieren/deaktivieren
Strg+Alt+F	Cache löschen
Strg+Alt+X	Superknoten erweitern
Strg+Alt+Z	Vergrößern/Verkleinern
Delete	Knoten oder Verbindung löschen

Drucke

Folgende Objekte können in IBM® SPSS® Modeler gedruckt werden:

- Stream-Diagramme
- Grafiken
- Tabellen
- Berichte (aus dem Berichtsknoten und den Projektberichten)
- Skripts (aus den Dialogfeldern “Stream-Eigenschaften”, “Standalone-Skript” oder “Superknoten-Skript”)
- Modelle (Modellbrowser, Registerkarten des Dialogfelds mit aktuellem Fokus, Baumansichten)
- Anmerkungen (unter Verwendung der Registerkarte “Anmerkungen” für die Ausgabe)

So drucken Sie ein Objekt:

- Um ohne Vorschau zu drucken, klicken Sie auf die Schaltfläche “Drucken” in der Symbolleiste.

- Um die Seite vor dem Drucken einzurichten, wählen Sie im Menü “Datei” die Option Seite einrichten.
- Um vor dem Drucken eine Vorschau anzuzeigen, wählen Sie im Menü “Datei” die Option Druckvorschau.
- Um das Standarddialogfeld zum Drucken mit Optionen zur Auswahl von Druckern anzuzeigen und die Optionen für das Aussehen festzulegen, wählen Sie im Menü “Datei” die Option Drucken.

Automatisieren von IBM SPSS Modeler

Da es sich beim erweiterten Data-Mining um einen komplexen und manchmal langwierigen Vorgang handeln kann, bietet IBM® SPSS® Modeler mehrere Kodierungsmöglichkeiten und Automatisierungsunterstützung.

- **Control Language for Expression Manipulation** (CLEM) ist eine Sprache zur Analyse und Bearbeitung der Daten, die SPSS Modeler-Streams durchlaufen. Data-Mining-Experten verwenden CLEM in hohem Maße für Stream-Operationen, um so einfache Aufgaben durchzuführen, wie das Ableiten von Gewinn aus Kosten- und Einkommensdaten, oder um so komplexe Aufgaben durchzuführen, wie das Umwandeln von Webprotokolldaten in ein Set von Feldern und Datensätzen mit nützlichen Informationen. Für weitere Informationen siehe Thema Informationen zu CLEM in Kapitel 7 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.
- Die **Skript** bildet ein leistungsstarkes Tool, mit dem Prozesse in der Benutzeroberfläche automatisiert werden. Mit Skripts können dieselben Aktionen durchgeführt werden, die die Benutzer mithilfe von Maus und Tastatur durchführen. Sie können für Knoten Optionen festlegen und mithilfe einer Untergruppe von CLEM Ableitungen durchführen. Sie können auch die Ausgabe angeben und generierte Modelle bearbeiten. Für weitere Informationen siehe Thema Skripterstellung – Überblick in Kapitel 2 in *IBM SPSS Modeler 14.2 – Handbuch für die Skripterstellung und Automatisierung*.

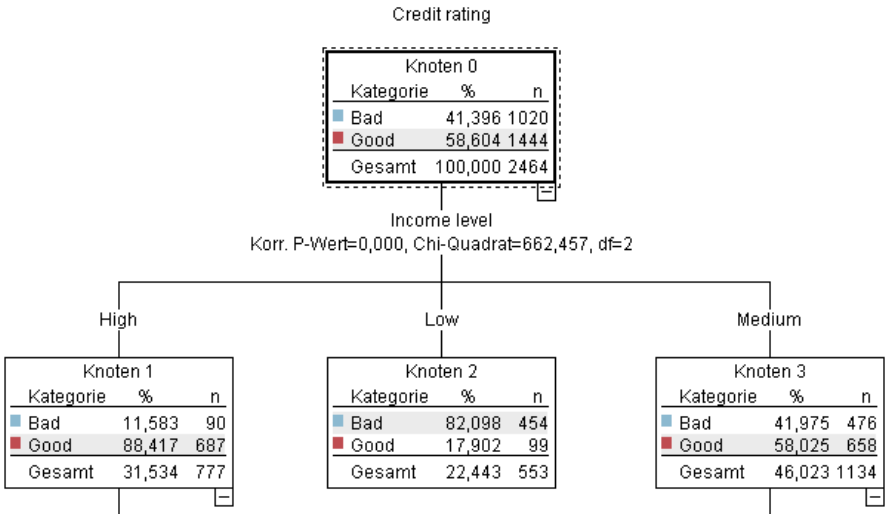
Kapitel
4

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. Eine Finanzinstitut verwendet z. B. möglicherweise ein Modell, um auf Basis von bekannten Informationen über vorherige Kreditantragstreller vorherzusagen, ob ein Kreditantragsteller ein geringes oder hohes Risiko darstellt.

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

Abbildung 4-1
Ein einfaches Entscheidungsbaum-Modell



In diesem Beispiel wird ein **Entscheidungsbaum**-Modell verwendet, das Datensätze aufzeichnet (und eine Reaktion vorhersagt), wobei eine Reihe von Entscheidungsregeln verwendet wird wie beispielsweise:

```
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 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	Einkommenstufe: 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 Entscheidungsbaum-Modells können Sie die Charakteristiken der beiden Kundengruppen analysieren und die Wahrscheinlichkeit von Kreditausfällen vorhersagen.

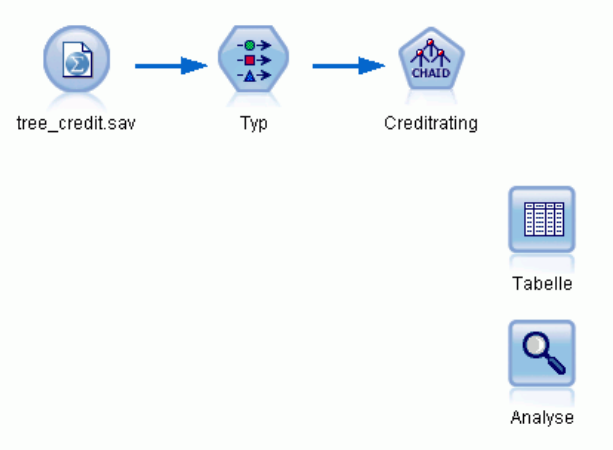
Für dieses Beispiel wird der Stream *modelingintro.str* verwendet, der im Ordner *Demos* unter dem Unterordner *streams* verfügbar ist. Die Datendatei ist *tree_credit.sav*. Für weitere Informationen siehe Thema Ordner “Demos” in Kapitel 1 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Werfen wir nun einen Blick auf den Stream.

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

Erstellen des Streams

Abbildung 4-2
Modellierungs-Stream



Um einen Stream zum Erzeugen eines Modells zu erstellen, benötigen Sie mindestens drei Elemente:

- Ein Quellenknoten, der Daten aus einer externen Quelle einliest, in diesem Fall eine IBM® SPSS® Statistics-Datendatei.
- Ein Quellen- oder Typknoten, der Feldeigenschaften wie beispielsweise 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 Modell-Nugget 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.

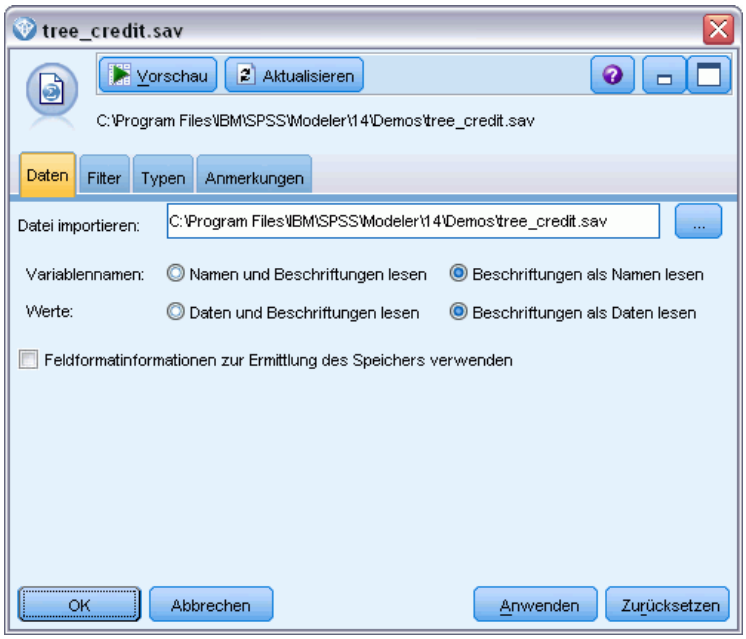
Wenn Messniveaus im Quellenknoten angegeben sind, kann auf den separaten Typknoten verzichtet werden. Hinsichtlich der Funktion ist das Ergebnis dasselbe.

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

Der Statistikdatei-Quellenknoten liest Daten in 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 diese 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 4-3
Einlesen von Daten mit einem Statistikdateiquellenknoten



Der Typknoten gibt das **Messniveau** für die einzelnen Felder an. Das Messniveau ist eine Kategorie, die den Datentyp für das Feld anzeigt. Unsere Quelldatendatei 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 4-4
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.

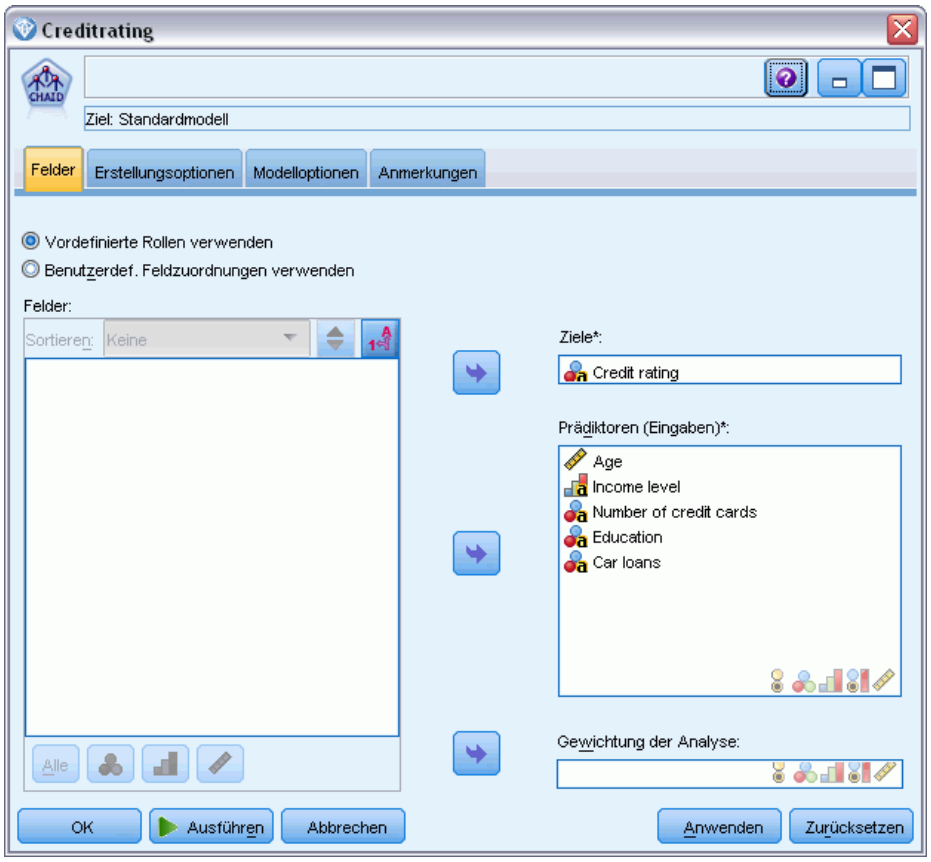
Für die anderen Felder ist die Rolle auf *Eingabe* eingestellt. Eingabefelder werden manchmal auch als **Prädiktoren** bezeichnet oder als Felder, mit deren Werten der Modellierungsalgorithmus den Wert des Zielfelds vorhersagt.

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.

▶ Klicken Sie auf die Registerkarte “Erstellungsoptionen”.

Abbildung 4-5
CHAID-Modellierungsknoten, Registerkarte “Felder”



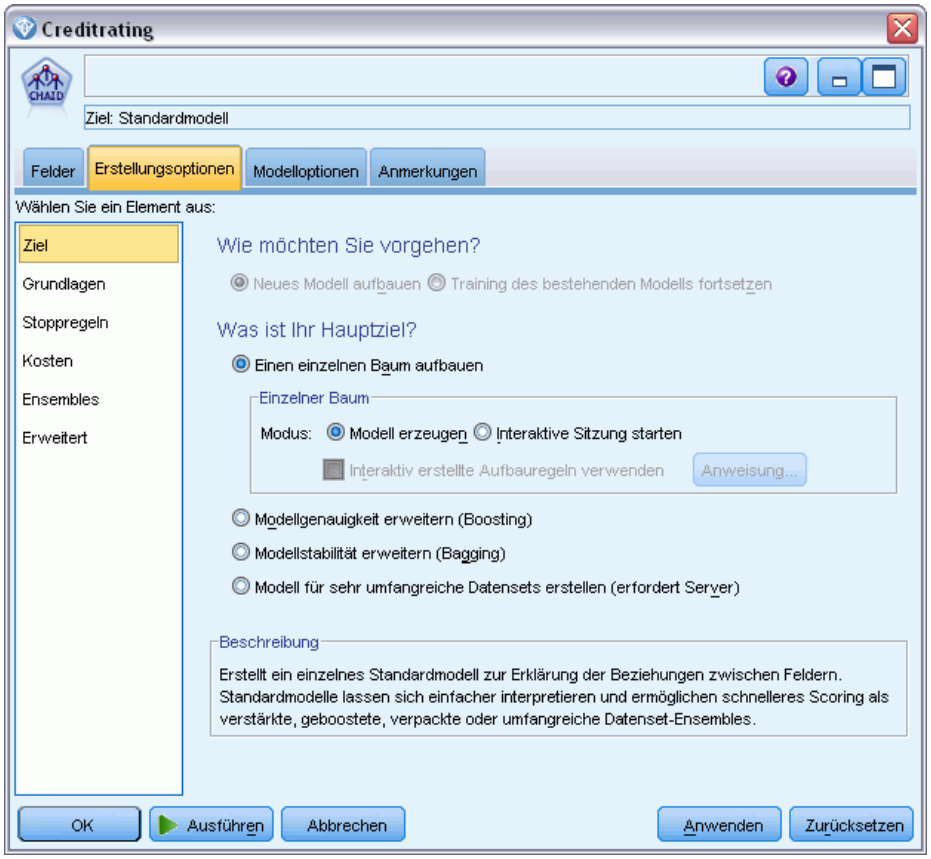
Hier finden Sie einige Optionen, in denen Sie die Art des aufzubauenden Modells festlegen können.

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 4-6
CHAID-Modellierungsknoten, Registerkarte "Erstellungsoptionen"

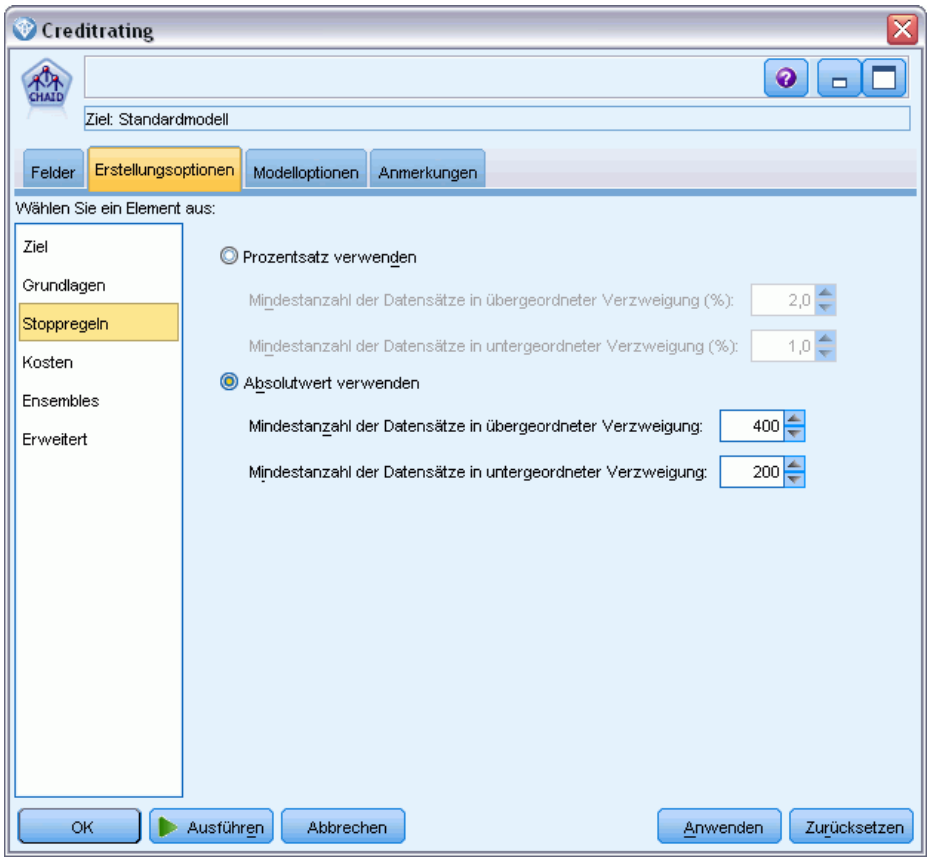


Für dieses Beispiel möchten wir einen einfach strukturierten Baum verwenden und begrenzen deshalb die Baumerweiterung, indem wir die minimale Anzahl der Fälle für über- und untergeordnete Knoten erhöhen.

- Wählen Sie auf der Registerkarte "Erstellungsoptionen" im linken Navigationsbereich Stoppregeln aus.
- Wählen Sie die Option Absolutwert verwenden aus.
- Legen Sie für Mindestanzahl der Datensätze in übergeordneter Verzweigung 400 fest.

- Legen Sie für Mindestanzahl der Datensätze in untergeordneter Verzweigung 200 fest.

Abbildung 4-7
Festlegen der Grenzkriterien beim Erstellen von Entscheidungsbäumen



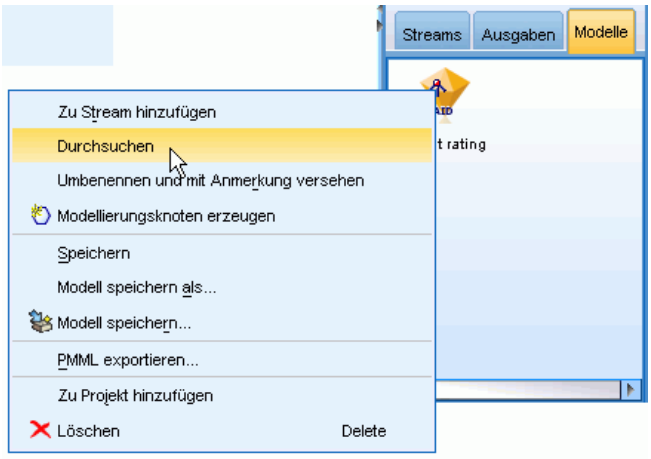
Wir können in diesem Beispiel alle anderen Standardoptionen verwenden, klicken Sie also auf Ausführen, um das Modell zu erstellen. (Alternativ können Sie mit der rechten Maustaste auf den Knoten klicken und im Kontextmenü Ausführen auswählen oder den Knoten auswählen und Ausführen im Menü “Extras” wählen.)

Durchsuchen des Modells

Nach Abschluss der Ausführung wird das Modell-Nugget der Modellpalette rechts oben im Anwendungsfenster hinzugefügt. Zusätzlich wird es in der Stream-Zeichenfläche mit einer Verknüpfung zum Modellierungsknoten gezeigt, von dem aus es erstellt wurde. Um die

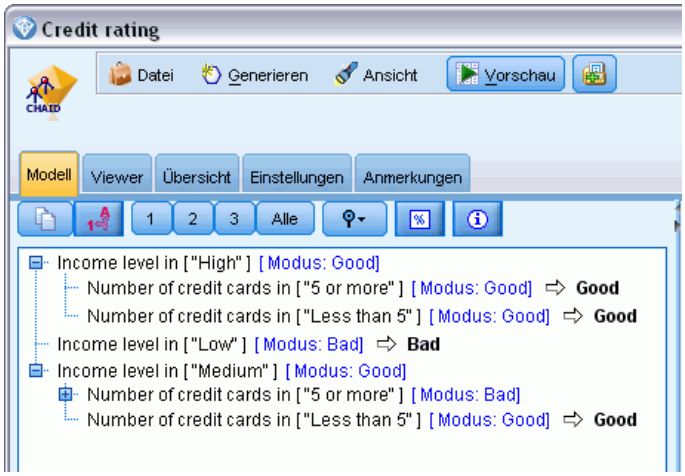
Modelldetails anzuzeigen, klicken Sie mit der rechten Maustaste auf den generierten Modellknoten und wählen Durchsuchen (in der Modellpalette) oder Bearbeiten (in der Zeichenfläche).

Abbildung 4-8
Modellpalette



Im Fall des CHAID-Nuggets zeigt die Registerkarte “Modell” die Details in Form einer Regelmenge. 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.

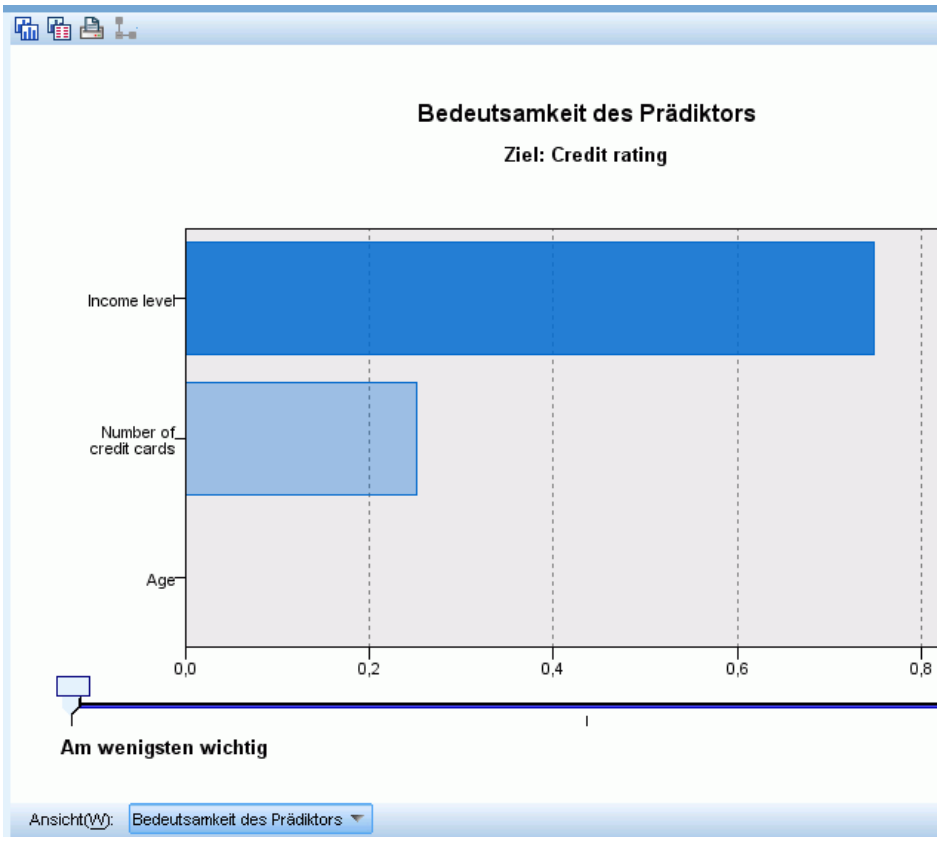
Abbildung 4-9
CHAID-Modell-Nugget, Regelmenge



Für jeden Entscheidungsbaum-Endknoten (also die Baumknoten, die nicht weiter aufgeteilt werden) wird die Vorgersage *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.

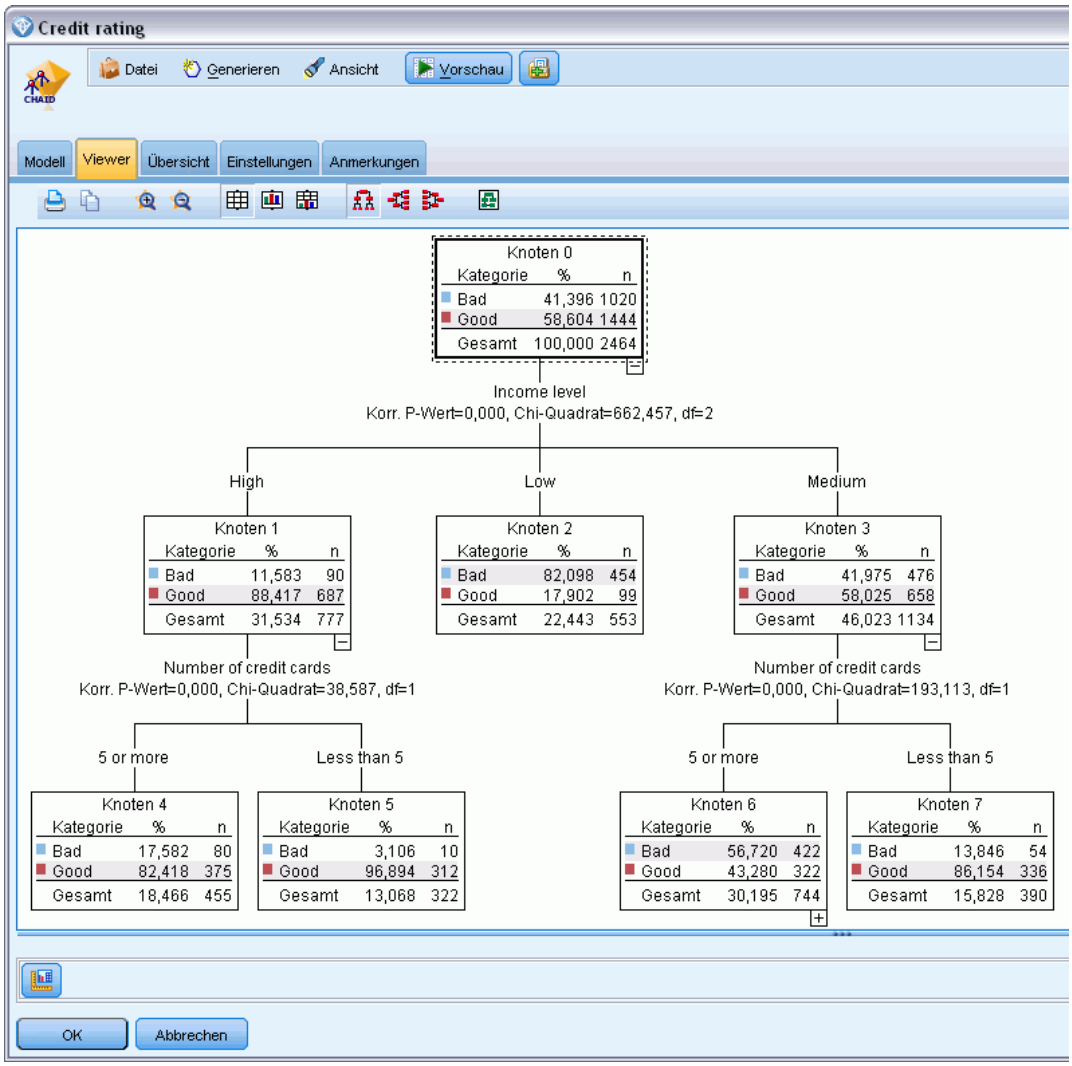
Rechts neben der Regelmenge zeigt die Registerkarte “Modell” das Diagramm “Bedeutsamkeit der Prädiktoren”, das die relative Wichtigkeit jedes Prädiktors beim Schätzen des Modells zeigt. Das zeigt uns, dass die *Einkommen in Kategorien* in diesem Fall eindeutig die größte Bedeutung hat, und dass der einzige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

Abbildung 4-10
Bedeutsamkeit der Prädiktoren – Diagramm



Auf der Registerkarte “Viewer” im Modell-Nugget wird dasselbe Modell in Form eines Baums angezeigt, mit einem Knoten bei jedem Entscheidungspunkt. Mit den Zoom-Steuerelementen auf der Symbolleiste können Sie die Ansicht eines bestimmten Knotens vergrößern bzw. die Ansicht verkleinern, um einen größeren Ausschnitt aus dem Baum zu sehen.

Abbildung 4-11
Registerkarte “Viewer” im Modell-Nugget, “Verkleinern” ausgewählt



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

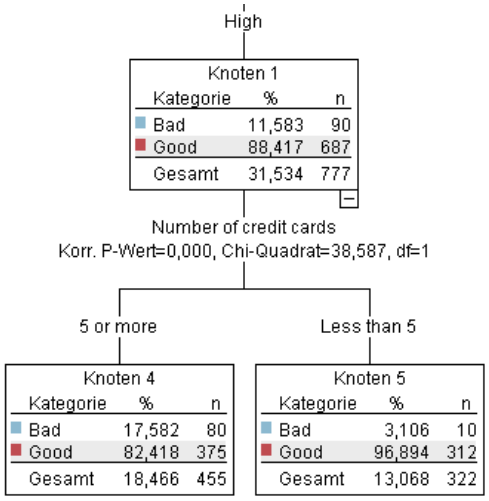
Wie wir sehen, findet die erste Aufteilung bei der *Einkommen in Kategorien* statt Datensätze, bei denen die Einkommensstufe in der Kategorie *Niedrig* liegt, werden Knoten 2 zugewiesen. Entsprechend enthält diese Kategorie den höchsten Prozentsatz an Kreditausfällen. Die Kreditvergabe an Kunden in dieser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

Bei 16 % der Kunden in dieser Kategorie kam es allerdings *nicht* zum Kreditausfall, die Prognose ist 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 zufriedenstellenderes Ergebnis erzielen.

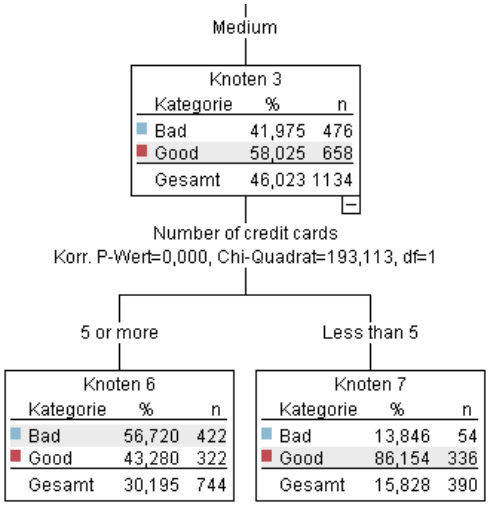
Abbildung 4-12
Baumansicht der Kunden mit hohem Einkommen



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 zufriedenstellenderes Ergebnis erzielen.

Abbildung 4-13
Baumansicht der Kunden mit mittlerem Einkommen



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

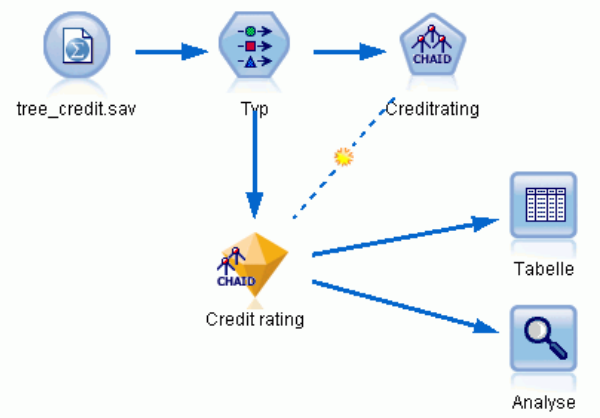
Dieser Vorgang der Zuweisung von Vorhersagen zu einzelnen Datensätzen wird als **Scoring** bezeichnet. Indem wir die Datensätze scoren, die auch zur Schätzung des Modells verwendet wurden, können wir evaluieren, mit welcher Genauigkeit das Modell für die Trainingsdaten (die Daten, für die das Ergebnis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

Bewertung 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 4-14
Anhängen des Modell-Nuggets an Ausgabeknoten zur Modellevaluation



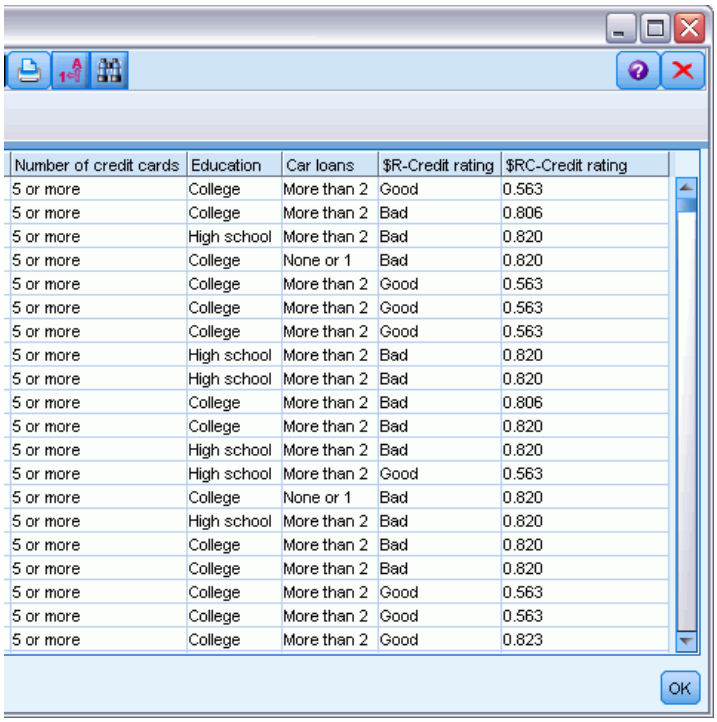
- Fügen Sie zur Anzeige der Scores bzw. Vorhersagen den Tabellenknoten zum Modell-Nugget hinzu, doppelklicken Sie auf den Tabellenknoten und klicken Sie auf Ausführen.

  Die Tabelle zeigt die vorhergesagten Scores unter einem Feldnamen (*$R-Credit rating*) an, 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, tragen jedoch ein Standard-Präfix, wie beispielsweise *$R-* für Vorhersagen oder *$RC-* für Konfidenzwerte. Verschiedene Modelltypen verwenden verschiedene Präfix-Sets. Ein

Konfidenzwert ist die Schätzung des Modells (auf einer Skala von 0,0 bis 1,0) bezüglich der Genauigkeit der einzelnen vorhergesagten Werte.

Abbildung 4-15
Tabelle mit generierten Scores und Konfidenzwerten



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

OK

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.)

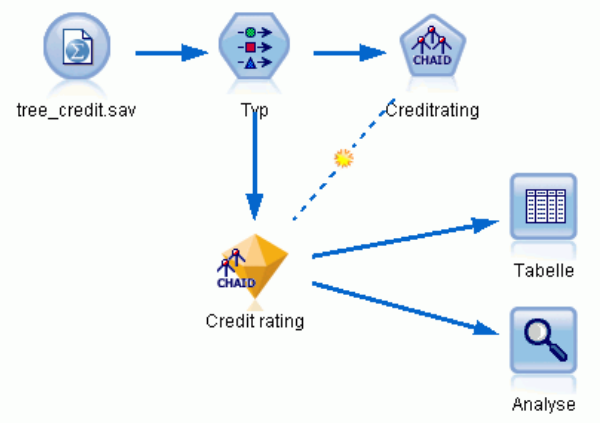
Um dies zu vermeiden, könnten wir damit fortfahren, den Baum in immer kleinere Verzweigungen aufzuspalten, bis jeder Knoten 100%ig einheitlich wäre – nur *Gut* oder nur *Schlecht*, ohne gemischte Antworten. Ein derartiges Modell wäre jedoch extrem kompliziert und ließe sich vermutlich nicht gut auf andere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

- Verbinden Sie das Modell-Nugget mit dem Analyseknoten.

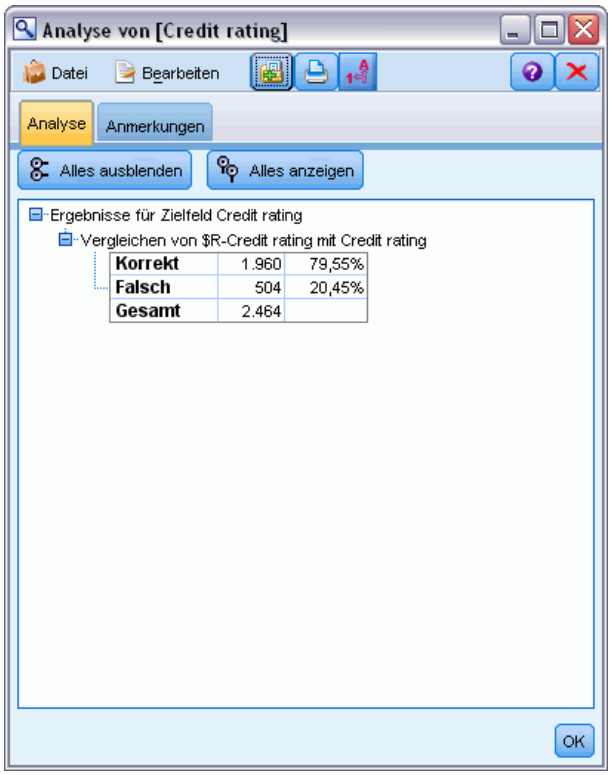
▶ Doppelklicken Sie auf den Analyseknoten und klicken Sie auf Ausführen.

Abbildung 4-16
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 4-17
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 Situationen könnten Sie einen Partitionsknoten verwenden, um die Daten in separate Stichproben für Training und Evaluierung aufzuteilen.

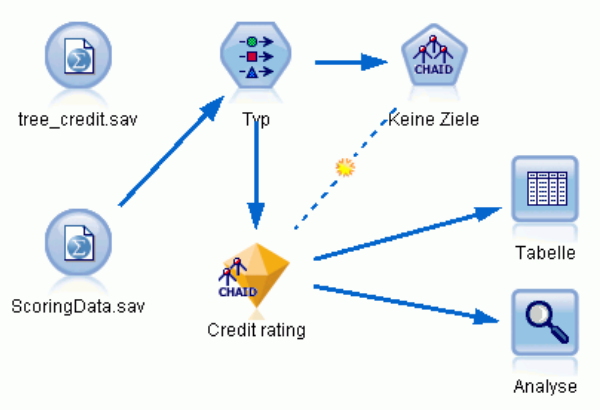
Durch Verwendung einer Stichprobenpartition zur Generierung des Modells und einer weiteren Stichprobenpartition zum Testen des Modells können Sie einen wesentlich besseren Anhaltspunkt dafür erhalten, wie gut sich das Modell auf andere Daten-Sets generalisieren lässt.

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 4-18
Angliedern neuer Daten zum Scoring



Sie können den Statistikdatei-Quellenknoten 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 Daten-Set 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 Modell-Nugget zu 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 den im Modell verwendeten übereinstimmen.

Sie können das Modell-Nugget auch als separate Datei speichern, das Modell im PMML-Format exportieren, um sie in anderen Anwendungen zu nutzen, die dieses Format unterstützen, oder das Modell in IBM® SPSS® Collaboration and Deployment Services speichern, was unternehmensweites Deployment, Scoring und unternehmensweite Verwaltung der Modelle 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 Modell-Nugget. Dieser Vorgang wird auch als Trainieren des Modells bezeichnet.
- Das Modell-Nugget kann zu 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 5

Automatische Modellierung für ein Flag-Ziel

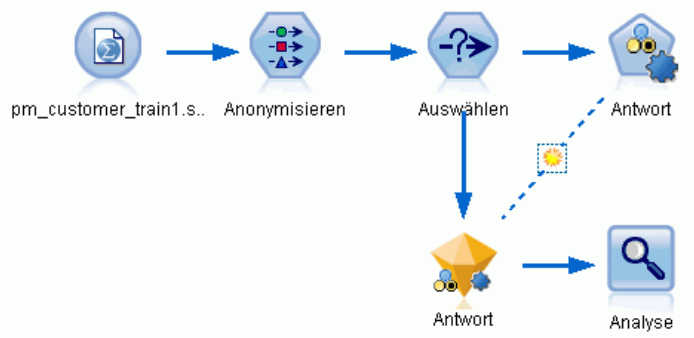
Modellieren der Kundenreaktion (Automatischer Klassifizierer)

Mit dem Knoten “Automatischer Klassifizierer” können Sie automatisch mehrere verschiedene Modelle für Flag-Ziele (beispielsweise ob ein bestimmter Kunde mit hoher Wahrscheinlichkeit einen Kredit nicht zurückzahlt oder auf ein bestimmtes Angebot eingeht) oder nominale (Set-) Ziele erstellen und vergleichen. In diesem Beispiel wird nach einem Flag-Ergebnis (“Ja” oder “Nein”) gesucht. In einem relativ einfachen Stream generiert der Knoten eine Gruppe infrage kommender Modelle und weist ihnen Ränge zu, wählt diejenigen aus, die die beste Leistung erbringen, und fasst sie zu einem einzigen aggregierten Modell (Ensemble-Modell) zusammen. Dieser Ansatz bietet gleichzeitig Automatisierung und die Vorteile der Kombination mehrerer Modelle, die häufiger genauere Vorhersagen erlaubt, als aus den einzelnen Modellen erzielt werden können.

Dieses Beispiel beruht auf einem fiktiven Unternehmen, das profitablere Ergebnisse erzielen möchte, indem jedem Kunden ein speziell für ihn geeignetes Angebot unterbreitet wird.

Bei diesem Ansatz werden die Vorteile der Automatisierung betont. Ein ähnliches Beispiel mit einem stetigen Ziel (numerischer Bereich) finden Sie hier: Kapitel 6 auf S. 57.

Abbildung 5-1
Automatischer Klassifizierer – Beispiel-Stream



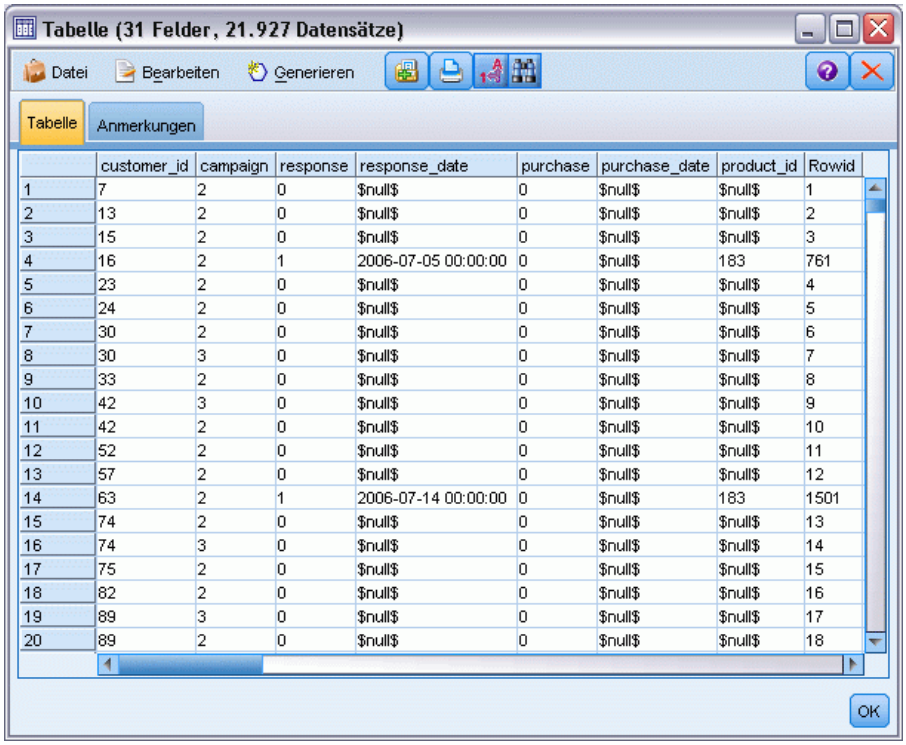
In diesem Beispiel wird der Stream *pm_binaryclassifier.str* verwendet, der im Ordner “Demo” unter *streams* installiert ist. Die verwendete Datendatei ist *pm_customer_train1.sav*. Für weitere Informationen siehe Thema Ordner “Demos” in Kapitel 1 auf S. 4.

Historische Daten

Die Datei *pm_customer_train1.sav* enthält historische Daten, die die Aufzeichnungen über die Angebote enthält, die bestimmten Kunden in früheren Kampagnen unterbreitet wurden, entsprechend dem Wert im Feld *campaign* (Kampagne). Die größte Anzahl an Datensätzen entfallen auf die Kampagne *Premium account* (Premium-Account).

Die Werte des Felds *campaign* (Kampagne) sind in den Daten tatsächlich als ganze Zahlen kodiert (Beispiel: *2 = Premium account*). Später werden Sie Beschriftungen für diese Werte definieren, um eine verständlichere Ausgabe zu erzielen.

Abbildung 5-2
Daten zu früheren Werbeaktionen



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

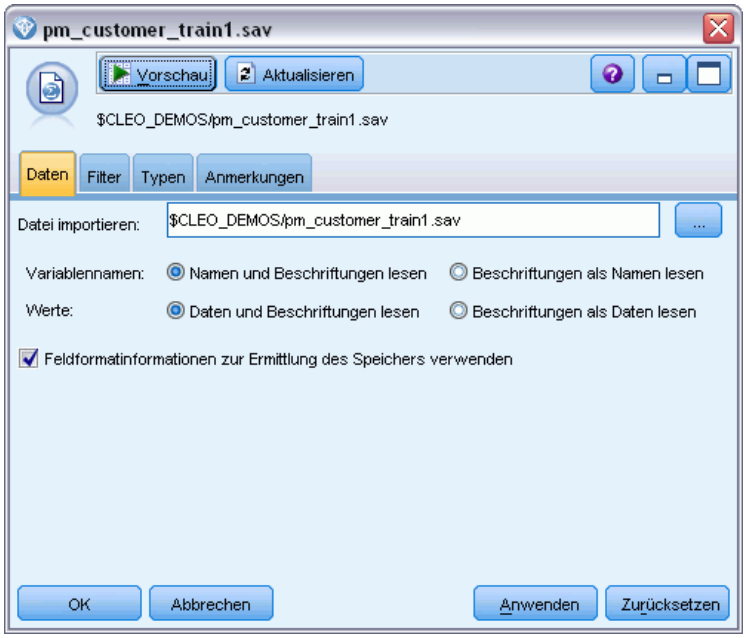
Außerdem enthält die Datei das Feld *response* (Antwort), das angibt, ob das Angebot angenommen wurde (0 = *nein*, 1 = *ja*). Dies ist das **Zielfeld**, also der vorherzusagende Wert. Außerdem ist eine Reihe von Feldern mit demografischen Informationen und Finanzdaten zu den einzelnen Kunden enthalten. Diese Felder können zum Erstellen bzw. "Trainieren" eines Modells verwendet werden, das die Antwortquoten für Einzelpersonen oder Gruppen auf der Grundlage bestimmter Merkmale, wie Einkommen, Alter oder Anzahl der Transaktionen pro Monat, vorhersagt.

Erstellen des Streams

- Fügen Sie einen Statistics-Dateiquellenknoten hinzu, der auf die Datei *pm_customer_train1.sav* im Ordner *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist. (Sie können `$CLEO_DEMOS/` im Dateipfad als Verknüpfung zur Referenzierung dieses Ordners angeben. Beachten Sie, dass

im Pfad, wie gezeigt, ein normaler Schrägstrich (und nicht etwa ein umgekehrter Schrägstrich) verwendet werden muss.)

Abbildung 5-3
Einlesen der Daten



- Fügen Sie einen Typknoten hinzu und wählen Sie *response* (Antwort) als Zielfeld (Rolle = Ziel) aus. Setzen Sie das Messniveau für dieses Feld auf Flag.

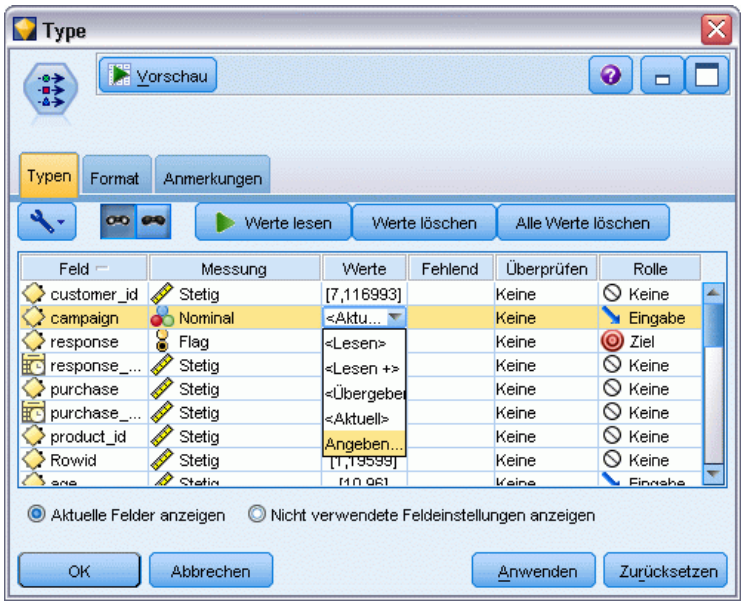
Abbildung 5-4
Festlegen von Messniveau und Rolle

- Setzen Sie die Rolle für folgende Felder auf Keine. *customer_id* (Kunden-ID), *campaign* (Kampagne), *response_date* (Antwortdatum), *purchase* (Einkauf), *purchase_date* (Einkaufsdatum), *product_id* (Produkt-ID), *Rowid* (Zeilen-ID), and *X_random* (X-Zufall). Diese Felder werden beim Erstellen des Modells ignoriert.

- Klicken Sie auf die Schaltfläche Werte lesen im Typknoten, um sicherzustellen, dass die Werte instanziiert werden.

  Wie bereits gesehen, umfassen unsere Quellendaten Informationen über vier verschiedene Kampagnen, von denen sich jede an eine andere Art von Kundenkonto richtet. Diese Kampagnen sind in den Daten als Ganzzahlen kodiert. Wir definieren nun Beschriftungen für jede Kampagne, um deutlicher zu sehen, welcher Kontotyp jeder Ganzzahl entspricht.

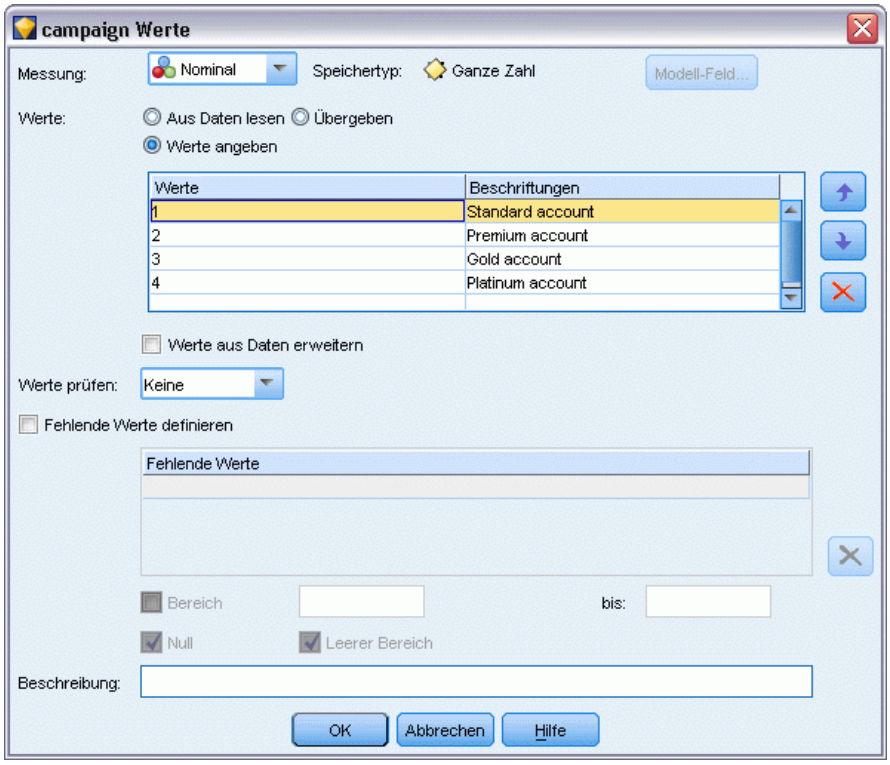
Abbildung 5-5
Auswahl zum Festlegen von Werten für ein Feld



- Klicken Sie in der Zeile für das Feld campaign (Kampagne) auf den Eintrag in der Spalte values (Werte).

- Wählen Sie Angeben aus der Dropdown-Liste.

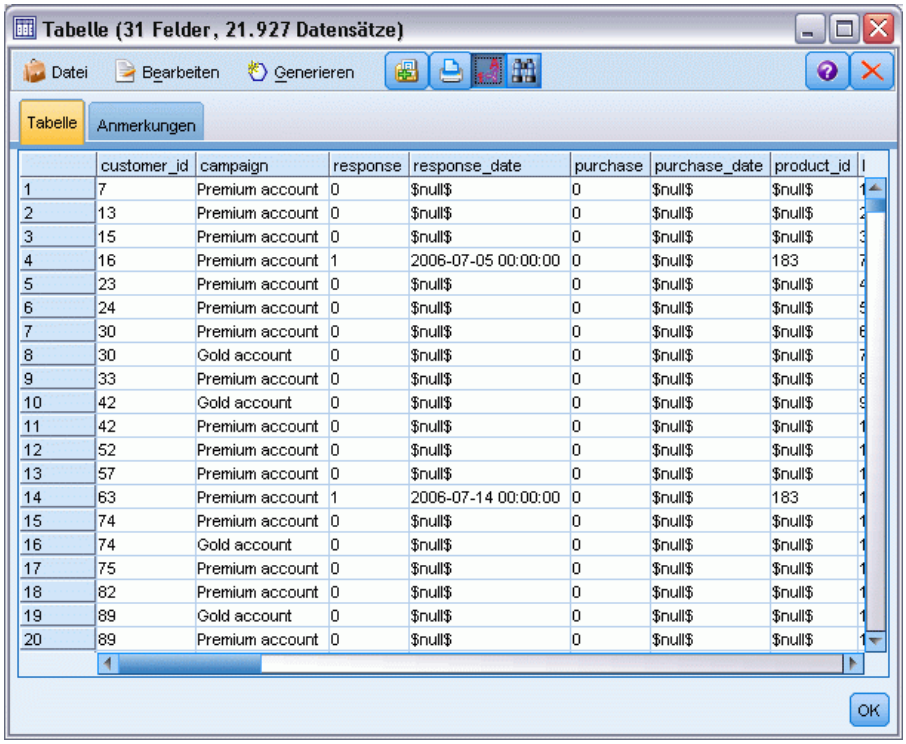
Abbildung 5-6
Definieren von Beschriftungen für die Feldwerte



- Geben Sie in der Spalte Labels die Beschriftungen für die vier Werte des Felds campaign (Kampagne) wie gezeigt ein.
- Klicken Sie auf OK.

Nun können Sie in Ausgabefenstern die Beschriftungen anstelle der Zahlen anzeigen.

Abbildung 5-7
Anzeigen der Feldwertbeschriftungen



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold account	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold account	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold account	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold account	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

- Verbinden Sie einen Tabellenknoten mit dem Typknoten.
- Öffnen Sie den Tabellenknoten und klicken Sie auf Ausführen.
- Klicken Sie im Ausgabefenster auf die Symbolleistenschaltfläche Feld- und Wertelabels anzeigen, um die Beschriftungen anzuzeigen.
- Klicken Sie auf OK, um das Ausgabefenster zu schließen.

Die Daten enthalten Informationen zu vier verschiedenen Kampagnen, Sie konzentrieren die Analyse jedoch jeweils nur auf eine Kampagne. Da die größte Anzahl an Datensätzen auf die Premium-Konten-Kampagne entfällt (in den Daten kodiert als *campaign=2*), können Sie einen Auswahlknoten verwenden, um nur die betreffenden Datensätze in den Stream aufzunehmen.

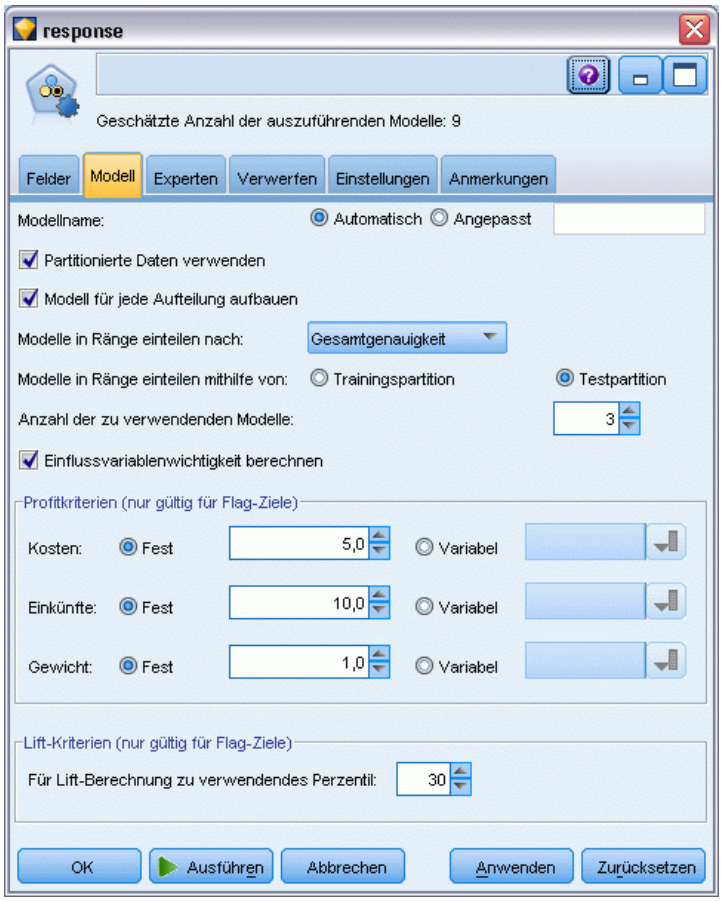
Abbildung 5-8
Auswählen von Datensätzen für eine einzelne Kampagne

Generieren und Vergleichen von Modellen

- Gliedern Sie einen Knoten vom Typ “Automatischer Klassifizierer” an und wählen Sie Gesamtgenauigkeit als Metrik für die Rangordnung der Modelle aus.

- Stellen Sie für Anzahl der zu verwendenden Modelle 3 ein. Das bedeutet, dass bei der Ausführung des Knotens die drei besten Modelle erstellt werden.

Abbildung 5-9
Knoten "Automatischer Klassifizierer" – Registerkarte "Modell"

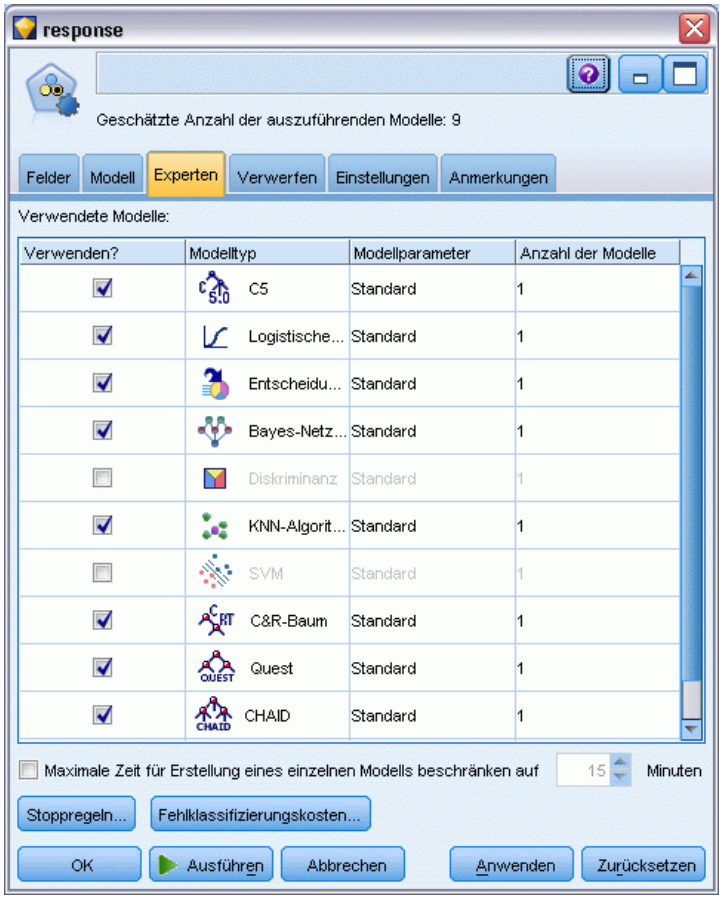


Auf der Registerkarte "Experten" können Sie aus bis zu 11 verschiedenen Modellalgorithmen auswählen.

- Deaktivieren Sie die Modelltypen Diskriminanz und SVM. (Bei diesen Modellen dauert das Training für die vorliegenden Daten länger. Durch den Verzicht darauf wird also die Durchführung des Beispiels beschleunigt.) Wenn es Ihnen nichts ausmacht zu warten, können Sie sie auch ausgewählt lassen.)

Da Sie auf der Registerkarte “Modell” für Anzahl der zu verwendenden Modelle 3 eingestellt haben, berechnet der Knoten die Genauigkeit der restlichen neun Algorithmen und erstellt ein einzelnes Modell-Nugget, in dem die drei genauesten enthalten sind.

Abbildung 5-10
Knoten “Automatischer Klassifizierer” – Registerkarte “Experten”

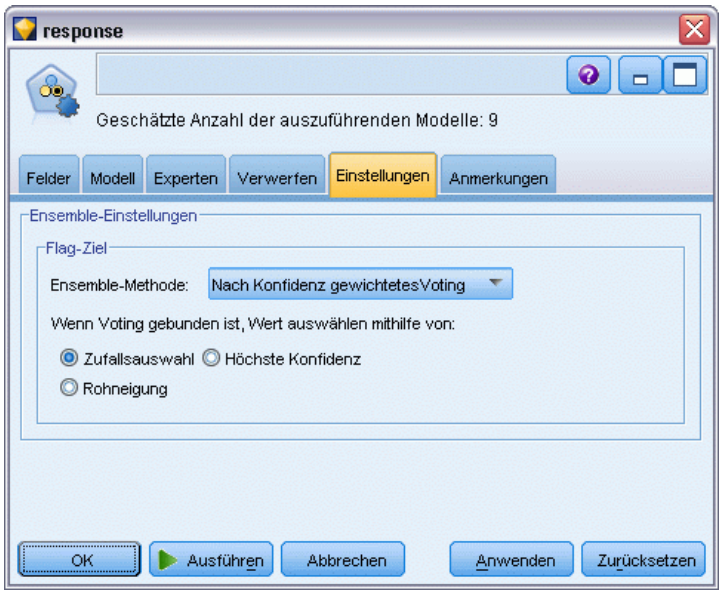


- Wählen Sie auf der Registerkarte “Einstellungen” als Ensemble-Methode Nach Konfidenz gewichtetes Voting aus. Dadurch wird bestimmt, auf welche Weise für jeden Datensatz ein einzelner aggregierter Score erstellt wird.

  Bei einfachem Voting gilt: Wenn zwei von drei Modellen *Ja* vorhersagen, dann “gewinnt” *Ja* mit einem Votum von 2 zu 1 “Stimmen”. Beim nach Konfidenz gewichteten Voting werden die Stimmen anhand des Konfidenzwerts für die einzelnen Vorhersagen gewichtet. Wenn

also ein Modell *Nein* mit einer höheren Konfidenz vorhersagt als die beiden *Ja*-Vorhersagen zusammengenommen, gewinnt *Nein*.

Abbildung 5-11
Knoten "Automatischer Klassifizierer" Registerkarte "Einstellungen"



▶ Klicken Sie auf Ausführen.

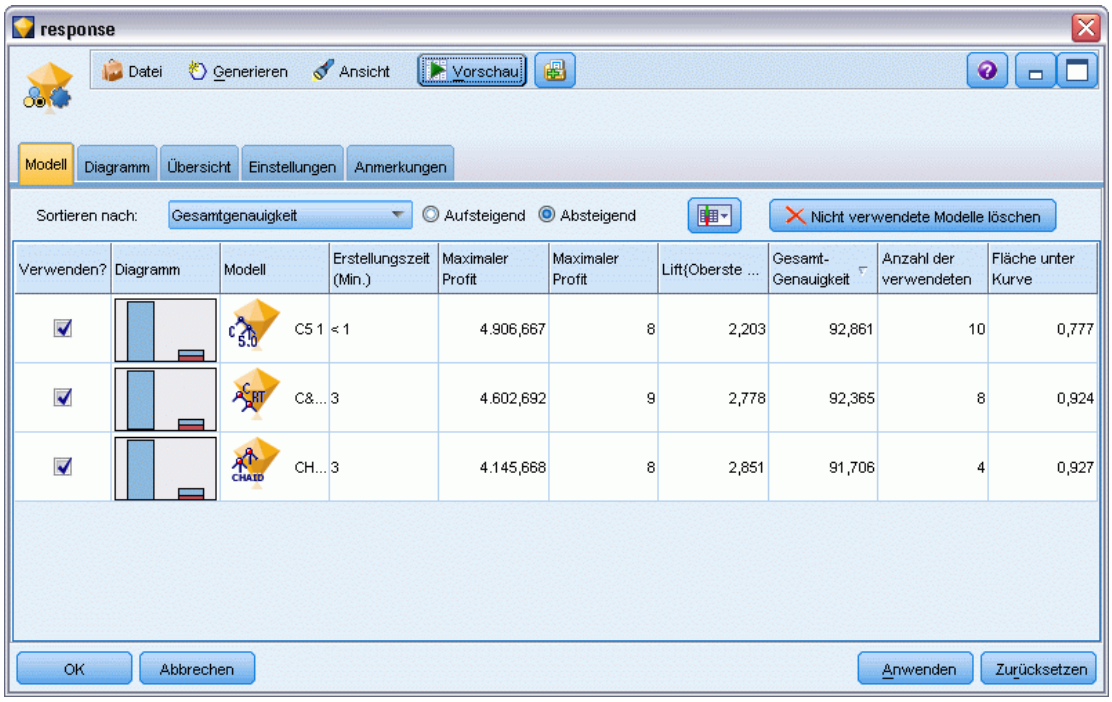
Nach einigen Minuten wird das generierte Modell-Nugget erstellt und auf die Zeichenfläche sowie in die Modellpalette in der rechten oberen Fensterecke platziert. Sie können das Modell-Nugget durchsuchen oder auf verschiedene Weise speichern bzw. bereitstellen.

Öffnen Sie das Modell-Nugget. Es zeigt Details zu den einzelnen Modellen an, die während der Ausführung erstellt wurden. (In einer realen Situation, in der unter Umständen Hunderte von Modellen für ein großes Daten-Set erstellt werden, kann dieser Vorgang etliche Stunden in Anspruch nehmen.) Unter Abbildung 5-1 auf S. 44.

Wenn Sie eines der Modelle eingehender untersuchen möchten, können Sie in der Spalte Modell auf ein Modell-Nugget-Symbol doppelklicken, um einen Drill-Down durchzuführen und die einzelnen Modellergebnisse zu durchsuchen. Anschließend können Sie Modellierungsknoten,

Modell-Nuggets oder Evaluationsdiagramme generieren. In der Spalte Diagramm können Sie ein Diagramm in voller Größe generieren, indem Sie auf eine Miniaturansicht doppelklicken.

Abbildung 5-12
Automatischer Klassifizierer - Ergebnisse



Standardmäßig werden die Modelle auf der Grundlage der Gesamtgenauigkeit sortiert, da dies das Maß ist, das Sie auf der Registerkarte “Modell” des Knotens “Automatischer Klassifizierer” ausgewählt haben. Unter Verwendung dieses Maßes erhält das Modell “C51” den besten Rang, die Modelle “C&RT-Baum” und “CHAID” sind jedoch fast ebenso genau.

Sie können die Sortierung anhand einer anderen Spalte durchführen, indem Sie auf die Kopfzeile der betreffenden Spalte klicken. Außerdem können Sie das gewünschte Maß in der Dropdown-Liste Sortieren nach in der Symbolleiste auswählen.

Basierend auf diesen Ergebnissen entscheiden Sie sich, jedes der drei exaktesten Modelle zu verwenden. Durch die Kombination der Vorhersagen aus mehreren Modellen können Beschränkungen in einzelnen Modellen vermieden werden, was zu einer höheren Gesamtgenauigkeit führt.

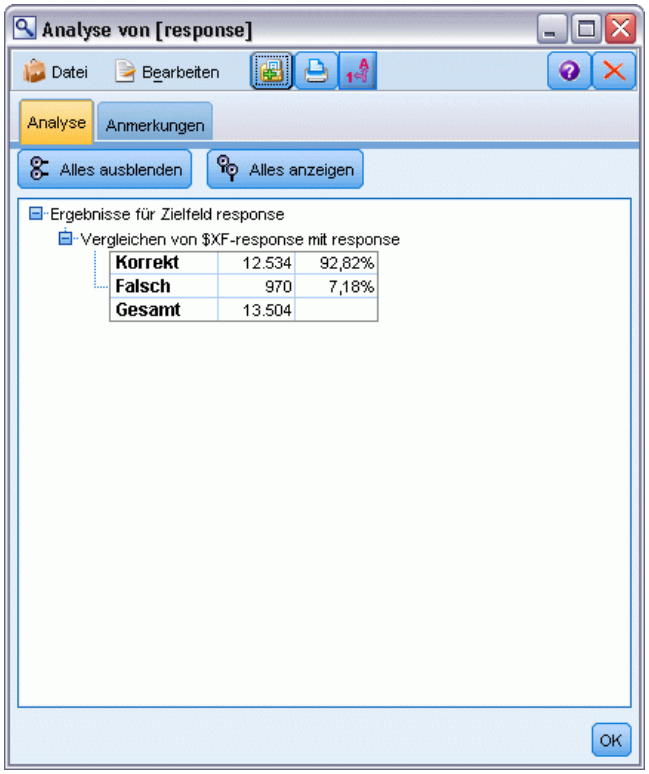
Wählen Sie in der Spalte Verwenden? die Modelle “C51”, “C&RT-Baum” und “CHAID” aus.

Gliedern Sie einen Analyseknoten (Ausgabepalette) nach dem Modell-Nugget an. Klicken Sie mit der rechten Maustaste auf den Analyseknoten und wählen Sie Ausführen aus, um den Stream auszuführen.

Die vom Ensemble-Modell generierten aggregierten Scores werden in einem Feld mit dem Namen *$XF-response* hinzugefügt. Beim Vergleich mit den Trainingsdaten stimmt der vorhergesagte Wert mit einer Gesamtgenauigkeit von 92,82% mit der tatsächlichen Antwort (die im ursprünglichen Feld *Antwort* aufgezeichnet ist) überein.

Das Ensemble-Modell ist in diesem Fall zwar nicht ganz so genau wie das beste der drei Einzelmodelle (92,86 % für das Modell "C51"), der Unterschied ist jedoch zu gering, um von Bedeutung zu sein. Im Allgemeinen bringt ein Ensemble-Modell mit höherer Wahrscheinlichkeit gute Leistungen, wenn es auf andere Daten-Sets als die Trainingsdaten angewendet wird.

Abbildung 5-13
Analyse der drei Modelle



Zusammenfassung

Sie haben mithilfe des Knotens "Automatischer Klassifizierer" eine Reihe verschiedener Modelle verglichen, die drei genausten Modelle verwendet und dem Stream hinzugefügt und diese Modelle schließlich in einem Modell-Nugget "Automatischer Klassifizierer" zusammengefasst.

- Hinsichtlich der Gesamtgenauigkeit erbrachten die Modelle "C51", "C&R-Baum" und "CHAID" die besten Leistungen bei den Trainingsdaten.
- Das Ensemble-Modell erzielte annähernd dieselbe Leistung wie das beste Einzelmodell und erbringt möglicherweise bei Anwendung auf andere Datensätze bessere Leistungen. Wenn Sie den Prozess so weit wie möglich automatisieren möchten, können Sie mit diesem Ansatz

unter den meisten Bedingungen ein robustes Modell erstellen, ohne sich allzu genau mit den spezifischen Eigenschaften der einzelnen Modelle befassen zu müssen.

Kapitel 6

Automatische Modellierung für ein stetiges Ziel

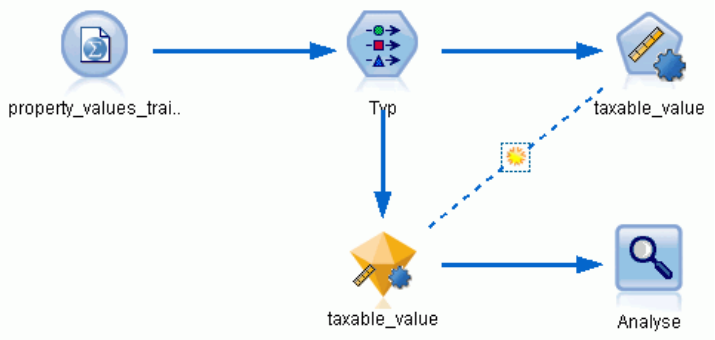
Eigenschaftswerte (Auto-Numerisch)

Mit dem Knoten “Auto-Numerisch” können Sie automatisch verschiedene Modelle für stetige Ergebnisse (numerischer Bereich) erstellen und vergleichen, beispielsweise wenn Sie den steuerlichen Wert einer Immobilie vorhersagen. Mit einem einzelnen Knoten können Sie eine Gruppe von infrage kommenden Modellen schätzen und vergleichen und eine Teilmenge der Modelle für die weitere Analyse erstellen. Der Knoten funktioniert ebenso wie der Knoten “Automatischer Klassifizierer”, ist jedoch für stetige und nicht für Flag- oder nominale Ziele gedacht.

Der Knoten kombiniert die besten der infrage kommenden Modelle in einem einzigen aggregierten Modell-Nugget. Dieser Ansatz bietet gleichzeitig Automatisierung und die Vorteile der Kombination mehrerer Modelle, die häufiger genauere Vorhersagen erlaubt, als aus den einzelnen Modellen erzielt werden können.

Das vorliegende Beispiel konzentriert sich auf eine fiktive Gemeinde, die Steuern auf Immobilien anpassen und einschätzen muss. Um hierbei eine größere Genauigkeit zu erzielen, erstellt die Gemeinde ein Modell, das Immobilienwerte auf der Grundlage von Gebäudetyp, Lage, Größe und anderer bekannter Faktoren vorhersagt.

Abbildung 6-1
Auto-Numerisch – Beispiel-Stream



In diesem Beispiel wird der Stream *property_values_numericpredictor.str* verwendet, der im Ordner “Demos” unter *streams* installiert ist. Die verwendete Datendatei ist *property_values_train.sav*.Für weitere Informationen siehe Thema Ordner “Demos” in Kapitel 1 auf S. 4.

Trainingsdaten

Die Datendatei enthält ein Feld mit der Bezeichnung *taxable_value* (steuerlicher Wert), das das **Zielfeld** bzw. den vorherzusagenden Wert darstellt. Die anderen Felder Enthalten Informationen wie Lage, Gebäudetyp und Innenvolumen und können als Prädiktoren verwendet werden.

Feldname	Label
property_id	Property ID (Eigentums-ID)
neighborhood	Area within the city (Wohngegend innerhalb des Ortes)
building_type	Type of building (Gebäudetyp)
year_built	Year built (Baujahr)
volume_interior	Volume of interior (Innenvolumen)
volume_other	Volume of garage and extra buildings (Volumen von Garage und Nebengebäude)
lot_size	Lot size (Grundstücksgröße)
taxable_value	Taxable value (Steuerlicher Wert)

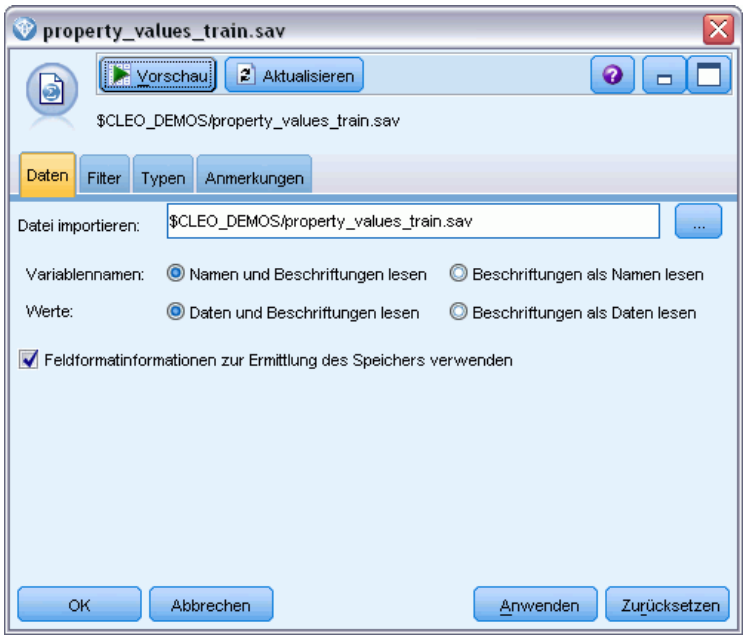
Eine Scoring-Datendatei mit dem Namen *property_values_score.sav* befindet sich ebenfalls im Ordner “Demos”. Sie enthält dieselben Felder, mit Ausnahme des Felds *taxable_value*. Nach dem Trainieren der Modelle mithilfe eines Daten-Sets bei dem der steuerliche Wert bekannt ist, können Sie Datensätze scoren, bei denen dieser Wert noch nicht bekannt ist.

Erstellen des Streams

- Fügen Sie einen Statistics-Dateiquellenknoten hinzu, der auf die Datei *property_values_train.sav* im Ordner *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist. (Sie können `$CLEO_DEMOS/` im Dateipfad als Verknüpfung zur Referenzierung dieses Ordners angeben. Beachten Sie, dass

im Pfad, wie gezeigt, ein normaler Schrägstrich (und nicht etwa ein umgekehrter Schrägstrich) verwendet werden muss.)

Abbildung 6-2
Einlesen der Daten

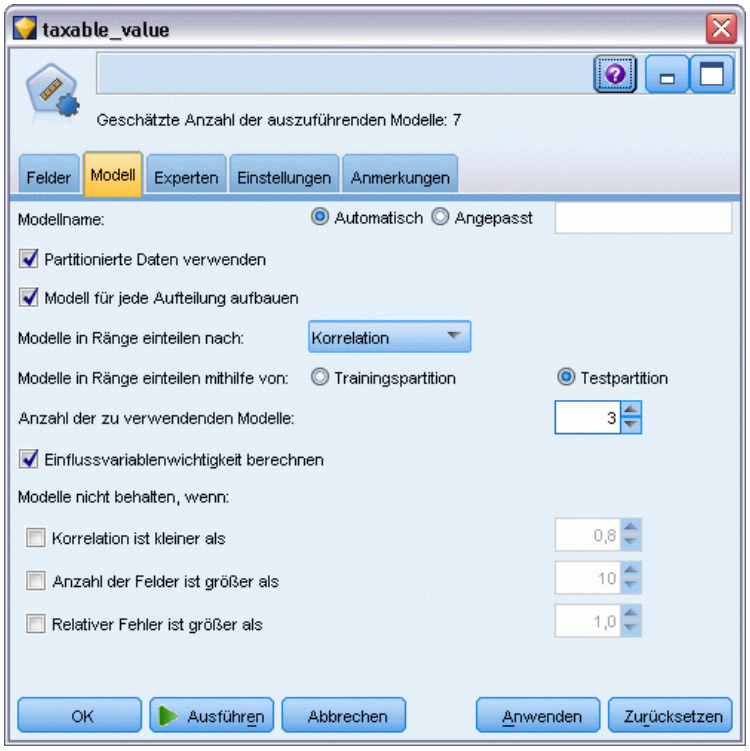


- Fügen Sie einen Typknoten hinzu und wählen Sie *taxable_value* (Antwort) als Zielfeld (Rolle = Ziel) aus. Für die anderen Felder sollte als Rolle Eingabe festgelegt werden, sodass sie als Prädiktoren verwendet werden.

Abbildung 6-3
Festlegen des Zielfelds

- Gliedern Sie einen Knoten vom Typ “Auto-Numerisch” an und wählen Sie Korrelation als Metrik für die Rangordnung der Modelle aus.
- Stellen Sie für Anzahl der zu verwendenden Modelle 3 ein. Das bedeutet, dass bei der Ausführung des Knotens die drei besten Modelle erstellt werden.

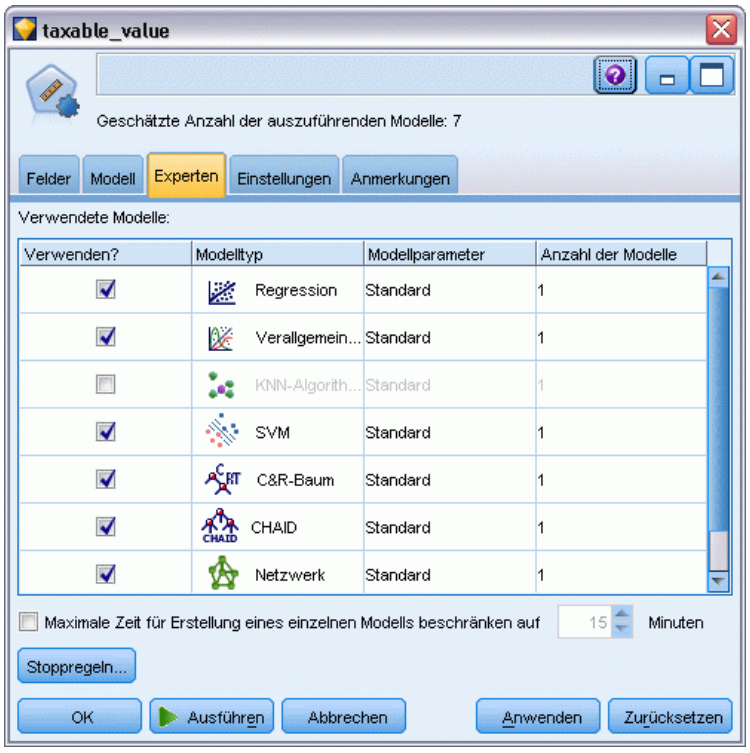
Abbildung 6-4
Knoten “Auto-Numerisch” – Registerkarte “Modell”



- Behalten Sie auf der Registerkarte “Experten” die Standardeinstellungen bei. Der Knoten schätzt ein einzelnes Modell für jeden Algorithmus (insgesamt sieben Modelle). (Alternativ können Sie diese Einstellungen ändern, um für jeden Modelltyp mehrere Varianten zu vergleichen.)

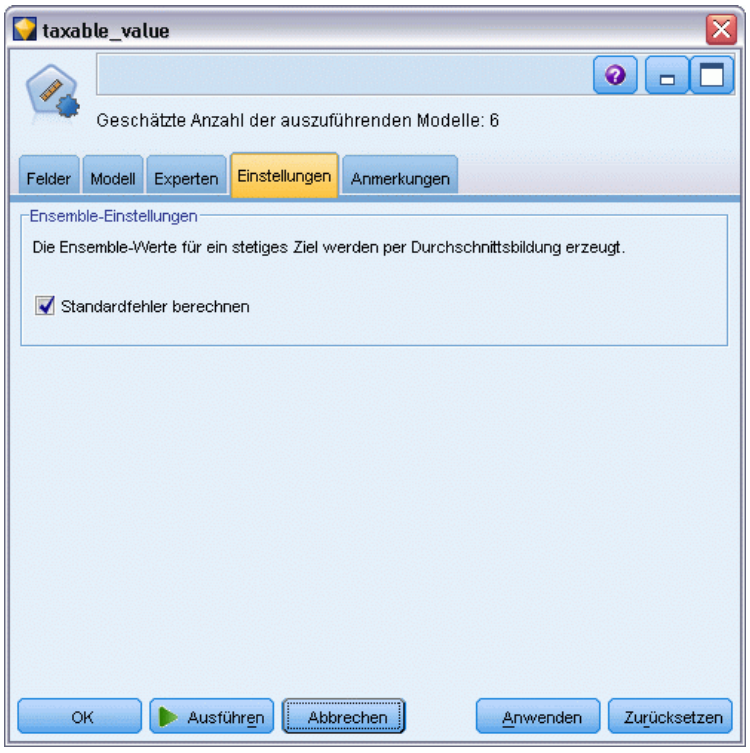
Da Sie auf der Registerkarte “Modell” für Anzahl der zu verwendenden Modelle 3 eingestellt haben, berechnet der Knoten die Genauigkeit der sieben Algorithmen und erstellt ein einzelnes Modell-Nugget, in dem die drei genauesten enthalten sind.

Abbildung 6-5
Knoten “Auto-Numerisch” – Registerkarte “Experten”



- Behalten Sie auf der Registerkarte “Einstellungen” die Standardeinstellungen bei. Da es sich hier um ein stetiges Ziel handelt, wird der Ensemble-Score aus dem Durchschnitt der Scores für die Einzelmodelle gebildet.

Abbildung 6-6
Knoten “Auto-Numerisch” – Registerkarte “Einstellungen”



Vergleichen der Modelle

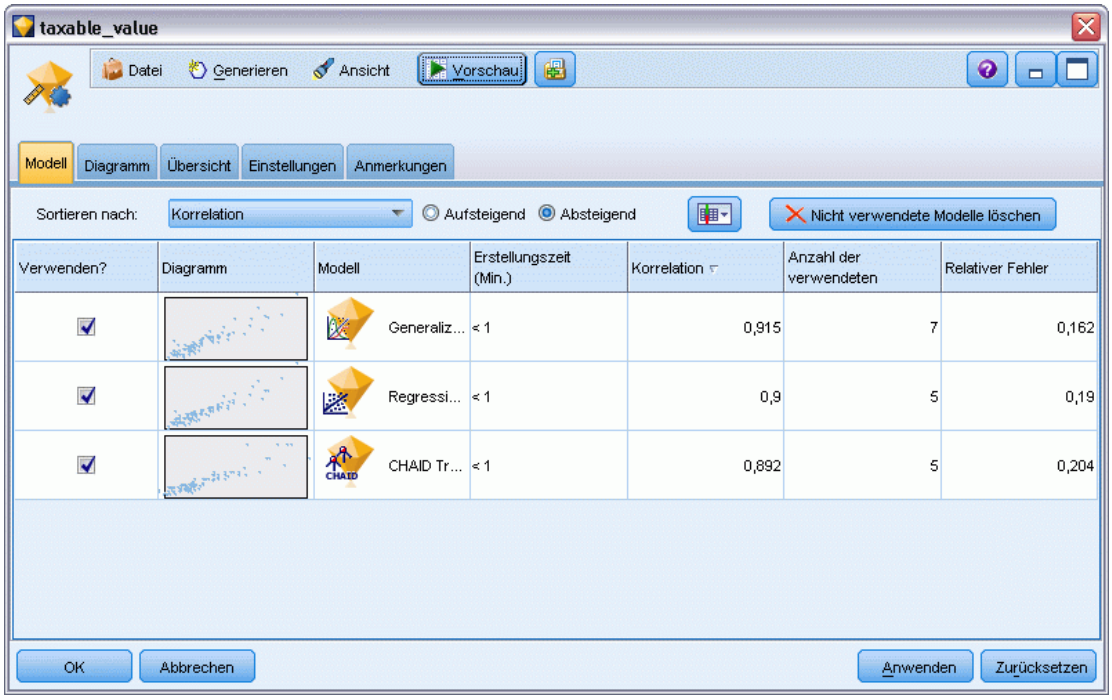
- Klicken Sie auf die Schaltfläche “Ausführen”.

Das Modell-Nugget wird erstellt und auf die Zeichenfläche sowie in die Modellpalette in der rechten oberen Fensterecke platziert. Sie können das Nugget durchsuchen oder auf verschiedene Weise speichern bzw. bereitstellen.

Öffnen Sie das Modell-Nugget. Es zeigt Details zu den einzelnen Modellen an, die während der Ausführung erstellt wurden. (In einer realen Situation, in der Hunderte von Modellen für ein großes Daten-Set geschätzt werden, kann dieser Vorgang etliche Stunden in Anspruch nehmen.) Unter Abbildung 6-1 auf S. 57.

Wenn Sie eines der Modelle eingehender untersuchen möchten, können Sie in der Spalte Modell auf ein Modell-Nugget-Symbol doppelklicken, um einen Drill-Down durchzuführen und die einzelnen Modellergebnisse zu durchsuchen. Anschließend können Sie Modellierungsknoten, Modell-Nuggets oder Evaluationsdiagramme generieren.

Abbildung 6-7
Auto-Numerisch - Ergebnisse



Standardmäßig werden die Modelle auf der Grundlage der Korrelation sortiert, da dies das Maß ist, das Sie im Knoten “Auto-Numerisch” ausgewählt haben. Für die Rangbildung wird der absolute Wert der Korrelation verwendet. Dabei deuten Werte nahe bei 1 auf eine stärkere Beziehung hin. Unter Verwendung dieses Maßes erhält das verallgemeinerte lineare Modell den besten Rang, mehrere andere sind jedoch fast ebenso genau. Das verallgemeinerte lineare Modell weist außerdem den geringsten relativen Fehler auf.

Sie können die Sortierung anhand einer anderen Spalte durchführen, indem Sie auf die Kopfzeile der betreffenden Spalte klicken. Außerdem können Sie das gewünschte Maß in der Liste Sortieren nach in der Symbolleiste auswählen.

Jedes Diagramm bietet für das Modell eine grafische Darstellung der beobachteten Werte in Abhängigkeit von den vorhergesagten Werten und ermöglicht dadurch einen schnellen Überblick über die Korrelation zwischen diesen Werten. Bei einem guten Modell sollten sich die Punkte entlang der Diagonale häufen, was bei allen Modellen in diesem Beispiel der Fall ist.

In der Spalte Diagramm können Sie ein Diagramm in voller Größe generieren, indem Sie auf eine Miniaturansicht doppelklicken.

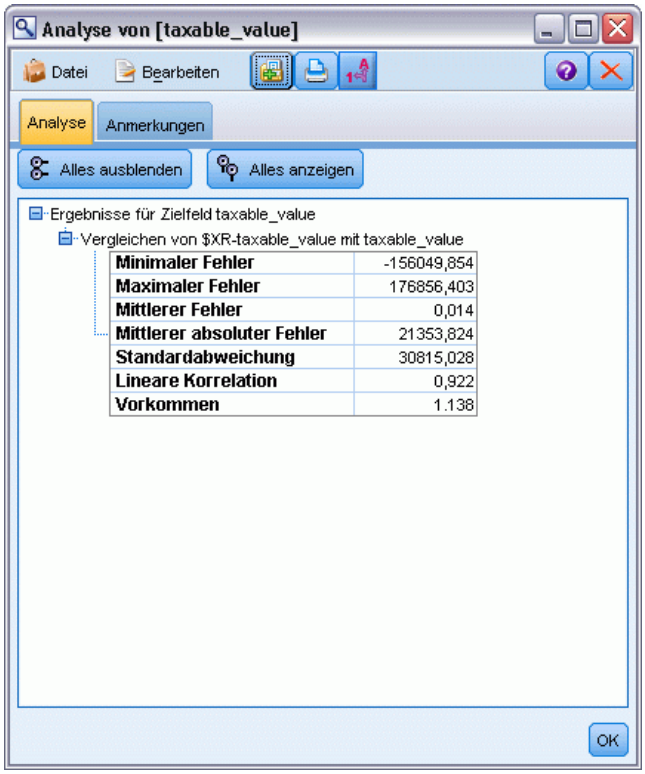
Basierend auf diesen Ergebnissen entscheiden Sie sich, jedes der drei exaktesten Modelle zu verwenden. Durch die Kombination der Vorhersagen aus mehreren Modellen können Beschränkungen in einzelnen Modellen vermieden werden, was zu einer höheren Gesamtgenauigkeit führt.

Stellen Sie sicher, dass in der Spalte Verwenden? alle drei Modelle ausgewählt sind.

Gliedern Sie einen Analyseknoten (Ausgabepalette) nach dem Modell-Nugget an. Klicken Sie mit der rechten Maustaste auf den Analyseknoten und wählen Sie Ausführen aus, um den Stream auszuführen.

Der vom Ensemble-Modell generierte Durchschnitts-Score wird im Feld *$XR-taxable_value* hinzugefügt. Die Korrelation liegt hier bei 0,922 und ist somit besser als bei den drei Einzelmodellen. Die Ensemble-Scores weisen außerdem einen niedrigeren mittleren absoluten Fehler auf und erzielen bei Anwendung auf andere Daten-Sets möglicherweise eine bessere Leistung als jedes der Einzelmodelle.

Abbildung 6-8
Auto-Numerisch – Beispiel-Stream



Zusammenfassung

Sie haben mithilfe des Knotens “Auto-Numerisch” eine Reihe verschiedener Modelle verglichen, die drei genauesten Modelle ausgewählt und dem Stream hinzugefügt und diese Modelle schließlich in einem Modell-Nugget “Auto-Numerisch” zusammengefasst.

- Hinsichtlich der Gesamtgenauigkeit erbrachten die verallgemeinerten linearen Modelle, die Regressionsmodelle und die CHAID-Modelle die besten Leistungen bei den Trainingsdaten.
- Das Modell-Ensemble erzielte eine Leistung, die zweien der Einzelmodelle überlegen war, und erbringt bei Anwendung auf andere Datensätze möglicherweise noch bessere Leistungen. Wenn Sie den Prozess so weit wie möglich automatisieren möchten, können Sie mit diesem Ansatz unter den meisten Bedingungen ein robustes Modell erstellen, ohne sich allzu genau mit den spezifischen Eigenschaften der einzelnen Modelle befassen zu müssen.

Teil II:
Beispiele für die Datenvorbereitung

Kapitel
7

Automatische Datenvorbereitung (ADP)

Die Vorbereitung von Daten für die Analyse ist einer der wichtigsten Schritte in jedem Data-Mining-Projekt – und traditionell auch einer der zeitaufwendigsten. Der ADP-Knoten (Automatische Datenvorbereitung) erledigt die Aufgabe für Sie, er analysiert Ihre Daten und identifiziert Korrekturen, schließt problematische oder wahrscheinlich überflüssige Felder aus, leitet falls erforderlich neue Attribute ab und verbessert die Leistung durch intelligente Prüfverfahren. Sie können den Knoten vollständig automatisiert nutzen, damit er Korrekturen wählen und anwenden kann. Sie können die Änderungen aber auch prüfen, bevor sie durchgeführt werden, und wie gewünscht akzeptieren oder ablehnen.

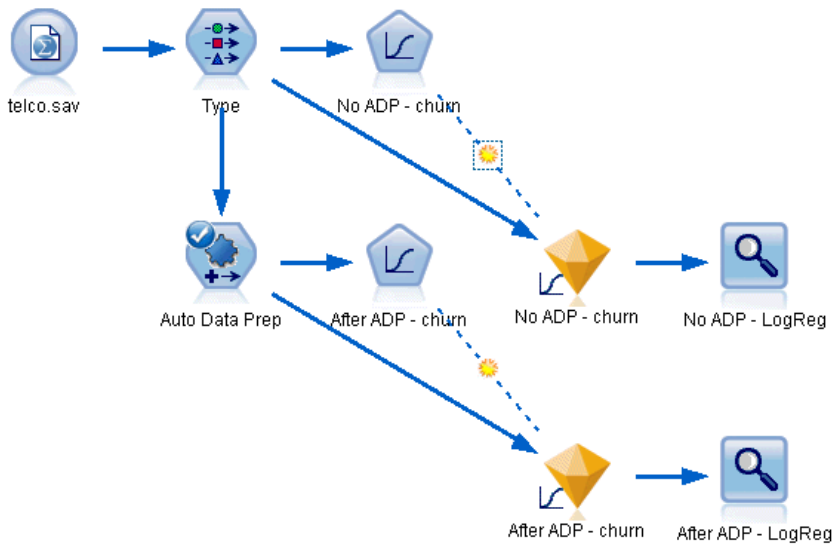
Mit dem ADP-Knoten können Sie Ihre Daten schnell und einfach für Data-Mining vorbereiten, ohne dass vorherige Kenntnisse zu den verwendeten Statistikkonzepten erforderlich sind. Wenn Sie den Knoten mit den Standardeinstellungen ausführen, werden Modelle gewöhnlich schneller erstellt und bewertet.

Dieses Beispiel verwendet den Stream *ADP_basic_demo.str*, der auf die Datendatei *telco.sav* verweist, um die höhere Genauigkeit zu demonstrieren, die beim Erstellen von Modellen mithilfe der Standardeinstellungen des ADP-Knotens erzielt werden kann. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *ADP_basic_demo.str* befindet sich im Verzeichnis *streams*.

Erstellen des Streams

- Um den Stream zu erstellen, fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf die Datei *telco.sav* im Verzeichnis *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist.

Abbildung 7-1
Erstellen des Streams

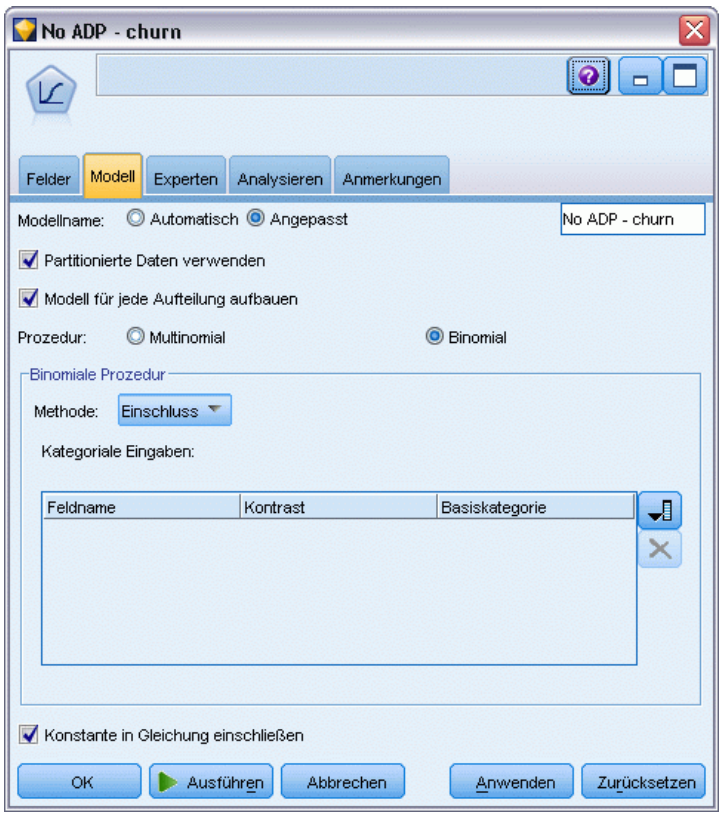


- Fügen Sie dem Quellenknoten einen Typknoten hinzu, stellen Sie das Messniveau für das Feld *churn* auf Flag und die Rolle auf Ziel ein. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

Abbildung 7-2
Auswahl des Ziels

- Verbinden Sie einen logistischen Knoten mit dem Typknoten.
- Klicken Sie im Logistikknoten auf die Registerkarte “Modell” und wählen Sie die Prozedur Binomial aus. Wählen Sie im Feld *Modellname* die Option Benutzerdefiniert aus und geben Sie Ohne ADP - churn an.

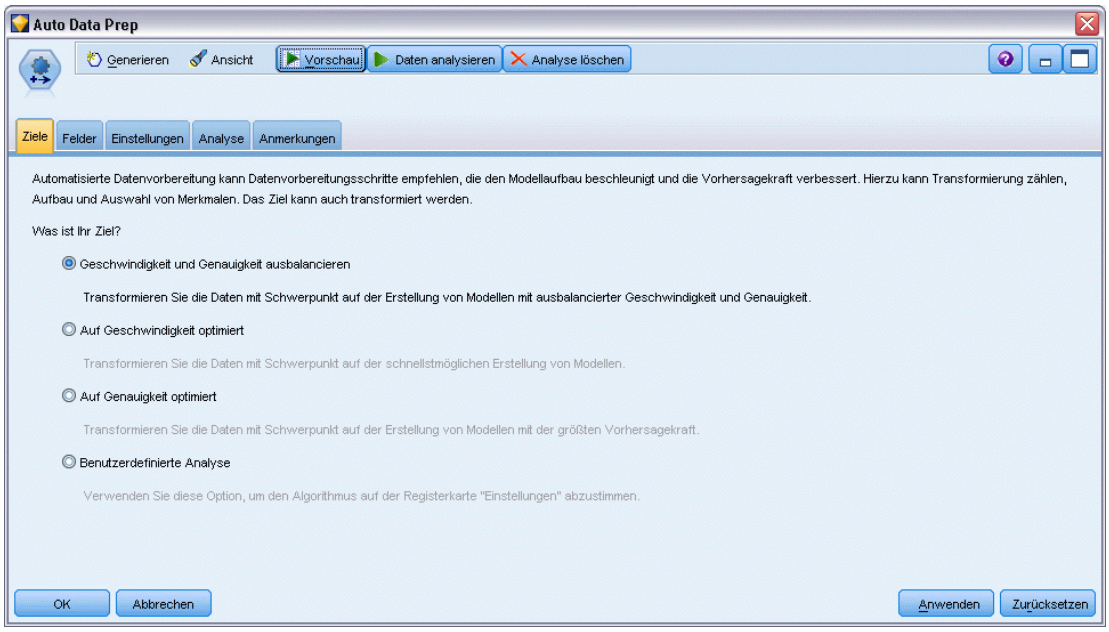
Abbildung 7-3
Auswählen Modelloptionen



- Verbinden Sie einen ADP-Knoten mit dem Typknoten. Belassen Sie in der Registerkarte “Ziele” die Standardeinstellungen, um Ihre Daten durch Ausgleich von Geschwindigkeit und Genauigkeit vorzubereiten.
- Klicken Sie am oberen Rand der Registerkarte “Ziele” auf Daten analysieren, um Ihre Daten zu analysieren und zu verarbeiten.

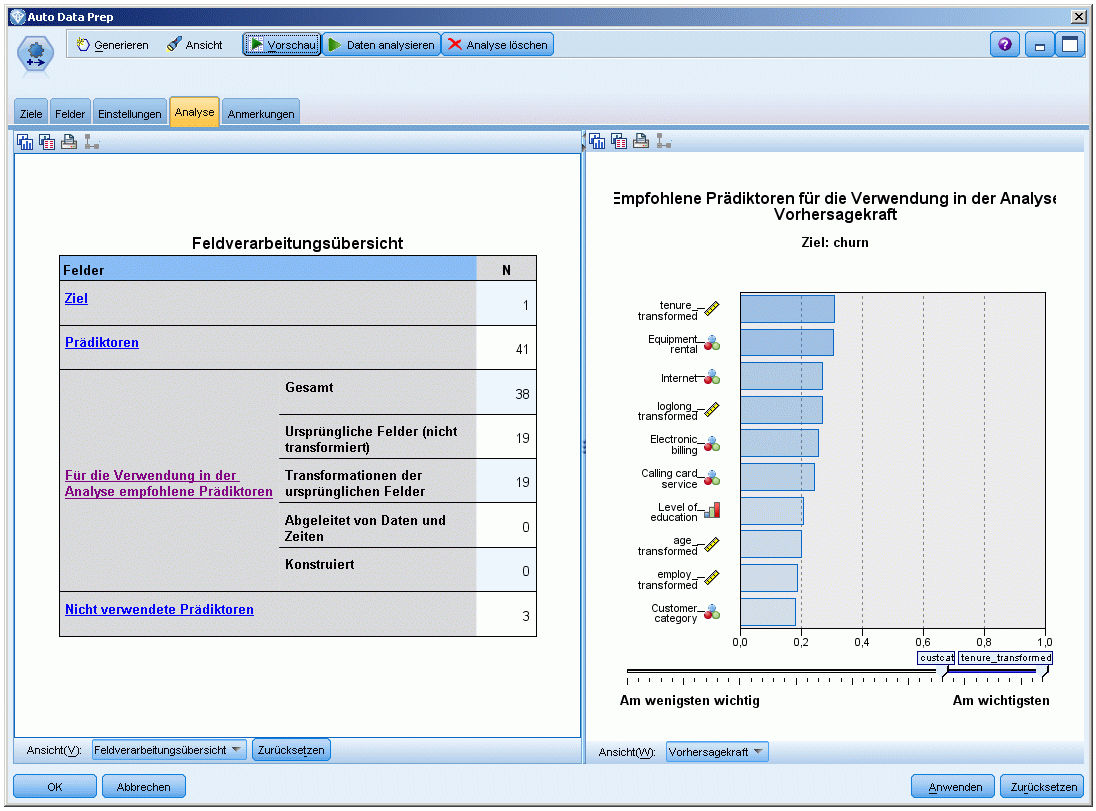
Mithilfe anderer Optionen des ADP-Knotens können Sie festlegen, dass größerer Wert auf Genauigkeit oder auf Verarbeitungsgeschwindigkeit gelegt werden soll, oder viele Verarbeitungsschritte der Datenvorbereitung präzise einstellen.

Abbildung 7-4
ADP-Standardziele



Die Ergebnisse der Datenverarbeitung werden in der Registerkarte “Analyse” angezeigt. Die Feldverarbeitungsübersicht zeigt, dass von den 41 Datenmerkmalen im ADP-Knoten 19 zur Unterstützung der Verarbeitung transformiert und 3 als nicht benutzt verworfen wurden.

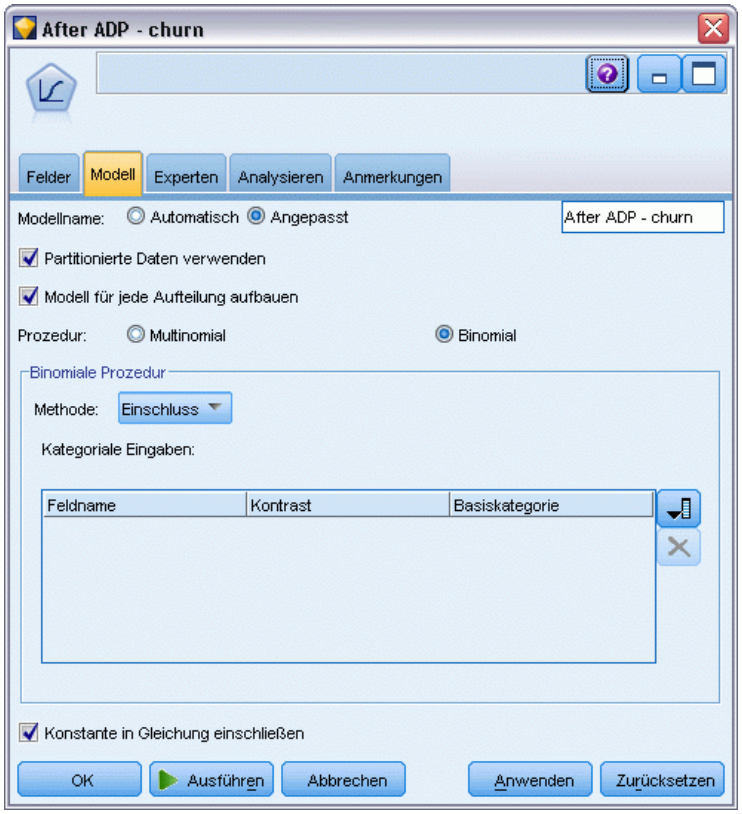
Abbildung 7-5
Übersicht der Datenverarbeitung



- Verbinden Sie einen logistischen Knoten mit dem ADP-Knoten.

- Klicken Sie im Logistikknoten auf die Registerkarte “Modell” und wählen Sie die Prozedur Binomial aus. Wählen Sie im Feld *Modellname* die Option Benutzerdefiniert aus und geben Sie Nach ADP - churn an.

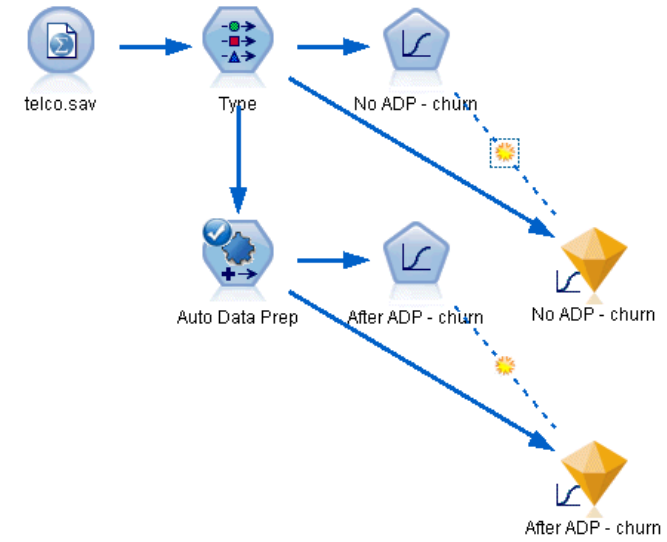
Abbildung 7-6
Auswählen Modelloptionen



Vergleichen der Modellgenauigkeit

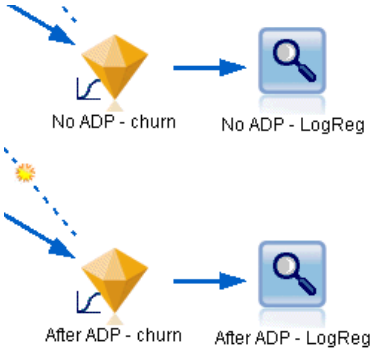
- Führen Sie beide Logistikknoten aus, um die Modell-Nuggets zu erstellen; diese werden dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt.

Abbildung 7-7
Anfügen der Modell-Nuggets



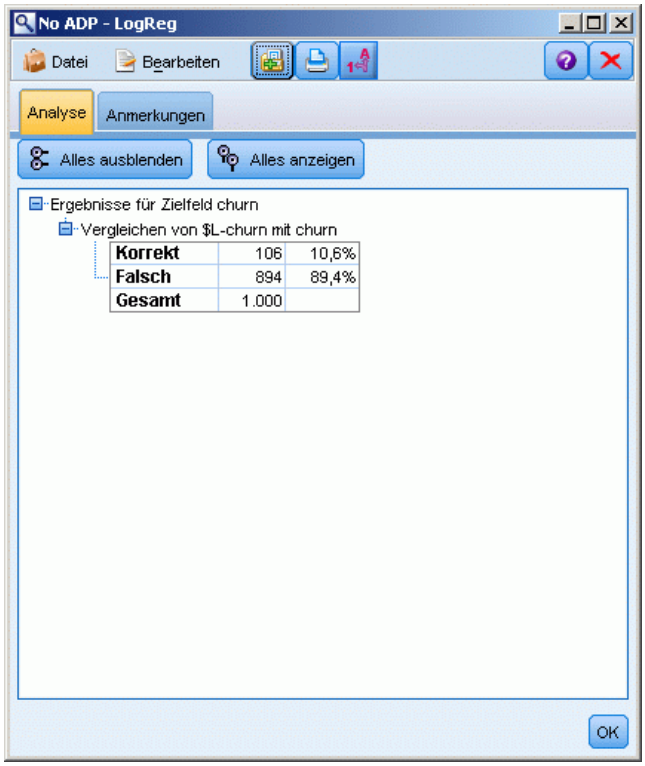
- Fügen Sie Analyseknoten an die Modell-Nuggets an und führen Sie die Analyseknoten mit den zugehörigen Standardeinstellungen aus.

Abbildung 7-8
Anfügen der Analyseknoten



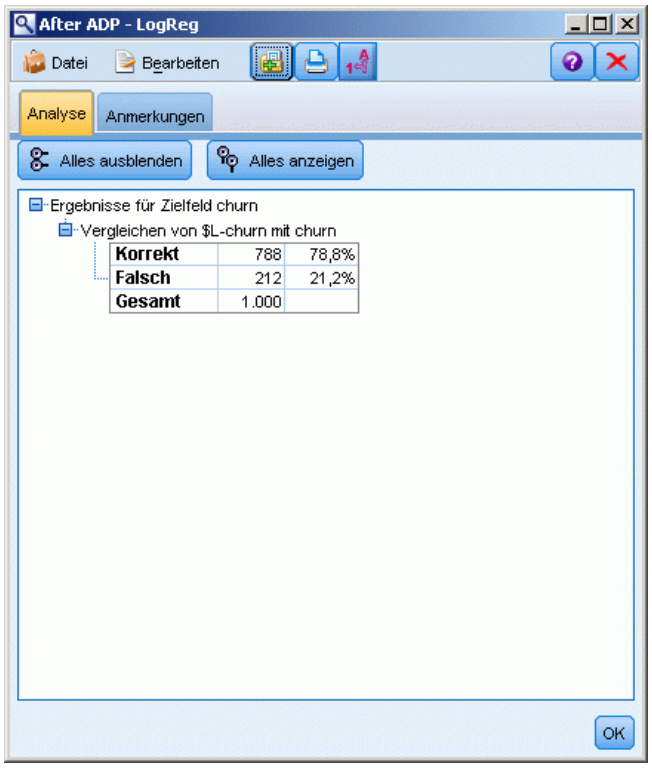
Die Analyse des Modells ohne ADP zeigt, dass der einfache Lauf der Daten durch den logistischen Regressionsknoten mit den Standardeinstellungen ein Modell mit geringer Genauigkeit ergibt – nur 10,6 %.

Abbildung 7-9
Ergebnisse des Modells ohne ADP



Die Analyse des Modells mit ADP zeigt, dass Sie durch den Lauf der Daten durch die ADP-Standardeinstellungen ein viel genaueres Modell erstellt haben, das zu 78,8 % korrekt ist.

Abbildung 7-10
Ergebnisse des Modells mit ADP



Zusammenfassend lässt sich sagen, dass Sie durch einfaches Ausführen des ADP-Knotens zur Feineinstellung der Verarbeitung Ihrer Daten ein genaueres Modell mit wenig direkter Datenmanipulation erstellen konnten.

Wenn Sie eine bestimmte Theorie beweisen oder widerlegen oder spezifische Modelle erstellen möchten, ist es offenbar nützlich, direkt mit den Modelleinstellungen zu arbeiten. Wenn Ihnen jedoch nur begrenzte Zeit zur Verfügung steht oder eine große Menge an Daten vorzubereiten ist, kann Ihnen der ADP-Knoten einen Vorteil liefern.

Erläuterungen der mathematischen Grundlagen für die in IBM® SPSS® Modeler verwendeten Modellierungsmethoden finden Sie im *SPSS Modeler-Algorithmushandbuch*, das sich im Verzeichnis *Documentation* des Installationsdatenträgers befindet.

Beachten Sie, dass die Ergebnisse in diesem Beispiel nur auf den Trainingsdaten beruhen. Um einzuschätzen, wie gut sich Modelle für andere Daten in der Praxis verallgemeinern lassen, könnten Sie mit einem Partitionsknoten eine Teilmenge der Datensätze für Test- und Validierungszwecke zurückhalten. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in *IBM SPSS Modeler 14.2- Quellen- , Prozess- und Ausgabeknoten*.

Kapitel 8

Vorbereiten von Daten für die Analyse (Data Audit)

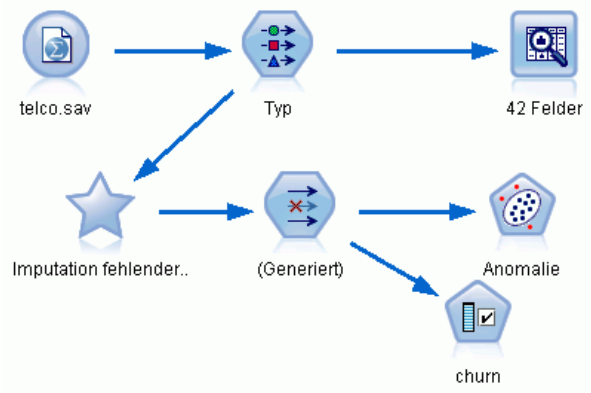
Der Data Audit-Knoten liefert einen umfassenden ersten Eindruck der Daten, die Sie in IBM® SPSS® Modeler einbringen. Der Data Audit-Bericht, der häufig im Rahmen der ersten Datenuntersuchung eingesetzt wird, zeigt Übersichtsstatistiken, Histogramme und Verteilungsdiagramme für die einzelnen Datenfelder. Außerdem können Sie hier angeben, wie fehlende Werte, Ausreißer und Extremwerte behandelt werden sollen.

In diesem Beispiel wird ein Stream namens *telco_dataaudit.str* verwendet, der Bezug nimmt auf die Datendatei *telco.sav*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "SPSS Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *telco_dataaudit.str* befindet sich im Verzeichnis *streams*.

Erstellen des Streams

- Um den Stream zu erstellen, fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf die Datei *telco.sav* im Verzeichnis *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist.

Abbildung 8-1
Erstellen des Streams



- Fügen Sie einen Typknoten hinzu, um Felder zu definieren, und geben Sie *churn* (Abwanderung) als Zielfeld (Rolle = Ziel) an. Für alle anderen Felder sollte die Rolle auf Eingabe gesetzt werden, sodass dies das einzige Zielfeld ist.

Abbildung 8-2
Festlegen des Ziels

- Vergewissern Sie sich, dass die Feldmessniveaus korrekt definiert wurden. So können beispielsweise die meisten Felder mit den Werten 0 und 1 als Flags betrachtet werden, manche

Felder, wie beispielsweise das Geschlecht, sollten jedoch besser als nominales Feld mit zwei Werten betrachtet werden.

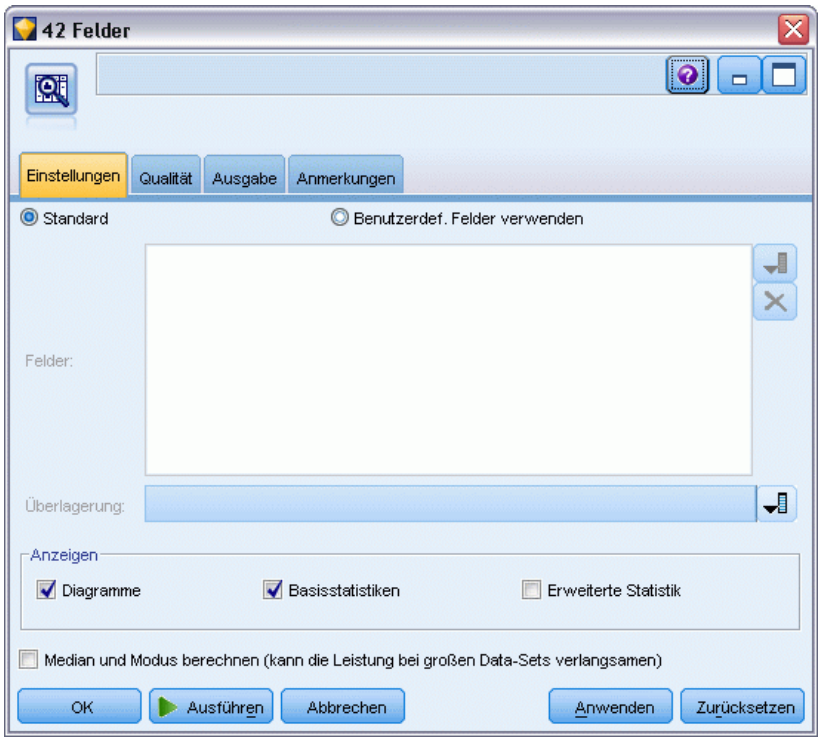
Abbildung 8-3
Festlegen von Messniveaus

Tipp: Um die Eigenschaften für mehrere Felder mit ähnlichen Werten (z. B. 0/1) zu ändern, klicken Sie auf die Überschrift der Spalte *Werte* (um die Felder nach dieser Spalte zu sortieren) und halten Sie anschließend die Umschalttaste gedrückt, um alle Felder auszuwählen, die geändert werden sollen. Anschließend können Sie mit der rechten Maustaste auf die Auswahl klicken, um das Messniveau oder andere Attribute für alle ausgewählten Felder zu ändern.

- Fügen Sie einen Data Audit-Knoten zum Stream hinzu. Behalten Sie auf der Registerkarte "Einstellungen" die Standardeinstellungen bei, um alle Felder in den Bericht aufzunehmen.

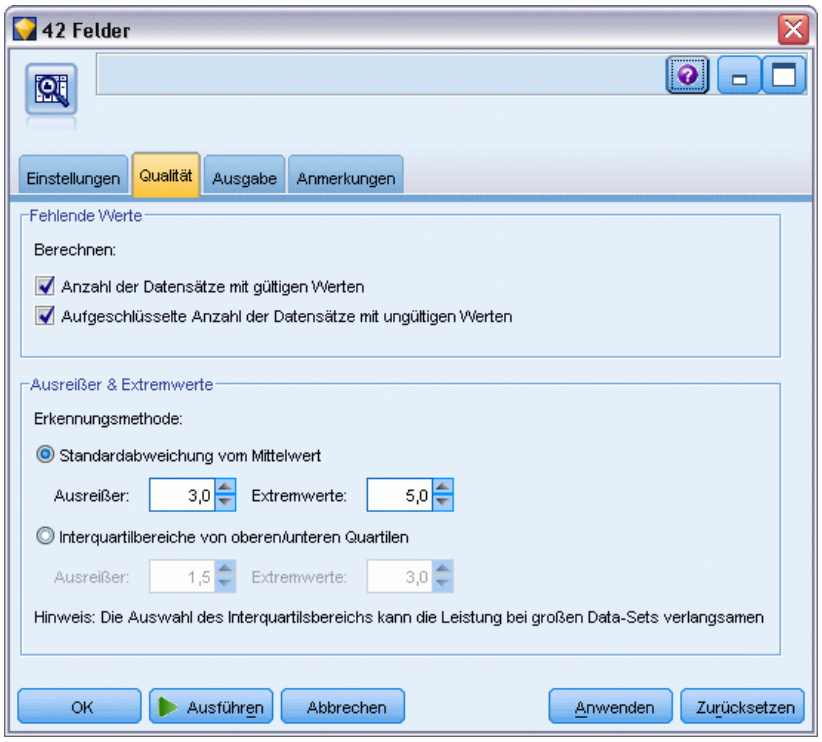
Da *churn* (Abwanderung) das einzige Zielfeld ist, das im Typknoten definiert wurde, wird es automatisch als Überlagerung verwendet.

Abbildung 8-4
Data Audit-Knoten – Registerkarte Einstellungen



Behalten Sie auf der Registerkarte “Qualität” die Standardeinstellungen für die Erkennung von fehlenden Werten, Ausreißern und Extremwerten bei und klicken Sie auf Ausführen.

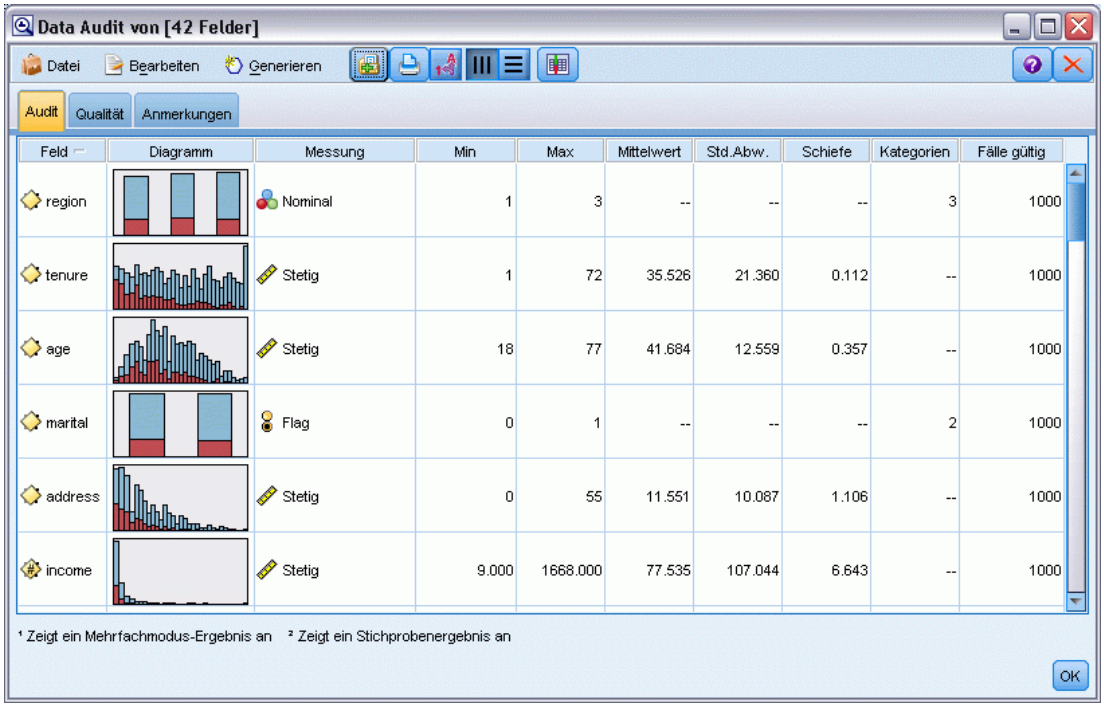
Abbildung 8-5
Data Audit-Knoten – Registerkarte “Qualität”



Durchsuchen von Statistiken und Diagrammen

Der Data Audit-Browser wird angezeigt. Er enthält Miniaturdiagramme und deskriptive Statistiken für die einzelnen Felder.

Abbildung 8-6
Data Audit-Browser



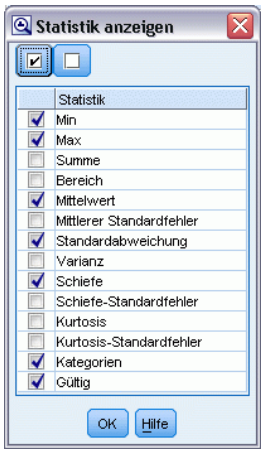
Feld	Diagramm	Messung	Min	Max	Mittelwert	Std.Abw.	Schiefe	Kategorien	Fälle gültig
region		Nominal	1	3	--	--	--	3	1000
tenure		Stetig	1	72	35.526	21.360	0.112	--	1000
age		Stetig	18	77	41.684	12.559	0.357	--	1000
marital		Flag	0	1	--	--	--	2	1000
address		Stetig	0	55	11.551	10.087	1.106	--	1000
income		Stetig	9.000	1668.000	77.535	107.044	6.643	--	1000

[1] Zeigt ein Mehrfachmodus-Ergebnis an [2] Zeigt ein Stichprobenergebnis an

Verwenden Sie die Symbolleiste, um Feld- und Wertelabels anzuzeigen und die Ausrichtung der Diagramme von horizontal in vertikal zu ändern (nur bei kategorialen Feldern).

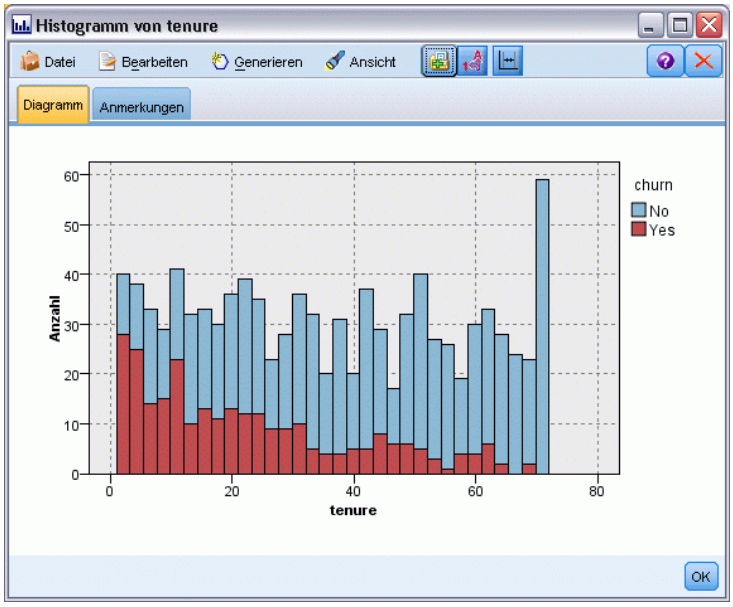
- Außerdem können Sie über die Symbolleiste oder das Menü “Bearbeiten” die anzuzeigenden Statistiken auswählen.

Abbildung 8-7
Statistik anzeigen



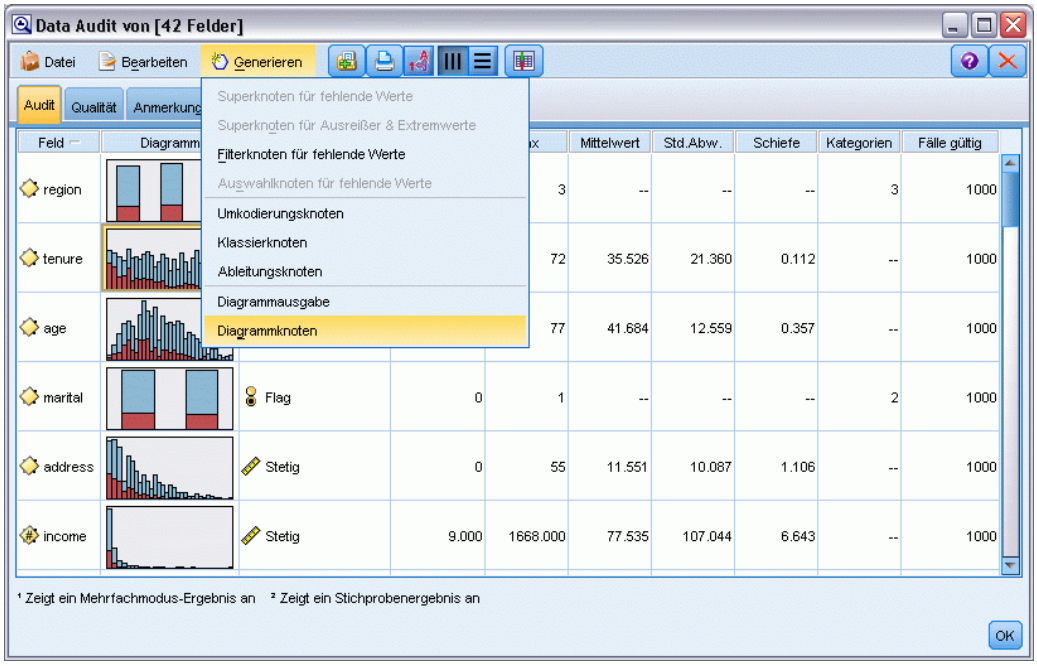
Doppelklicken Sie auf ein Miniaturdiagramm im Audit-Bericht, um eine Version des betreffenden Diagramms in voller Größe anzuzeigen. Da *churn* (Abwanderung) das einzige Zielfeld im Stream ist, wird es automatisch als Überlagerung verwendet. Mit der Symbolleiste des Diagrammfensters können Sie zwischen der Anzeige von Feld- und Wertelabels umschalten. Alternativ können Sie auf die Schaltfläche “Bearbeitungsmodus” klicken, um das Diagramm weiter anzupassen.

Abbildung 8-8
Histogramm der Beschäftigungsdauer



Alternativ können Sie eine oder mehrere Miniaturansichten auswählen und dafür jeweils einen Diagrammknoten generieren. Die generierten Knoten werden im Stream-Zeichenbereich platziert und können zum Stream hinzugefügt werden, um das betreffende Diagramm neu zu erstellen.

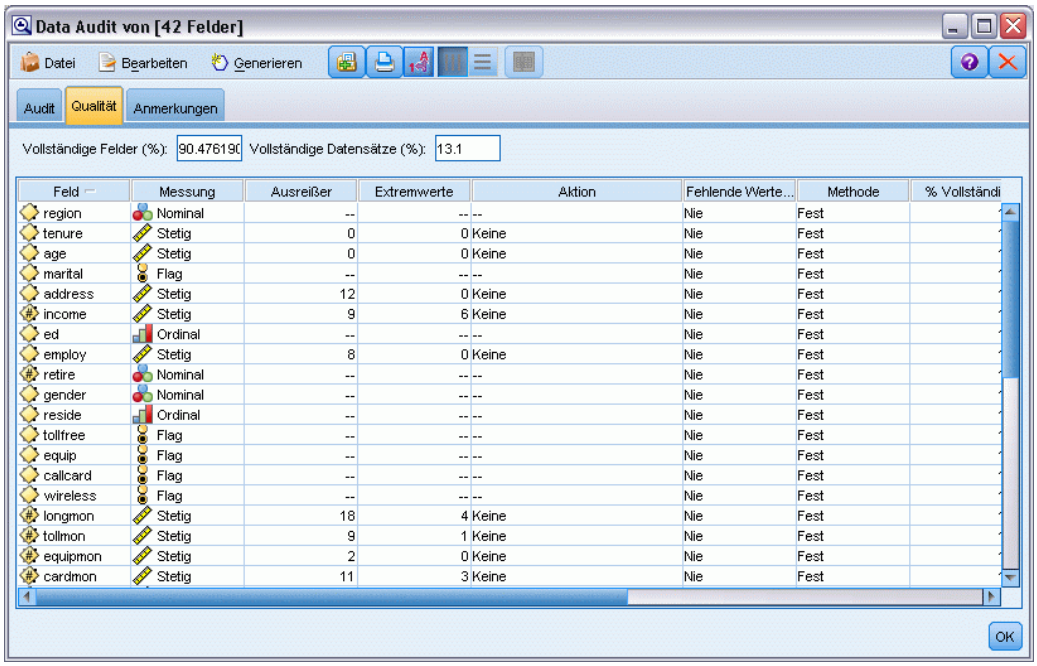
Abbildung 8-9
Generieren eines Diagrammknotens



Umgang mit Ausreißern und fehlenden Werten

Auf der Registerkarte “Qualität” des Audit-Berichts finden Sie Informationen zu Ausreißern, Extremwerten und fehlenden Werten.

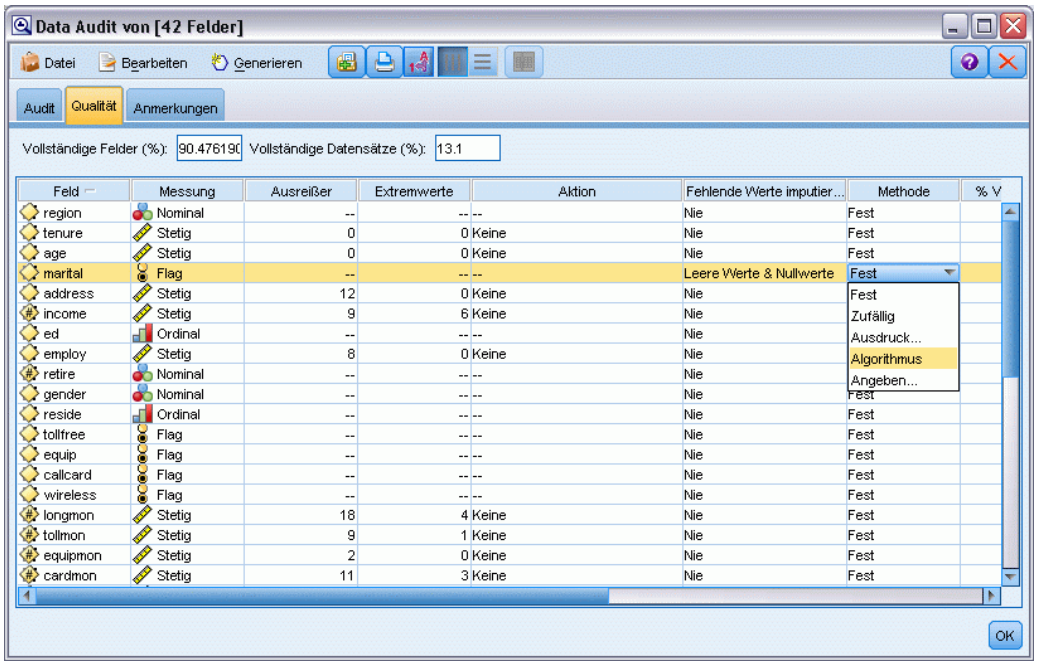
Abbildung 8-10
Data Audit-Browser – Registerkarte “Qualität”



Feld	Messung	Ausreißer	Extremwerte	Aktion	Fehlende Werte...	Methode	% Vollständi
region	Nominal	--	--	--	Nie	Fest	
tenure	Stetig	0	0	Keine	Nie	Fest	
age	Stetig	0	0	Keine	Nie	Fest	
marital	Flag	--	--	--	Nie	Fest	
address	Stetig	12	0	Keine	Nie	Fest	
income	Stetig	9	6	Keine	Nie	Fest	
ed	Ordinal	--	--	--	Nie	Fest	
employ	Stetig	8	0	Keine	Nie	Fest	
retire	Nominal	--	--	--	Nie	Fest	
gender	Nominal	--	--	--	Nie	Fest	
reside	Ordinal	--	--	--	Nie	Fest	
tollfree	Flag	--	--	--	Nie	Fest	
equip	Flag	--	--	--	Nie	Fest	
callcard	Flag	--	--	--	Nie	Fest	
wireless	Flag	--	--	--	Nie	Fest	
longmon	Stetig	18	4	Keine	Nie	Fest	
tollmon	Stetig	9	1	Keine	Nie	Fest	
equipmon	Stetig	2	0	Keine	Nie	Fest	
cardmon	Stetig	11	3	Keine	Nie	Fest	

Sie können Methoden für den Umgang mit diesen Werten angeben und Superknoten generieren, mit denen diese Transformationen automatisch angewendet werden können. Sie können beispielsweise ein oder mehrere Felder auswählen und fehlende Werte für diese Felder mit einer Reihe von Methoden imputieren bzw. ersetzen, beispielsweise mit dem C&RT-Algorithmus.

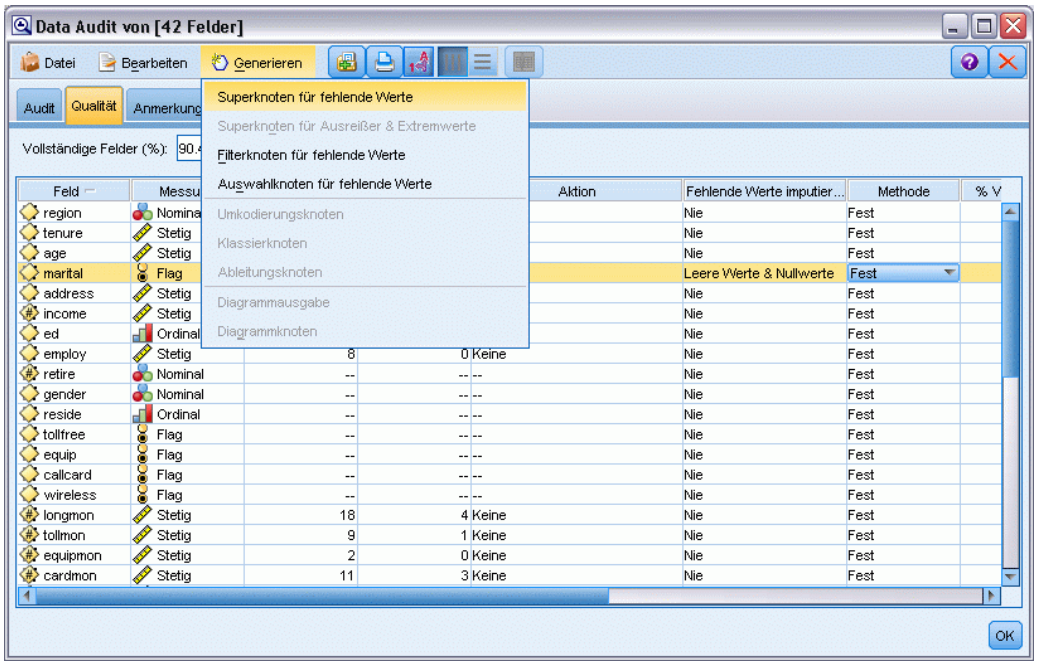
Abbildung 8-11
Auswahl einer Imputationsmethode



Nach der Angabe einer Eingabemethode für ein oder mehrere Felder können Sie einen Superknoten für fehlende Werte generieren. Wählen Sie dazu folgende Optionen in den Menüs aus:

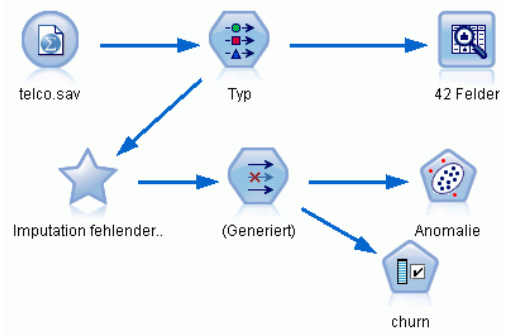
Erzeugen > Superknoten für fehlende Werte

Abbildung 8-12
Generieren des Superknotens



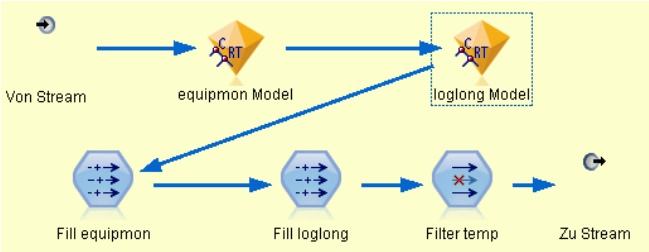
Der generierte Superknoten wird zum Stream-Zeichenbereich hinzugefügt. Dort können Sie ihn an den Stream anfügen, um die Transformationen anzuwenden.

Abbildung 8-13
Stream mit Superknoten für fehlende Werte



Der Superknoten enthält eine Reihe von Knoten, die die angeforderten Transformationen durchführen. Um einen Einblick in die Funktionsweise des Superknotens zu erhalten, können Sie ihn bearbeiten und auf Vergrößern klicken.

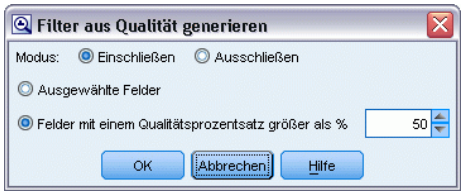
Abbildung 8-14
Vergrößern des Superknotens.



Für jedes Feld, das beispielsweise unter Verwendung der Algorithmusmethode imputiert wurde, gibt es ein separates C&RT-Modell sowie einen Füllerknoten, der Leerstellen und Nullen durch den vom Modell vorhergesagten Wert ersetzt. Sie können einzelne Knoten innerhalb des Superknotens hinzufügen, bearbeiten bzw. entfernen, um das Verhalten weiter anzupassen.

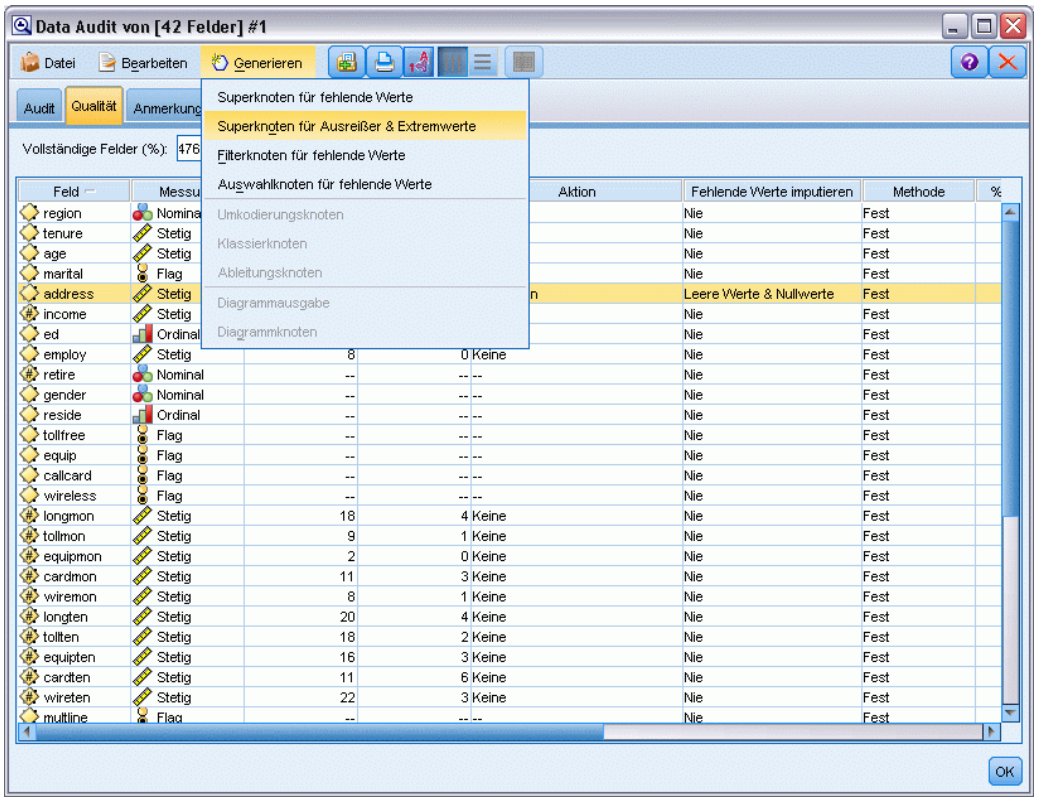
Alternativ können Sie einen Auswahl- oder Filterknoten generieren, um Felder oder Datensätze mit fehlenden Werten zu entfernen. Sie können beispielsweise alle Felder herausfiltern, deren Qualitätsprozentsatz unter einem angegebenen Schwellenwert liegt.

Abbildung 8-15
Generieren eines Filterknotens



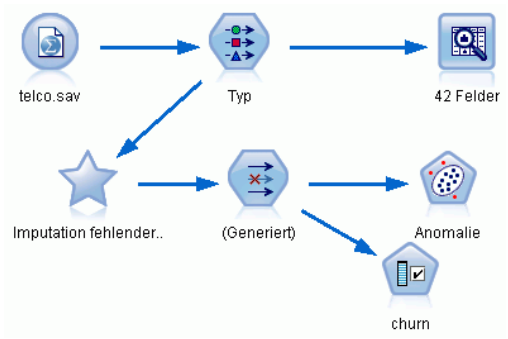
Ausreißer und Extremwerte können auf ähnliche Weise behandelt werden. Geben Sie die Aktion an, die Sie für die einzelnen Felder durchführen möchten – erzwingen, verwerfen oder auf null setzen – und generieren Sie einen Superknoten zur Anwendung der Transformationen.

Abbildung 8-16
Generieren eines Filterknotens



Nachdem der Audit abgeschlossen wurde und die generierten Knoten zum Stream hinzugefügt wurden, können Sie mit Ihrer Analyse fortfahren. Optional können Sie eine weitere Sichtung der Daten mithilfe der Anomalieerkennung, der Merkmalsauswahl bzw. einer Reihe anderer Methoden vornehmen.

Abbildung 8-17
Stream mit Superknoten für fehlende Werte



Kapitel
9

Medikamentöse Behandlung (Explorative Diagramme/C5.0)

In diesem Abschnitt schlüpfen Sie in die Rolle eines Medizinforschers, der Daten für eine Studie zusammenstellen soll. Sie haben 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. Ihre Aufgabe besteht u. a. darin, mithilfe von Data-Mining herauszufinden, welches Medikament in Zukunft für einen Patienten geeignet sein kann, der an derselben Krankheit leidet.

In diesem Beispiel wird ein Stream namens *druglearn.str* verwendet, der Bezug nimmt auf die Datendatei *DRUG1n*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *druglearn.str* befindet sich im Verzeichnis *streams*.

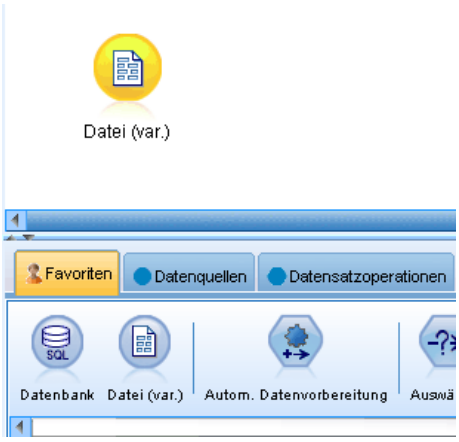
In dieser Demonstration werden die folgenden Datenfelder verwendet:

Datenfeld	Beschreibung
Alter	(Zahl)
Sex	*M* oder *F*
BP	Blutdruck: *HIGH*, *NORMAL* oder *LOW*
Cholesterol	Cholesterinspiegel im Blut: *NORMAL* oder *HIGH*
Na	Natriumkonzentration im Blut
C	Kaliumkonzentration im Blut
Drug	Medikament, auf das ein Patient ansprach

Einlesen von Textdaten

Textdaten mit Trennzeichen können mithilfe eines **Knotens für variable Dateien** eingelesen werden. Sie können einen Variablendateiknoten aus den Paletten hinzufügen: Klicken Sie entweder auf die Registerkarte Datenquellen, um den Knoten zu suchen, oder verwenden Sie die Registerkarte Favoriten, auf der dieser Knoten standardmäßig enthalten ist. Doppelklicken Sie dann auf den neu eingefügten Knoten, um das zugehörige Dialogfeld zu öffnen.

Abbildung 9-1
Hinzufügen eines Knotens vom Typ "Datei (var.)"



Klicken Sie auf die Schaltfläche, die sich direkt rechts neben dem Feld "Datei" befindet und mit Auslassungszeichen (...) gekennzeichnet ist, um in das Verzeichnis zu wechseln, in dem IBM® SPSS® Modeler auf Ihrem System installiert ist. Öffnen Sie das Verzeichnis *Demos* und wählen Sie die Datei *DRUG1n* aus.

Stellen Sie sicher, dass Feldnamen aus Datei lesen ausgewählt ist, und achten Sie auf die Felder und Werte, die gerade in das Dialogfeld geladen wurden.

Abbildung 9-2
Dialogfeld "Datei (var.)"

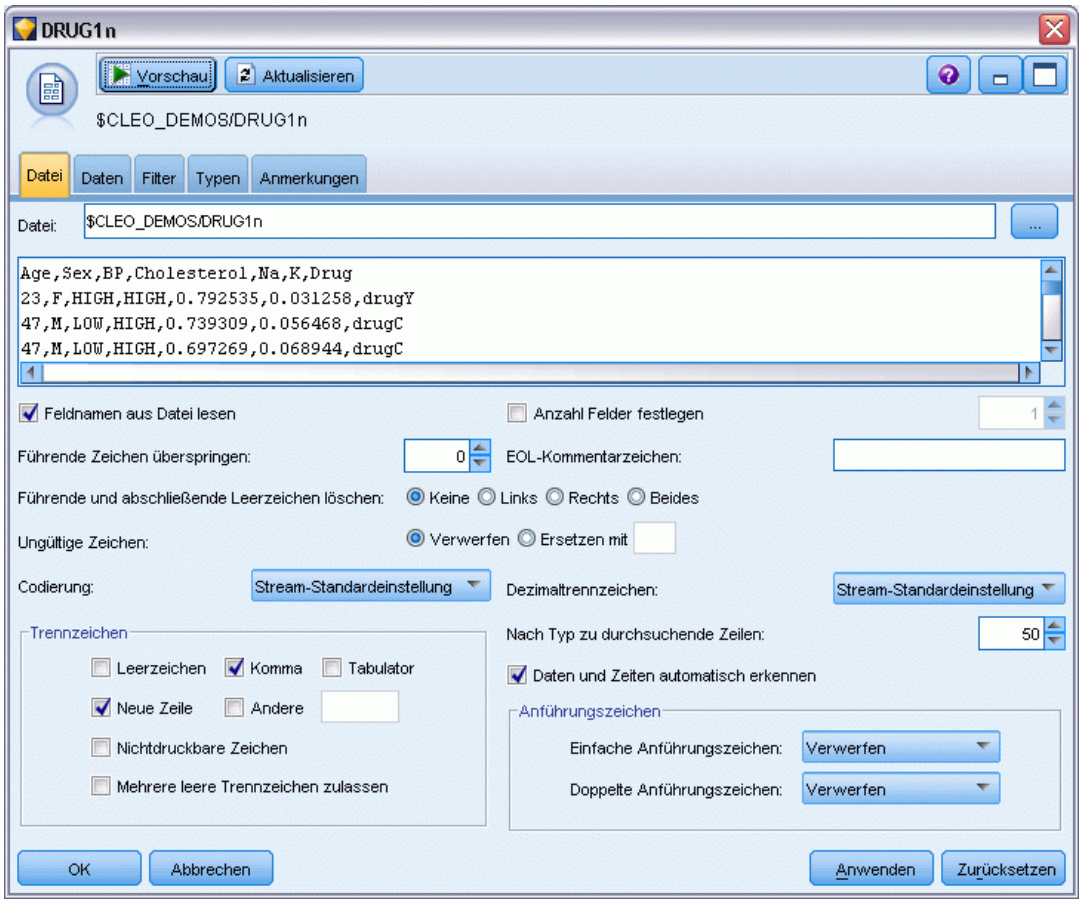


Abbildung 9-3
Ändern des Speichertyps für ein Feld

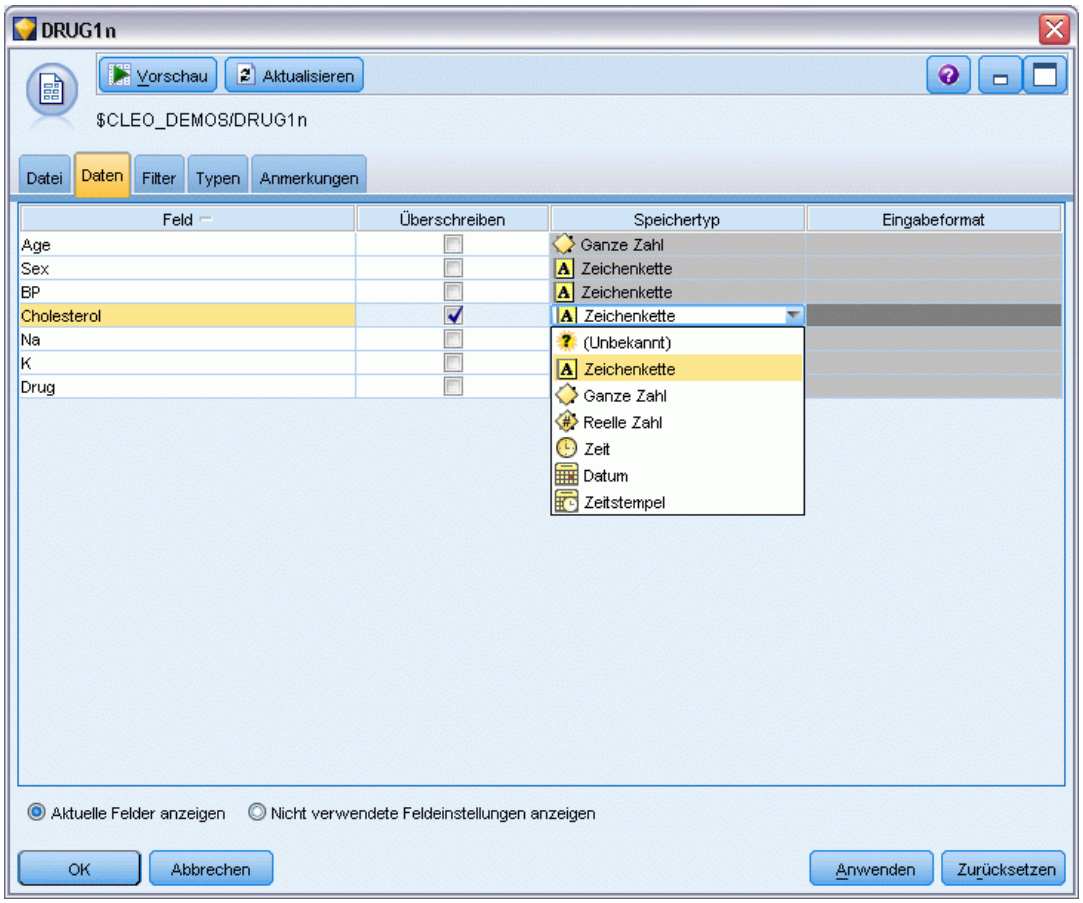
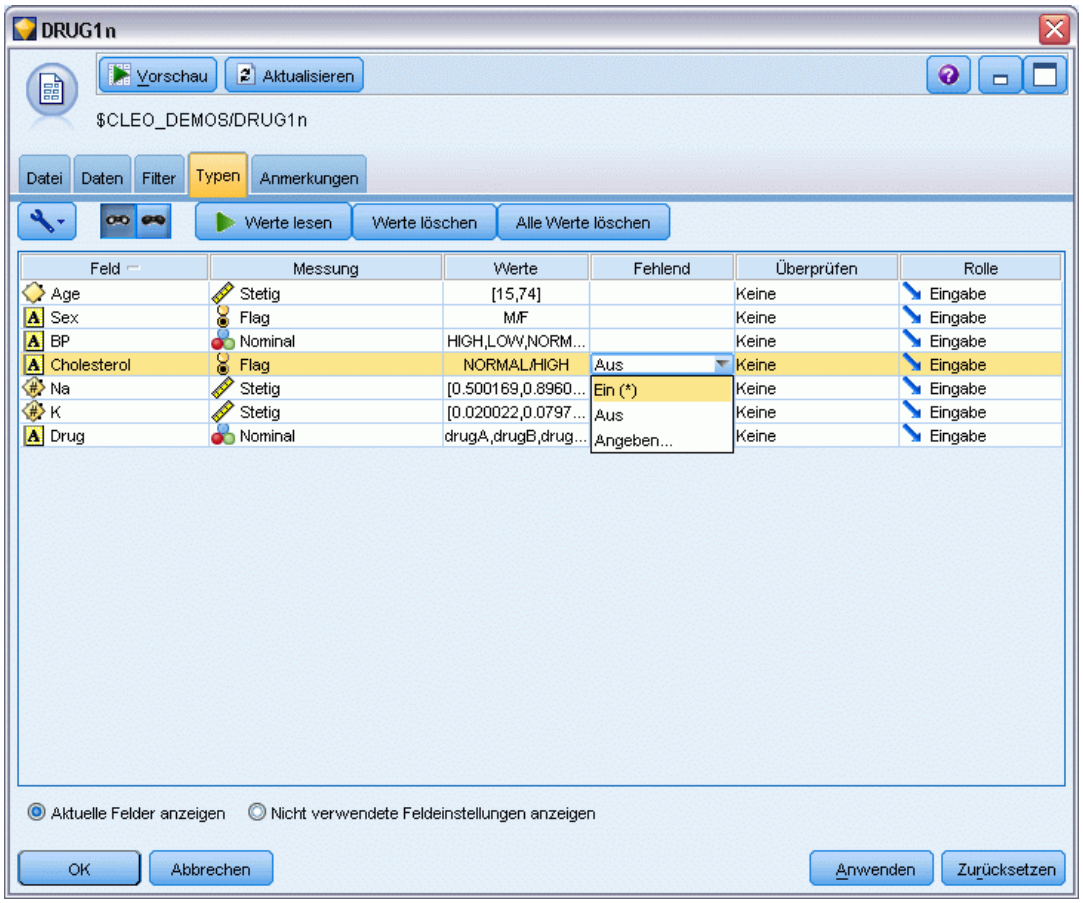


Abbildung 9-4
Auswählen von Werteoptionen auf der Registerkarte "Typen"



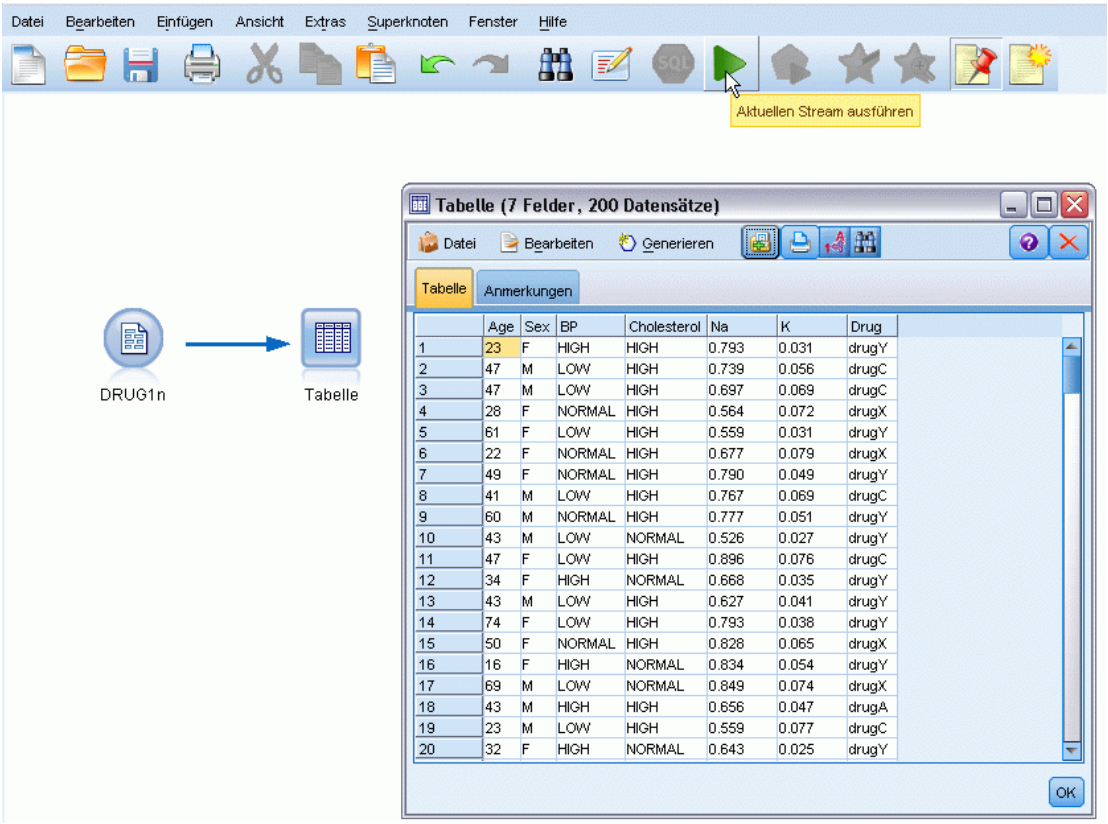
Klicken Sie auf die Registerkarte Daten, um den **Speichertyp** eines Felds zu überschreiben und zu ändern. Beachten Sie, dass sich der Speichertyp von **Messung**, d. h. dem Messniveau (oder Verwendungstyp) des Datenfelds unterscheidet. Auf der Registerkarte Typen können Sie Näheres zu den Feldtypen in Ihren Daten erfahren. Sie können auch Werte lesen wählen, um basierend auf der Auswahl, die Sie in der Spalte *Werte* vorgenommen haben, die tatsächlichen Werte für jedes Feld anzuzeigen. Dieser Prozess wird als **Instanziierung** bezeichnet.

Hinzufügen von Tabellen

Nachdem Sie nun die Datendatei geladen haben, möchten Sie vielleicht einen Blick auf die Werte einiger Datensätze werfen. Eine Möglichkeit hierfür besteht darin, einen Stream zu erstellen, der einen Tabellenknoten enthält. Um einen Tabellenknoten im Stream zu platzieren, doppelklicken Sie entweder in der Palette auf das entsprechende Symbol oder ziehen Sie es mit Ziehen und Ablegen auf den Zeichenbereich.

Abbildung 9-5
Mit der Datenquelle verbundener Tabellenknoten

Abbildung 9-6
Ausführen eines Streams über die Symbolleiste



	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

Wenn Sie auf einen Knoten in der Palette doppelklicken, wird dieser automatisch mit dem ausgewählten Knoten im Stream-Zeichenbereich verbunden. Falls die Knoten noch nicht bereits verbunden sind, können Sie mithilfe der mittleren Maustaste den Quellenknoten mit dem Tabellenknoten verbinden. Sie können die mittlere Maustaste simulieren, indem Sie die Alt-Taste gedrückt halten, während Sie die Maus verwenden. Wenn Sie die Tabelle anzeigen möchten, klicken Sie in der Symbolleiste auf die Schaltfläche mit dem grünen Pfeil, um den Stream auszuführen, oder klicken Sie mit der rechten Maustaste auf den Tabellenknoten und wählen Sie die Option Ausführen.

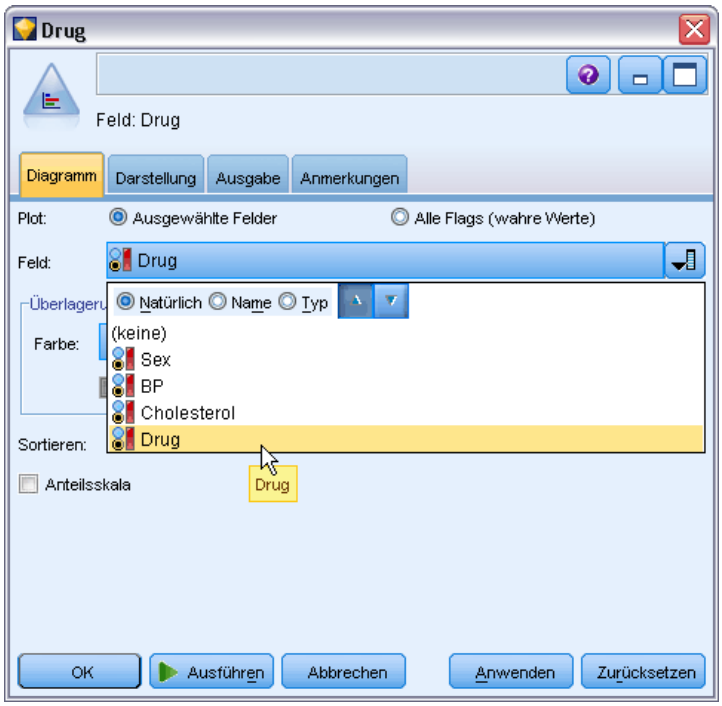
Erstellen eines Verteilungsdiagramms

Während des Data-Minings ist es häufig hilfreich, die Daten anhand einer visuellen Übersicht zu untersuchen. IBM® SPSS® Modeler stellt je nach Art der Daten, die Sie zusammenfassen möchten, verschiedene Diagrammtypen zur Auswahl. Wenn Sie z. B. herausfinden möchten, welcher Anteil der Patienten jeweils auf ein Medikament reagiert hat, verwenden Sie einen Verteilungsknoten.

Fügen Sie einen Verteilungsknoten zum Stream hinzu und verbinden Sie ihn mit dem Quellenknoten, doppelklicken Sie dann auf den Knoten, um die Anzeigeoptionen zu bearbeiten.

Wählen Sie *Drug* (Medikament) als das Zielfeld aus, dessen Verteilung Sie zeigen möchten. Klicken Sie dann im Dialogfeld auf Ausführen.

Abbildung 9-7
Auswählen von "drug" als Zielfeld



Aus dem so enstandenden Diagramm können Sie die "Form" der Daten erkennen. Diese zeigt, dass Patienten am häufigsten auf Medikament *Y* und am wenigsten auf Medikament *B* und *C* ansprachen.

Abbildung 9-8
Verteilung der Ansprechquoten auf den Medikamententyp

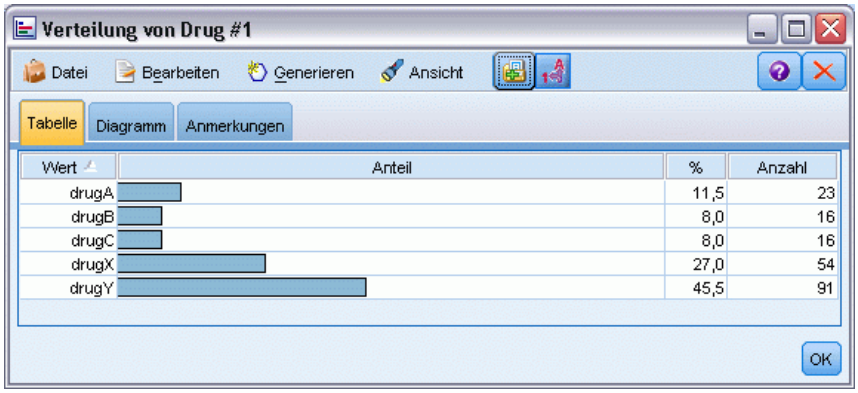
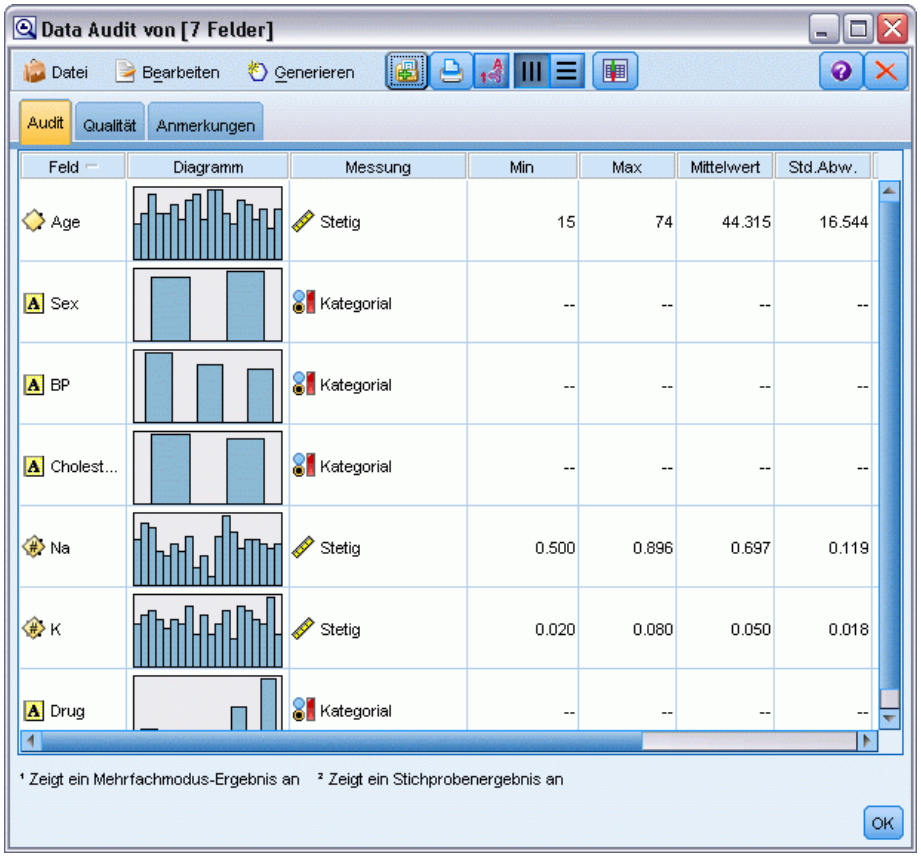


Abbildung 9-9
Ergebnisse eines Data Audit



Alternativ können Sie auch einen Data Audit-Knoten anfügen und ausführen, um sich einen raschen Überblick über die Verteilungen und Histogramme für alle Felder auf einmal zu verschaffen. Der Data Audit-Knoten ist auf der Registerkarte “Ausgabe” verfügbar.

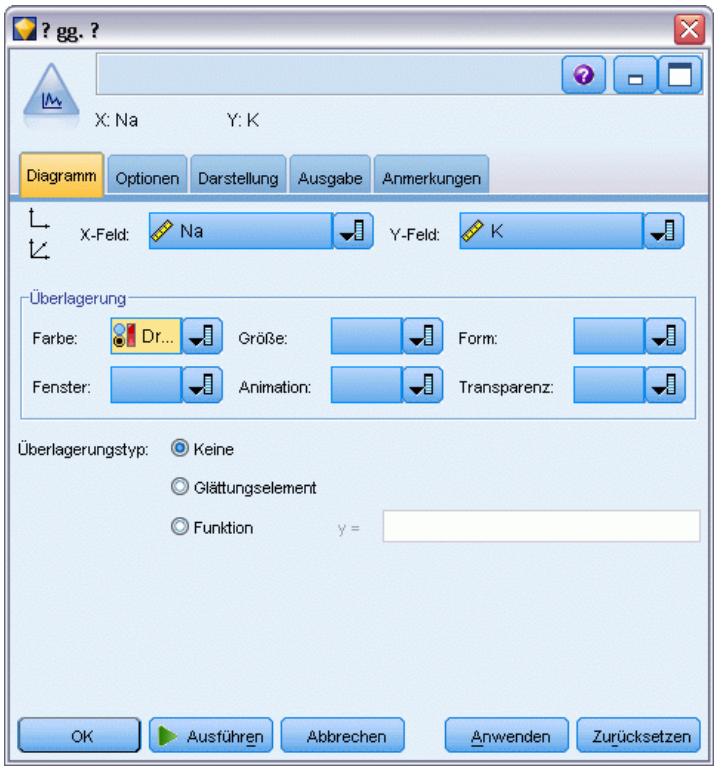
Erstellen eines Streudiagramms

Nun sehen wir uns an, welche Faktoren die Zielvariable *Drug* (Medikament) beeinflussen könnten. Als Medizinforscher wissen Sie, dass die Konzentration von Natrium und Kalium im Blut wichtige Faktoren sind. Da es sich beide Male um numerische Werte handelt, können Sie die Natrium/Kalium-Gegenüberstellung als Streudiagramm darstellen, in dem die Medikamentenkategorien farblich überlagert werden.

Platzieren Sie einen Plotknoten im Arbeitsbereich und verbinden Sie ihn mit dem Quellenknoten, doppelklicken Sie dann auf den Knoten, um ihn zu bearbeiten.

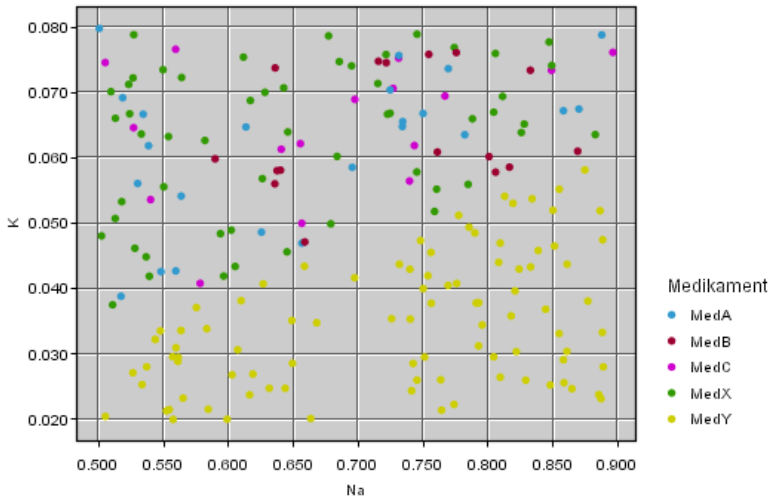
Wählen Sie auf der Registerkarte "Plot" *Na* als X-Feld, *K* als Y-Feld und *Drug* (Medikament) als Überlagerungsfeld aus. Klicken Sie dann auf Ausführen.

Abbildung 9-10
Erstellen eines Streudiagramms



Das Diagramm zeigt eindeutig einen Grenzwert auf, über dem das richtige Medikament immer Medikament *Y* und unter dem das richtige Medikament niemals Medikament *Y* ist. Bei diesem Grenzwert handelt es sich um den Quotienten, der sich aus dem Verhältnis von Natrium (*Na*) zu Kalium (*K*) ergibt.

Abbildung 9-11
Streudiagramm der Medikamentenverteilung

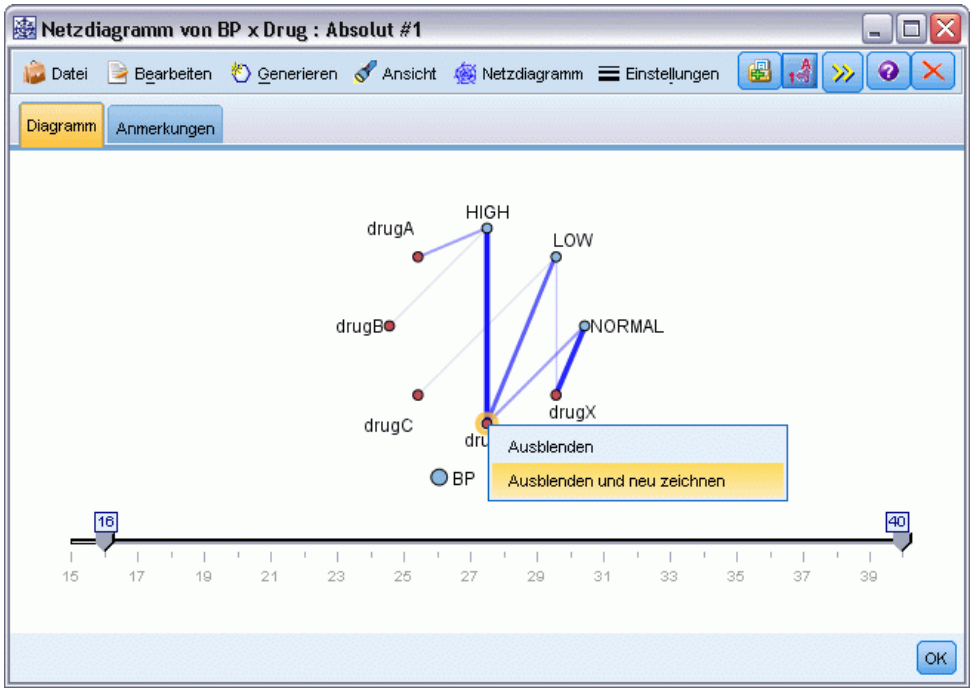


Erstellen eines Netzdiagramms

Da viele der Datenfelder kategorial sind, können sie auch versuchen, ein Netzdiagramm zu erstellen. Dieses stellt die Assoziationen zwischen verschiedenen Kategorien dar. Beginnen Sie, indem Sie einen Netzdiagrammknoten mit dem Quellenknoten in Ihrem Arbeitsbereich verbinden. Wählen Sie im Dialogfeld des Netzdiagrammknotens *BP* (Blutdruck) und *Drug* (Medikament). Klicken Sie dann auf Ausführen.

Aus dem Plot können wir entnehmen, dass Medikament *Y* mit allen drei Blutdruckstufen assoziiert ist. Dies ist nicht weiter überraschend, da Sie ja bereits die für Medikament *Y* geeignetste Situation ermittelt haben. Um sich auf die anderen Medikamente zu konzentrieren, können Sie Medikament *Y* ausblenden. Wählen Sie aus dem Menü Ansicht die Option Bearbeitungsmodus, klicken Sie dann mit der rechten Maustaste auf den Punkt für Medikament *Y* und wählen Sie Ausblenden und neu zeichnen.

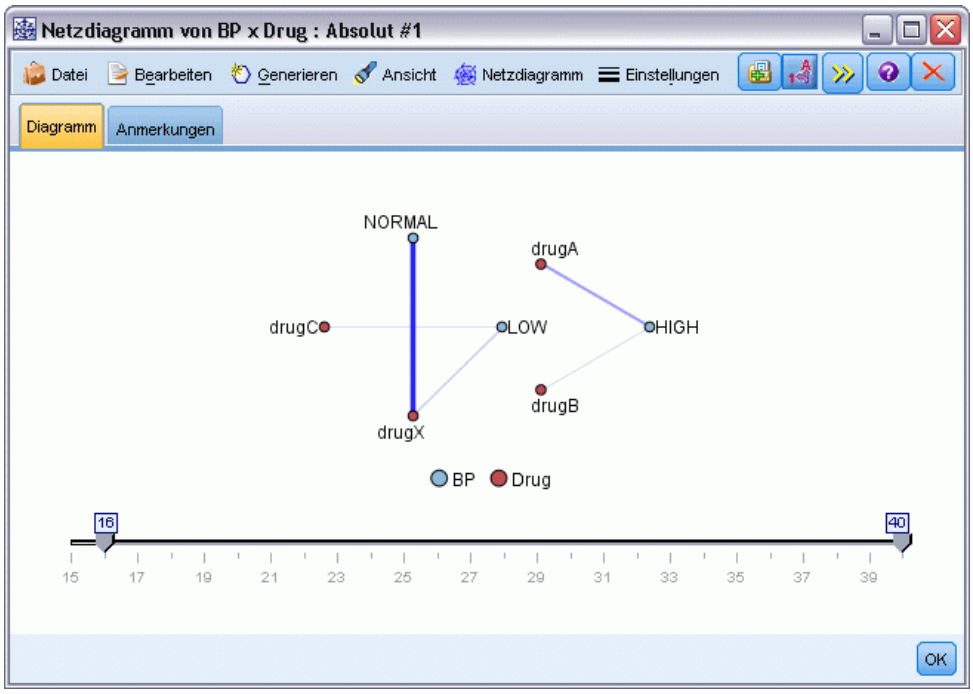
Abbildung 9-12
Netzdiagramm: Medikamente im Vergleich zum Blutdruck



In der vereinfachten Darstellung sind Medikament *Y* und alle zugehörigen Verbindungen ausgeblendet. Nun können Sie klar erkennen, dass nur Medikament *A* und Medikament *B* mit hohen Blutdruckwerten assoziiert sind. Nur die Medikamente *C* und *X* sind mit niedrigen Blutdruckwerten assoziiert. Und normale Blutdruckwerte werden nur mit Medikament *X* assoziiert. An diesem Punkt wissen Sie jedoch noch immer nicht, wie Sie bei einem Patienten

zwischen Medikament *A* und *B* bzw. zwischen Medikament *C* und *X* entscheiden sollen. Hierbei kann sich die Modellierung als hilfreich erweisen.

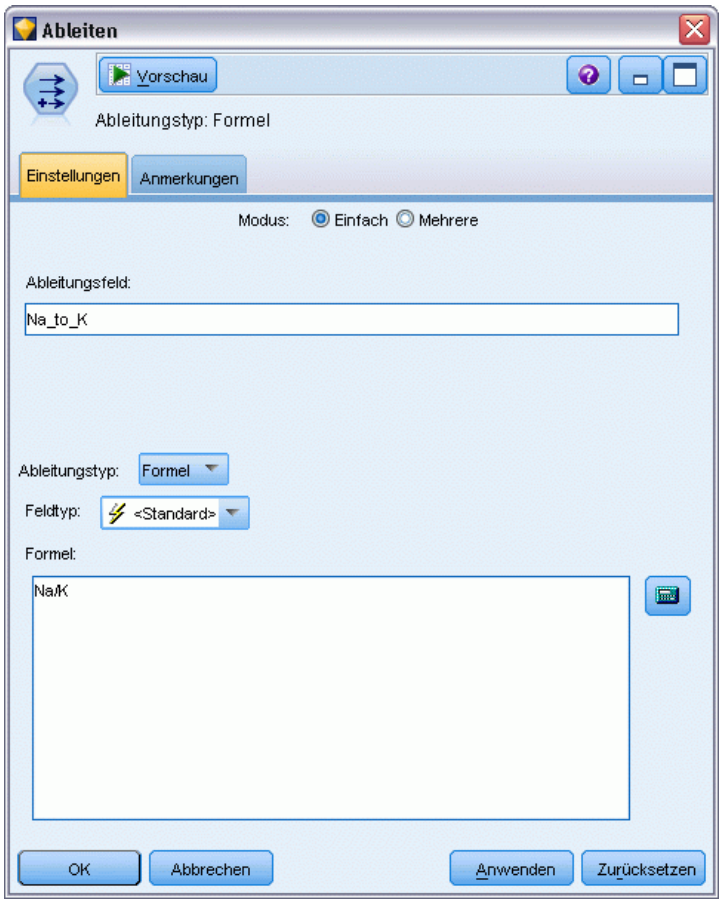
Abbildung 9-13
Netzdiagramm, Medikament Y und alle zugehörigen Verbindungen sind ausgeblendet



Ableiten neuer Felder

Da das Verhältnis von Natrium zu Kalium eine Vorhersage zu ermöglichen scheint, wann Medikament *Y* zu verwenden ist, können Sie ein Feld ableiten, das für jeden Datensatz den Wert dieses Verhältnisses enthält. Dieses Feld kann für später nützlich sein, wenn Sie ein Modell erstellen, um vorherzusagen, in welchen Fällen jedes der fünf Medikamente eingesetzt werden soll. Um das Stream-Layout zu vereinfachen, beginnen Sie damit, dass Sie alle Knoten mit Ausnahme des Quellenknotens DRUG1n löschen. Hängen Sie einen Ableitungsknoten (Registerkarte “Feldoperationen”) an DRUG1n an und doppelklicken Sie dann auf den Ableitungsknoten, um ihn zu bearbeiten.

Abbildung 9-14
Bearbeiten des Ableitungsknotens



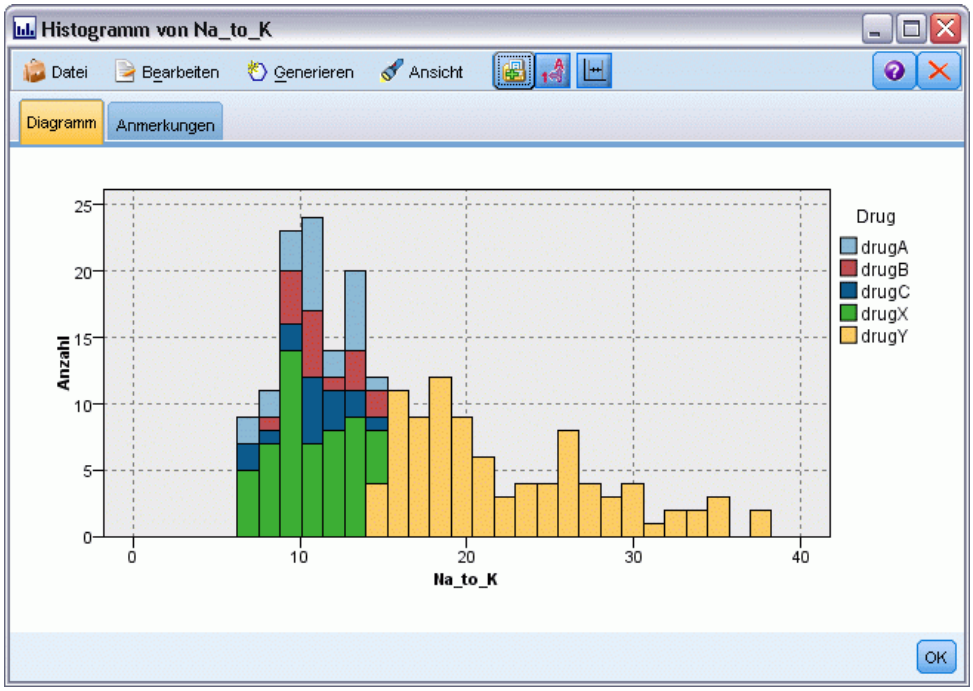
Benennen Sie das neue Feld *Na_zu_K*. Da sich das neue Feld durch Dividieren des Natriumwerts durch den Kaliumwert ergibt, geben Sie für die Formel Na/K ein. Sie können eine Formel auch durch Klicken auf das Symbol gleich rechts neben dem Feld erstellen. Dadurch wird der Expression Builder geöffnet, in dem Sie Ausdrücke mithilfe von integrierten Funktionslisten und Operanden sowie mit Feldern und deren Werten interaktiv erstellen können.

Sie können die Verteilung des neuen Felds überprüfen, indem Sie an den Ableitungsknoten einen Histogrammknoten anfügen. Geben Sie im Histogrammknoten *Na_zu_K* als darzustellendes Feld und *Drug* (Medikament) als Überlagerungsfeld an.

Abbildung 9-15
Bearbeiten des Histogrammknotens

Wenn Sie den Stream ausführen, erhalten Sie das hier gezeigte Diagramm. Die Diagrammdarstellung ermöglicht die Schlussfolgerung, dass bei einem *Na_zu_K*-Wert von 15 und darüber Medikament *Y* das Medikament der Wahl ist.

Abbildung 9-16
Histogrammanzeige

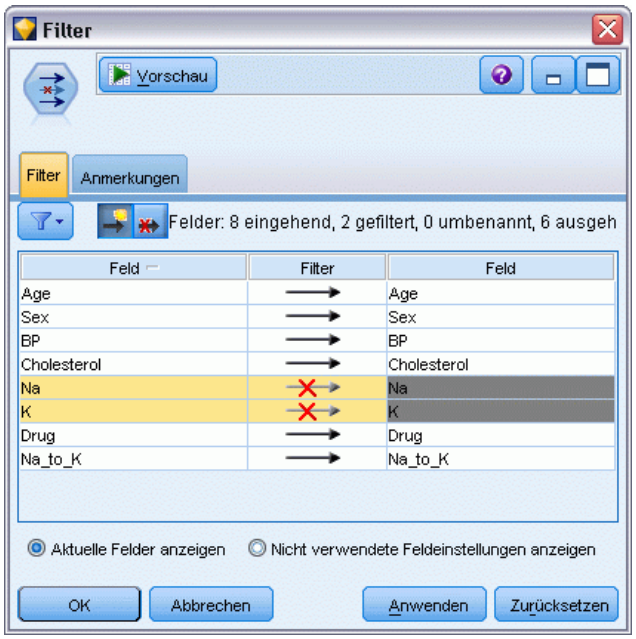


Erstellen eines Modells

Durch Untersuchen und Manipulieren der Daten konnten Sie bereits einige Hypothesen aufstellen. Das Verhältnis der Natriumkonzentration zur Kaliumkonzentration im Blut scheint, wie auch der Blutdruck, einen Einfluss auf die Wahl des Medikaments zu haben. Sie sind jedoch noch nicht in der Lage, alle Beziehungen vollständig zu erklären. Hier liefert die Modellierung wahrscheinlich einige Antworten. In diesem Fall versuchen Sie die Daten mithilfe von C5.0, einem Regel bildenden Modell, anzupassen.

Da Sie ein abgeleitetes Feld, *Na_zu_K*, verwenden, können Sie die ursprünglichen Felder *Na* und *K* ausfiltern, damit sie im Modellierungsalgorithmus nicht zweimal verwendet werden. Dies können Sie mithilfe eines Filterknotens durchführen.

Abbildung 9-17
Bearbeiten des Filterknotens

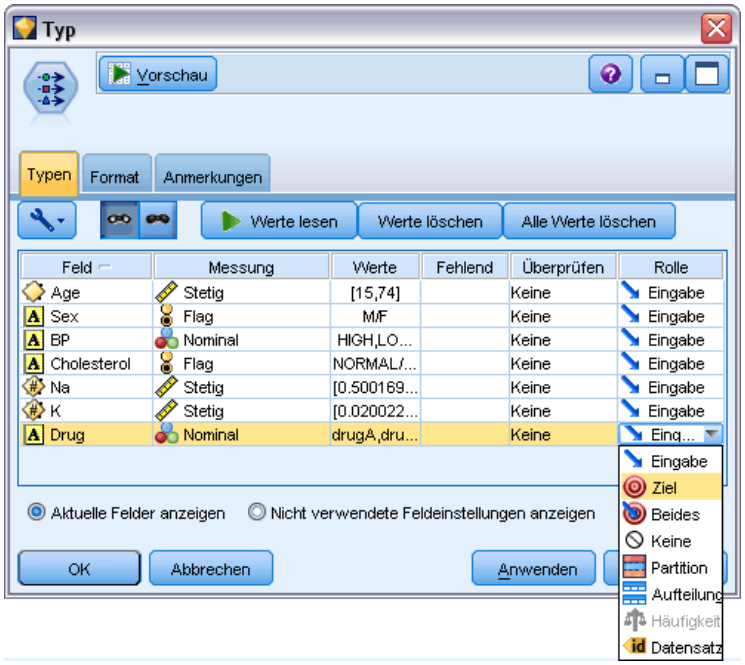


Klicken Sie auf der Registerkarte “Filter” auf die Pfeile neben *Na* und *K*. Über jedem Pfeil erscheint ein rotes X, das anzeigt, dass die Felder nun ausgefiltert sind.

Fügen Sie als Nächstes einen Typknoten an, der mit dem Filterknoten verbunden ist. Der Typknoten ermöglicht Ihnen, anzugeben, welche Feldtypen Sie verwenden und wie sie zur Vorhersage der Ergebnisse verwendet werden.

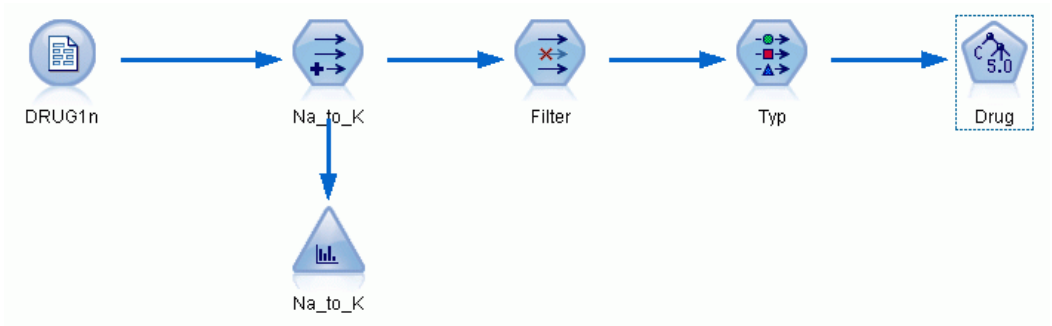
Setzen Sie auf der Registerkarte “Typen” die Rolle für das Feld *Drug* (Medikament) auf Ziel, um anzugeben, dass *Drug* (Medikament) das Feld ist, das vorhergesagt werden soll. Belassen Sie die Rolle für die anderen Felder bei Eingabe.

Abbildung 9-18
Bearbeiten des Typknotens



Fügen Sie zur Abschätzung des Modells einen C5.0-Knoten in den Arbeitsbereich ein und fügen Sie ihn, wie in der Abbildung gezeigt, an das Ende des Streams an. Klicken Sie dann auf die grüne Schaltfläche Ausführen, um den Stream auszuführen.

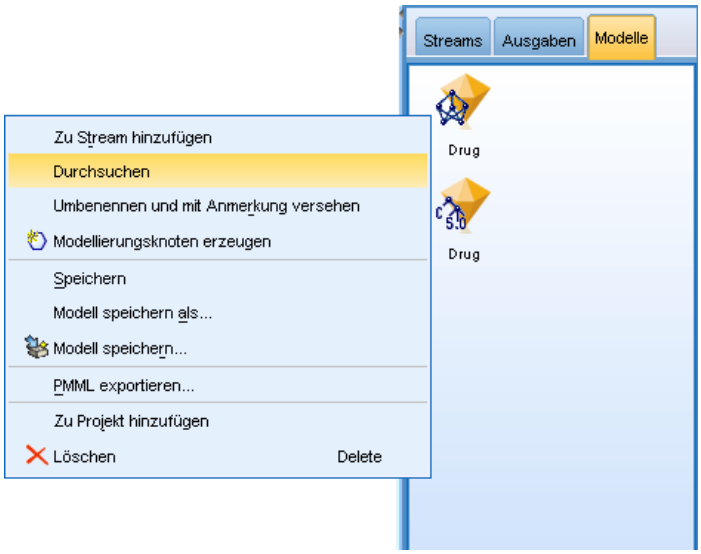
Abbildung 9-19
Hinzufügen eines C5.0-Knotens



Durchsuchen des Modells

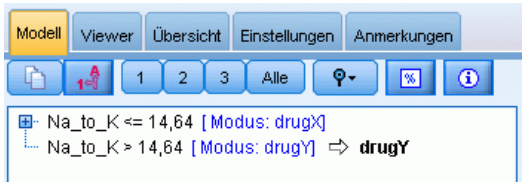
Wenn der Knoten C5.0 ausgeführt wird, wird das generierte Modell-Nugget zum Stream und zur Modellpalette in der rechten oberen Fensterecke hinzugefügt. Um das Modell zu durchsuchen, klicken Sie mit der rechten Maustaste auf eines der Symbole und wählen Sie Bearbeiten oder Durchsuchen aus dem Kontextmenü.

Abbildung 9-20
Durchsuchen des Modells



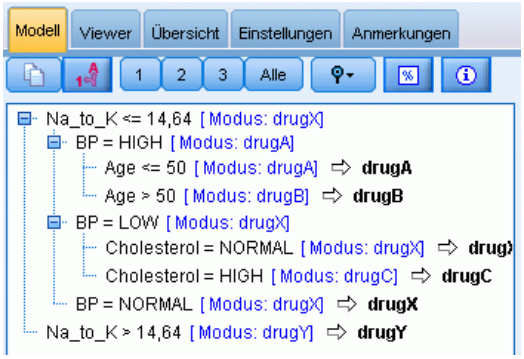
Der Regelbrowser zeigt die vom C5.0-Knoten generierten Regeln in einem Entscheidungsbaumformat an. Der Entscheidungsbaum ist zunächst noch reduziert. Um ihn zu erweitern und alle Ebenen anzuzeigen, klicken Sie auf die Schaltfläche Alle.

Abbildung 9-21
Regelbrowser



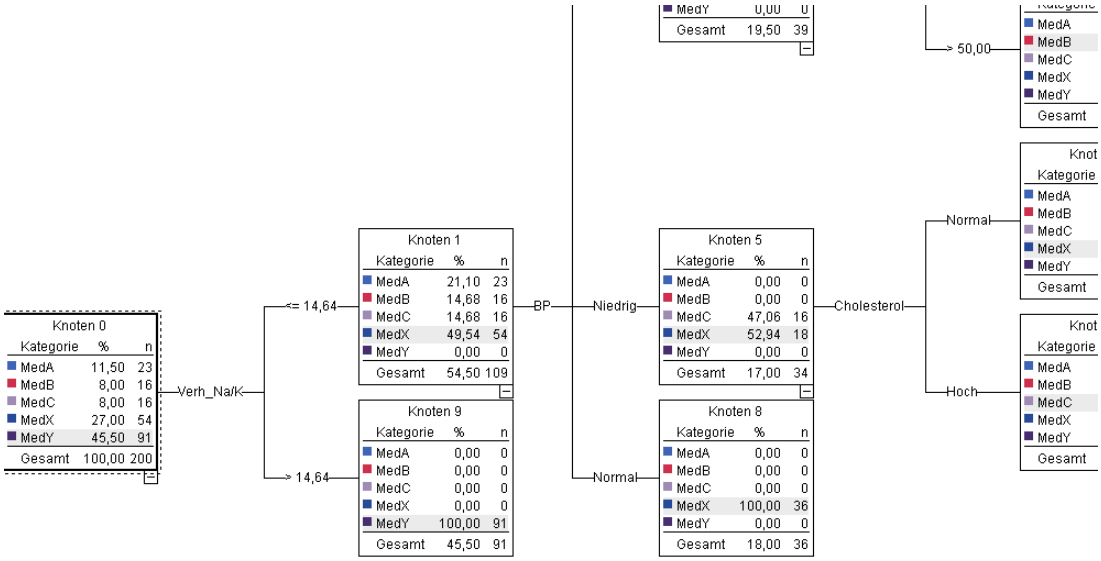
Nun können Sie die fehlenden Teile des Puzzles sehen. Bei Personen mit einem *Na*-zu-*K*-Verhältnis von unter 14,64 und hohem Blutdruck bestimmt das Alter die Wahl des Medikaments. Bei Personen mit niedrigem Blutdruck scheint der Cholesterinspiegel der beste Prädiktor zu sein.

Abbildung 9-22
Vollständig erweiterter Regelbrowser



Der gleiche Entscheidungsbaum kann in einem anspruchsvolleren grafischen Format angezeigt werden, indem Sie auf die Registerkarte Viewer klicken. Hier können Sie deutlicher die Anzahl der Fälle für jede Blutdruckkategorie sowie den Prozentsatz der einzelnen Fälle sehen.

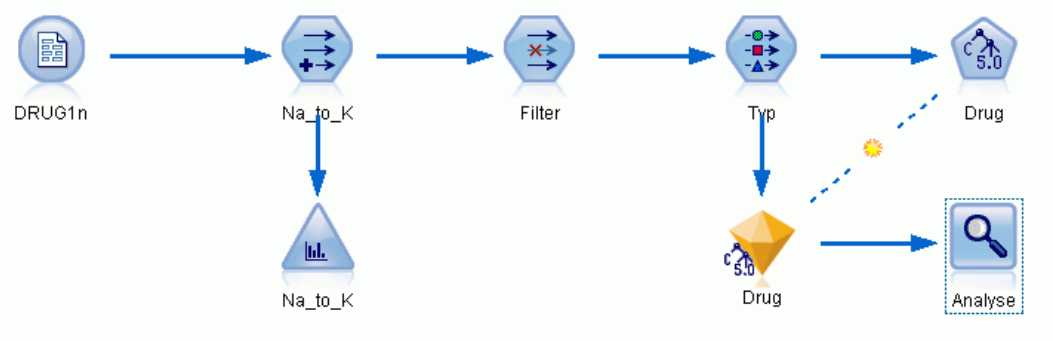
Abbildung 9-23
Entscheidungsbaum in grafischem Format



Verwenden eines Analyseknotens

Sie können die Genauigkeit des Modells mithilfe eines Analyseknotens bewerten. Verbinden Sie einen Analyseknoten (aus der Ausgabeknoten-Palette) mit dem Modell-Nugget, öffnen Sie den Analyseknoten und klicken Sie auf Ausführen.

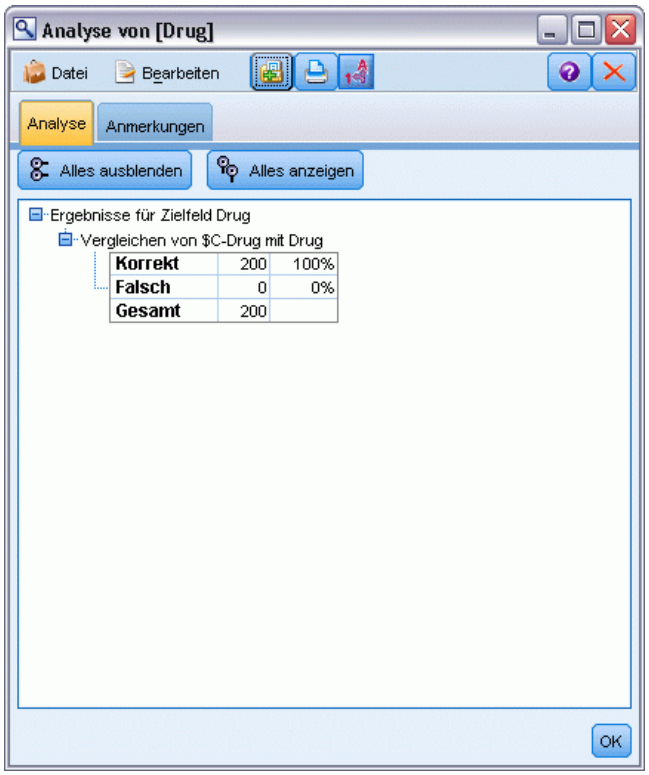
Abbildung 9-24
Hinzufügen eines Analyseknotens



Die Ausgabe des Analyseknotens zeigt, dass das Modell bei diesem künstlichen Daten-Set die Wahl des Medikaments für jeden Datensatz im Daten-Set korrekt vorhergesagt hat. Mit einem realen Daten-Set werden Sie kaum eine 100%ige Genauigkeit erreichen, Sie können jedoch

mithilfe des Analyseknotens bestimmen, ob das Modell für Ihre spezielle Anwendung über eine ausreichende Genauigkeit verfügt.

Abbildung 9-25
Analyseknoten – Ausgabe



Kapitel
10

Screening von Prädiktoren (Merkmalsauswahl)

Mit dem Merkmalsauswahlknoten können Sie die Felder identifizieren, denen bei der Vorhersage eines bestimmten Ergebnisses die größte Bedeutung zukommt. Aus einem Set von Hunderten oder sogar Tausenden von Prädiktoren führt der Merkmalsauswahlknoten ein Screening, eine Rangeinordnung und eine Auswahl der Prädiktoren durch, die voraussichtlich am wichtigsten sind. Letztlich können Sie so ein schnelleres und effizienteres Modell erreichen, ein Modell, das weniger Prediktoren verwendet, schneller ausgeführt werden kann und leichter verständlich ist.

Bei den in diesem Beispiel verwendeten Daten handelt es sich um ein Data Warehouse für eine hypothetische Telefongesellschaft. Sie enthalten Informationen zu Reaktionen auf eine spezielle Werbeaktion, 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 "Ziel"-Felder 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 auf ähnliche Angebote in Zukunft reagieren.

In diesem Beispiel wird ein Stream namens *featureselection.str* verwendet, der Bezug nimmt auf die Datendatei *customer_dbase.sav*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *featureselection.str* befindet sich im Verzeichnis *streams*.

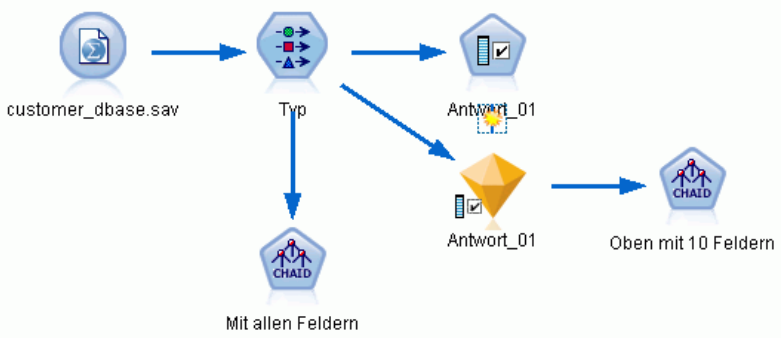
Dieses Beispiel konzentriert sich auf nur eines der Angebote als Ziel. Mithilfe des CHAID-Baumerstellungsknotens wird ein Modell entwickelt, das beschreibt, welche Kunden mit der größten Wahrscheinlichkeit auf die Werbeaktion reagieren. Es werden zwei Ansätze gegenübergestellt:

- Ohne Merkmalsauswahl. Alle Prädiktorfelder im Daten-Set dienen als Eingaben für den CHAID-Baum.
- Mit Merkmalsauswahl. Der Merkmalsauswahlknoten dient zur Auswahl der besten 10 Prädiktoren. Diese werden dann als Eingabe für den CHAID-Baum verwendet.

Wenn wir die zwei resultierenden Baummodelle vergleichen, sehen wir die effektiven Ergebnisse, die mithilfe der Merkmalsauswahl erzielt werden können.

Erstellen des Streams

Abbildung 10-1
Beispiel-Stream für die Merkmalsauswahl



- Platzieren Sie einen Statistikdatei-Quellenknoten in einem leeren Stream-Zeichenbereich. Richten Sie diesen Knoten auf die Beispieldatendatei *customer_dbase.sav*, die im Verzeichnis *Demos* des IBM® SPSS® Modeler-Installationsordners verfügbar ist. (Alternativ können Sie die Beispiel-Stream-Datei *featureselection.str* im Verzeichnis *streams* öffnen.)

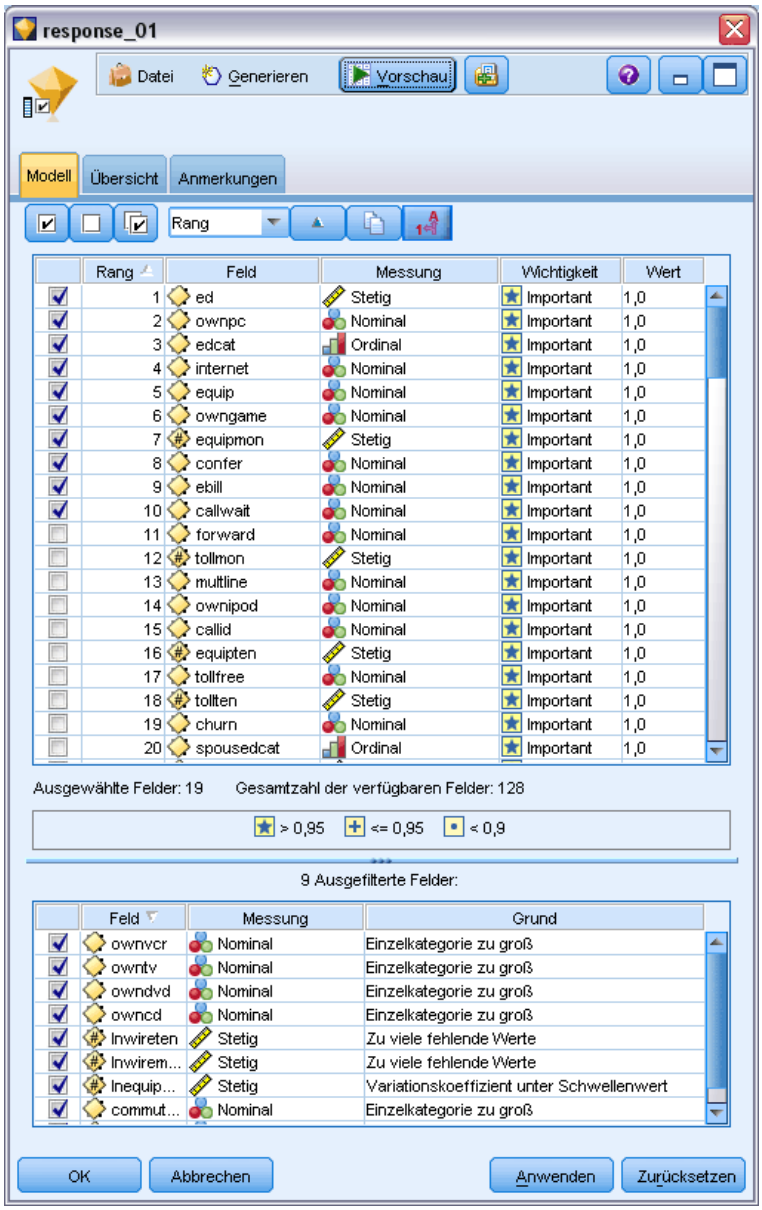
- Fügen Sie einen Typknoten hinzu. Führen Sie auf der Registerkarte "Typen" einen Bildlauf nach ganz unten durch und ändern Sie die Rolle für *response_01* in *Ziel*. Ändern Sie die Rolle für die anderen Antwortfelder (*response_02* und *response_03*) sowie für die Kunden-ID (*custid*) am Beginn der Liste in *Keine*. Lassen Sie die Rolle für alle anderen Felder auf *Eingabe* gesetzt, klicken Sie auf die Schaltfläche Werte lesen und anschließend auf OK.

- Fügen Sie einen Merkmalsauswahl-Modellierungsknoten zum Stream hinzu. An diesem Knoten können Sie die Regeln und Kriterien für das Screening bzw. Ausschließen von Feldern angeben.

- Führen Sie den Stream aus, um das Modell-Nugget vom Typ "Merkmalsauswahl" zu erstellen.

- Klicken Sie mit der rechten Maustaste auf das Modell-Nugget im Stream oder in der Modellpalette und wählen Sie Bearbeiten oder Durchsuchen, um die Ergebnisse zu betrachten.

Abbildung 10-2
Registerkarte "Modell" im Modell-Nugget "Merkmalsauswahl"

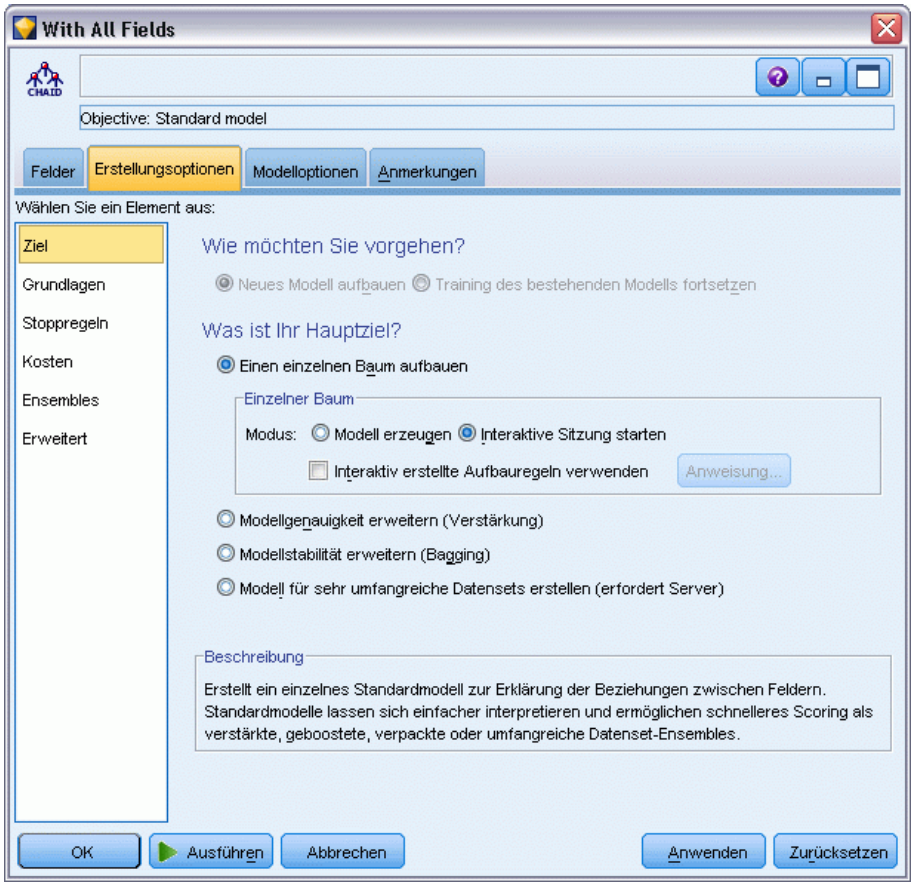


Im oberen Fensterbereich werden die Felder angezeigt, die für die Vorhersage als nützlich erachtet werden. Diese sind nach Wichtigkeit angeordnet. Im unteren Fensterbereich wird angegeben, welche Felder beim Screening aus der Analyse entfernt wurden und warum. Wenn Sie die Felder im oberen Fensterbereich untersuchen, können Sie entscheiden, welche davon in den späteren Modellierungssitzungen verwendet werden sollen.

- Jetzt können die Felder ausgewählt werden, die weiter unten im Stream verwendet werden sollen. Ursprünglich wurden 34 Felder als bedeutsam identifiziert. Wir möchten jedoch das Set an Prädiktoren weiter verkleinern.
- Wählen Sie mithilfe der Markierungen in der ersten Spalte nur die obersten 10 Prädiktoren aus, um die Auswahl der nicht gewünschten Prädiktoren aufzuheben. (Klicken Sie auf das Häkchen in Zeile 11, halten Sie die Umschalttaste gedrückt und klicken Sie auf das Häkchen in Zeile 34.) Schließen Sie das Modell-Nugget.
- Um Ergebnisse ohne Merkmalsauswahl zu vergleichen, müssen Sie zwei CHAID-Modellierungsknoten in den Stream aufnehmen: einen, bei dem die Merkmalsauswahl verwendet wird, und einen bei dem auf Merkmalsauswahl verzichtet wird.
- Verbinden Sie einen CHAID-Knoten mit dem Typknoten und den anderen mit dem generierten Merkmalsauswahlmodell-Nugget.
- Öffnen Sie jeden CHAID-Knoten, klicken Sie auf die Registerkarte “Erstellungsoptionen” und vergewissern Sie sich, dass die Optionen Neues Modell aufbauen, Einzelnen Baum aufbauen und Interaktive Sitzung starten im Fensterbereich “Ziele” ausgewählt sind.

  Vergewissern Sie sich, dass die Maximale Baumtiefe im Fensterbereich “Basis” auf 5 gesetzt ist.

Abbildung 10-3
Zieleinstellungen für den CHAID-Modellierungsknoten für alle Prädiktor-Felder

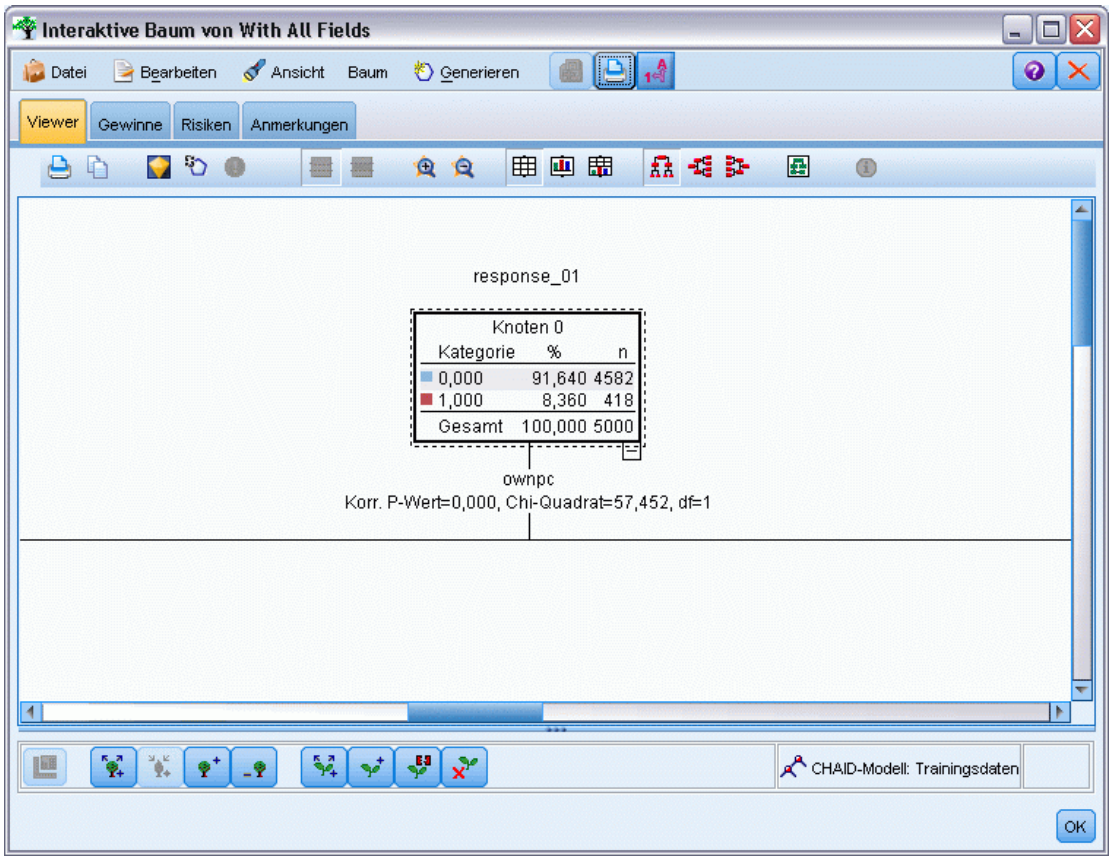


Erstellen der Modelle

- Führen Sie den CHAID-Knoten aus, der alle Prädiktoren im Daten-Set verwendet (den Knoten, der mit dem Typknoten verbunden ist). Achten Sie darauf, wie lange die Ausführung dauert. Im Ergebnisfenster wird eine Tabelle angezeigt.

- Wählen Sie in den Menüs die Optionsfolge Baum > Baum erweitern aus, um den Baum zu erweitern und anzuzeigen.

Abbildung 10-4
Erweitern des Baums im Tree Builder



- Führen Sie nun dieselben Schritte mit dem anderen CHAID-Knoten aus, der nur 10 Prädiktoren verwendet. Erweitern Sie auch hier den Baum, wenn der Tree Builder geöffnet wird.

  Das zweite Modell müsste schneller ausgeführt worden sein als das erste. Da dieses Daten-Set recht klein ist, beträgt der Unterschied in der Ausführungsdauer vermutlich nur wenige Sekunden. Bei größeren, realen Daten-Sets kann der Unterschied jedoch beträchtlich sein – es kann sich um Minuten oder sogar Stunden handeln. Mithilfe der Merkmalsauswahl können Sie die Verarbeitungsgeschwindigkeit drastisch erhöhen.

Der zweite Baum enthält außerdem weniger Baumknoten als der erste. Er ist leichter verständlich. Doch bevor Sie sich entschließen, ihn zu verwenden, sollten Sie herausfinden, ob er effektiv ist und wie er im Vergleich mit dem Modell abschneidet, bei dem alle Prädiktoren verwendet werden.

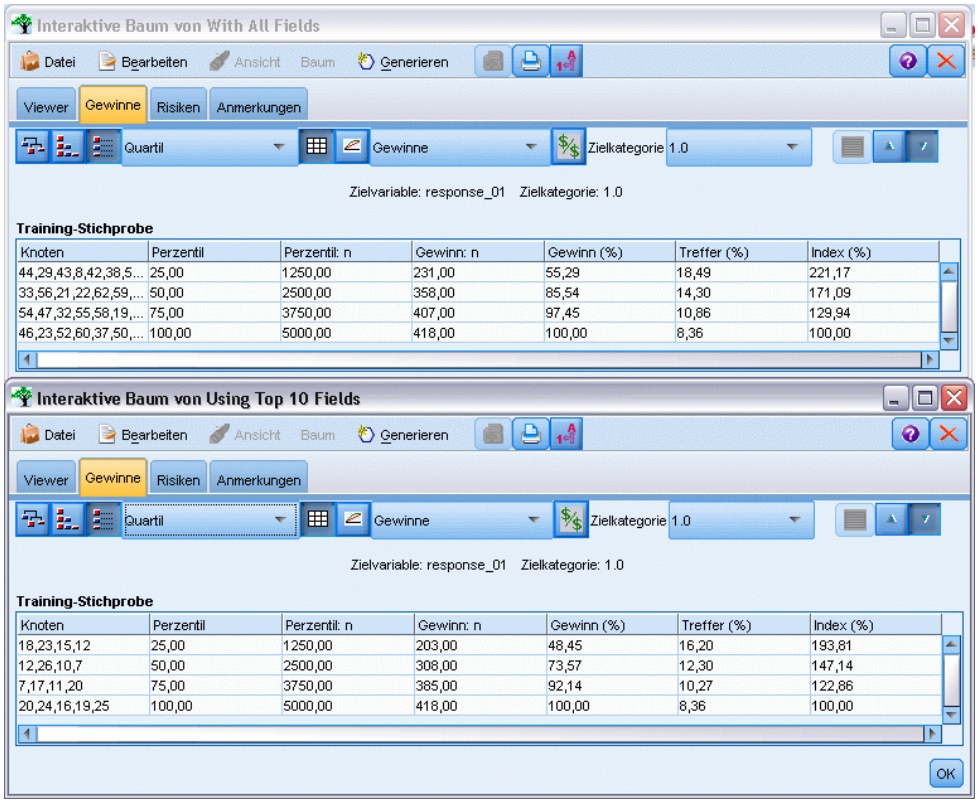
Vergleichen der Ergebnisse

Um die beiden Ergebnisse zu vergleichen, benötigen wir ein Maß für die Effektivität. Dazu verwenden wir die Registerkarte “Gewinne” im Tree Builder. Wir untersuchen den **Lift**, der misst, mit um wie viel höherer Wahrscheinlichkeit die Datensätze in einem Knoten in die Zielkategorie fallen im Vergleich zu allen Datensätzen im Daten-Set. Ein Lift-Wert von 148 % beispielsweise gibt an, dass die Datensätze im Knoten mit 1,48-mal höherer Wahrscheinlichkeit in die Zielkategorie fallen als alle Datensätze im Daten-Set. Der Lift wird auf der Registerkarte “Gewinne” in der Spalte *Index* angegeben.

- Klicken Sie im Tree Builder für das vollständige Prädiktoren-Set auf die Registerkarte “Gewinne”. Ändern Sie die Zielkategorie in 1,0. Ändern Sie die Anzeige in Quartile. Klicken Sie dazu zuerst auf die Symbolleisten-Schaltfläche “Quantile”. Wählen Sie anschließend im Dropdown-Menü neben dieser Schaltfläche die Option Quartil aus.

- Wiederholen Sie diesen Vorgang im Tree Builder für das Set mit 10 Prädiktoren, sodass Sie zwei ähnliche Gewinntabellen zum Vergleich erhalten, wie in den folgenden Abbildungen dargestellt.

Abbildung 10-5
Gewinntabellen für beide CHAID-Modelle



Jede Gewinntabelle gruppiert die Endknoten für den zugehörigen Baum in Quartile Um die Effektivität der beiden Modelle zu vergleichen, betrachten wir den Lift (*Index*-Wert) des obersten Quartils in jeder Tabelle.

Wenn alle Prädiktoren enthalten sind, zeigt das Modell einen Lift von 221%. Fälle mit den Merkmalen in diesen Knoten reagieren also mit 2,2mal höherer Wahrscheinlichkeit auf die Ziel-Werbeaktion. Um zu sehen, um welche Merkmale es sich dabei handelt, wählen Sie die oberste Zeile durch Klicken aus. Wechseln Sie dann zur Registerkarte “Viewer”, auf der die entsprechenden Knoten nun schwarz hervorgehoben sind. Gehen Sie den Baum hinunter zu den einzelnen hervorgehobenen Endknoten, um zu sehen, wie die Prädiktoren aufgeteilt wurden. Allein das oberste Quartile enthält 10 Knoten. Wenn wir diese auf reale Scoring-Modelle übertragen, können 10 verschiedene Kundenprofile recht schwer zu verwalten sein.

Wenn nur die besten 10 Prädiktoren (durch die Merkmalsauswahl ermittelt) enthalten sind, beträgt der Lift fast 194 %. Dieses Modell ist zwar nicht ganz so gut wie das Modell mit allen Prädiktoren, aber es ist definitiv brauchbar. Hier sind im obersten Quartile nur vier Knoten enthalten. Es ist also einfacher. Daher können wir davon ausgehen, dass das Merkmalsauswahlmodell dem Modell mit allen Prädiktoren vorzuziehen ist.

Zusammenfassung

Fassen wir die Vorteile der Merkmalsauswahl noch einmal zusammen. Durch die Verwendung weniger Prädiktoren wird der Aufwand verringert. Dies bedeutet, dass Sie weniger Daten sammeln, verarbeiten und in die Modelle einspeisen müssen. Die Berechnungszeit wird reduziert. In diesem Beispiel war trotz des zusätzlichen Merkmalsauswahlschritts die Modellerstellung mit dem kleineren Prädiktoren-Set merklich schneller. Mit einem größeren realen Daten-Set sollten die Zeiteinsparungen noch erheblich deutlicher ausfallen.

Durch die Verwendung weniger Prädiktoren wird das Scoring vereinfacht. Wie das Beispiel zeigt, könnten eventuell nur vier Profile von Kunden ermittelt werden, die mit hoher Wahrscheinlichkeit auf die Werbeaktion ansprechen. Beachten Sie, dass bei einer größeren Anzahl an Prädiktoren das Risiko einer Überanpassung des Modells besteht. Das einfachere Modell lässt sich möglicherweise besser auf andere Daten-Sets verallgemeinern (allerdings müsste dies sicherheitshalber getestet werden).

Sie könnten einen Baumerstellungsalgorithmus für die Merkmalsauswahl verwenden, sodass der Baum die wichtigsten Prädiktoren automatisch ermittelt. Tatsächlich wird der CHAID-Algorithmus häufig zu diesem Zweck verwendet. Es ist sogar möglich, den Baum Ebene für Ebene zu erweitern, um seine Tiefe und Komplexität steuern zu können. Der Merkmalsauswahlknoten ist jedoch schneller und benutzerfreundlicher. Er erstellt eine Rangordnung aller Prädiktoren in einem einzigen schnellen Schritt, sodass Sie schnell die wichtigsten Felder ermitteln können. Außerdem können Sie damit angeben, wie viele Prädiktoren aufgenommen werden sollen. Sie können dieses Beispiel problemlos erneut ausführen und statt der 10 wichtigsten Prädiktoren die 15 oder 20 wichtigsten Prädiktoren verwenden. Anschließend können Sie die Ergebnisse vergleichen, um das optimale Modell zu ermitteln.

Kapitel
11

Reduzieren der Länge der Zeichenkette für die Eingabedaten (Umkodierungsknoten)

Reduzieren der Länge der Zeichenkette für die Eingabedaten (Umkodierung)

Bei Modellen vom Typ “Binomiale logistische Regression” und Modellen vom Typ “Automatischer Klassifizierer”, die ein Modell vom Typ “Binomiale logistische Regression” enthalten, sind die Zeichenkettenfelder auf maximal acht Zeichen begrenzt. Zeichenketten mit mehr als acht Zeichen können mithilfe eines Umkodierungsknotens neu kodiert werden.

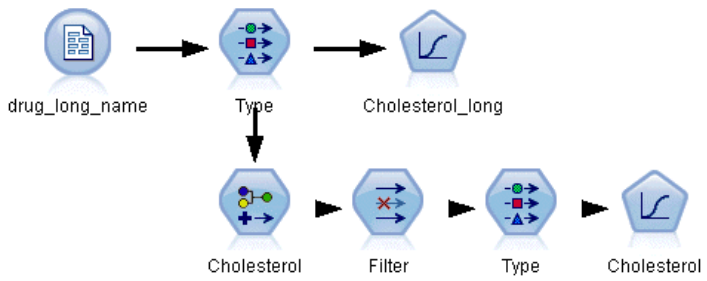
In diesem Beispiel wird ein Stream namens *reclassify_strings.str* verwendet, der Bezug nimmt auf die Datendatei *drug_long_name*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe “IBM® SPSS® Modeler” im Windows-Startmenü aufgerufen werden. Die Datei *reclassify_strings.str* befindet sich im Verzeichnis *streams*.

Dieses Beispiel konzentriert sich auf einen kleinen Teil eines Streams, um zu zeigen, welche Art von Fehlern durch übermäßig lange Zeichenketten entstehen können. Außerdem wird erläutert, wie die Zeichenkettendetails mithilfe des Umkodierungsknotens auf eine akzeptable Länge gekürzt werden können. In diesem Beispiel wird ein Beispiel eines Knotens vom Typ “Binomiale logistische Regression” verwendet, er ist jedoch gleichermaßen gültig, wenn mithilfe des Knotens “Automatischer Klassifizierer” ein Modell vom Typ “Binomiale logistische Regression” erstellt wird.

Umkodieren der Daten

- Stellen Sie mithilfe eines Quellenknotens für variable Dateien eine Verbindung mit dem Daten-Set *drug_long_name* im Ordner *Demos* her.

Abbildung 11-1
Beispiel-Stream zur Umkodierung von Zeichenketten für die binomiale logistische Regression



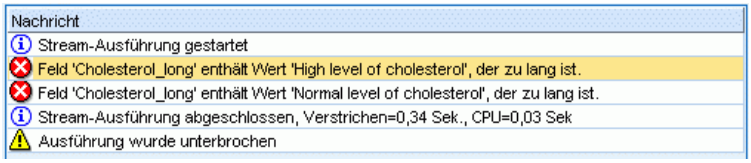
- Fügen Sie einen Typknoten zum Quellenknoten hinzu und wählen Sie Cholesterol_long als Ziel aus.
- Fügen Sie einen Knoten vom Typ “Logistische Regression” zum Typknoten hinzu.
- Klicken Sie im Knoten “Logistische Regression” auf die Registerkarte “Modell” und wählen Sie die Prozedur Binomial aus.

Abbildung 11-2
Lange Zeichenkettendetails im Feld “Cholesterol_long”

- Wenn Sie den Knoten “Logistische Regression” in *reclassify_strings.str* ausführen, wird eine Fehlermeldung angezeigt, die Sie darauf hinweist, dass die Werte der Zeichenkette Cholesterol_long zu lang sind.

  Wenn diese Art von Fehlermeldung auftritt, sollten Sie Ihre Daten mithilfe des im Folgenden in diesem Beispiel erläuterten Verfahrens ändern.

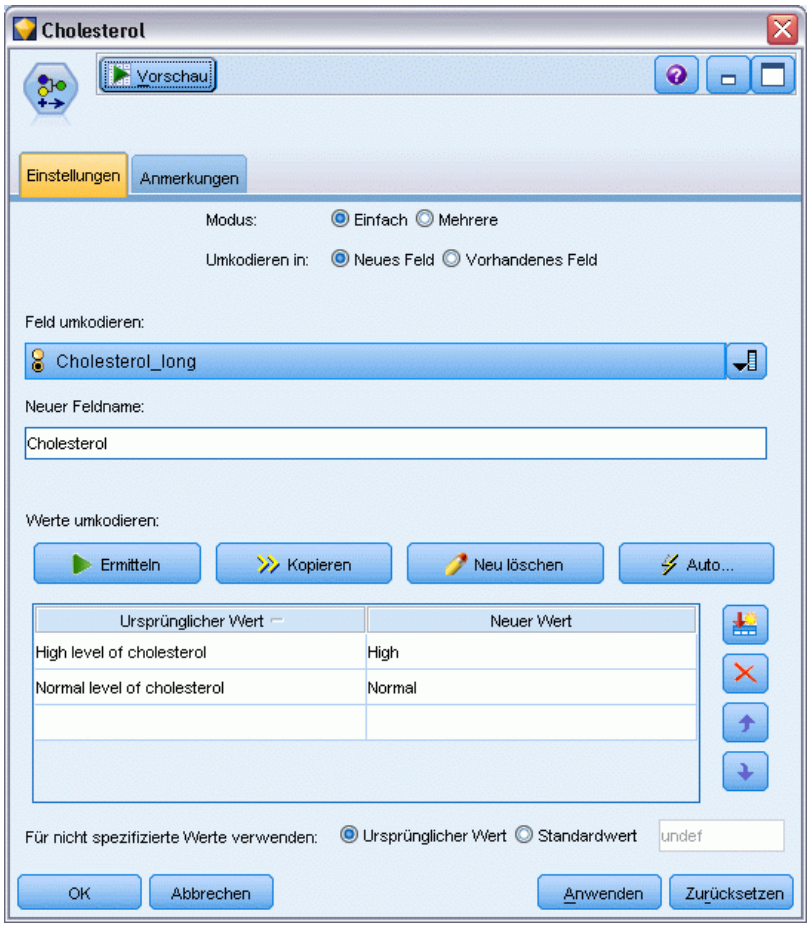
Abbildung 11-3
Fehlermeldung bei Ausführung des Knotens “Binomiale logistische Regression”



- Fügen Sie einen Umkodierungsknoten zum Typknoten hinzu.
- Wählen Sie im Feld “Umkodieren” den Eintrag Cholesterol_long aus.
- Geben sie Cholesterol als neuen Feldnamen ein.
- Klicken Sie auf die Schaltfläche Ermitteln, um die Werte von Cholesterol_long zur Spalte “Ursprünglicher Wert” hinzuzufügen.

- Geben Sie in der Spalte “Neuer Wert” High neben dem ursprünglichen Wert High level of cholesterol und Normal neben dem ursprünglichen Wert Normal level of cholesterol ein.

Abbildung 11-4
Umkodieren der langen Zeichenketten



- Fügen Sie einen Filterknoten zum Umkodierungsknoten hinzu.

▶ Klicken Sie in der Spalte “Filter”, um Cholesterol_long zu entfernen.

Abbildung 11-5
Filtern des Felds “Cholesterol_long” aus den Daten

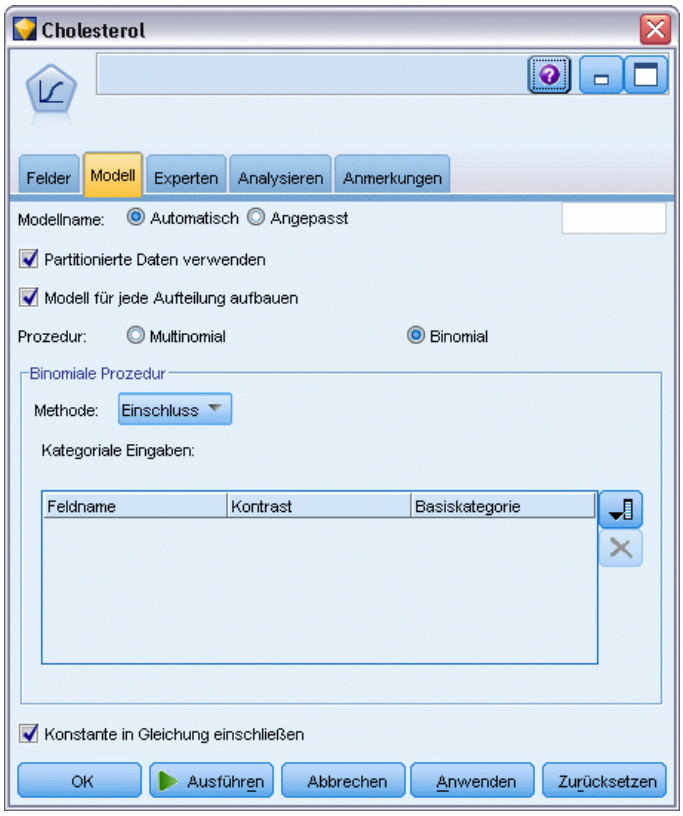
▶ Fügen Sie einen Typknoten zum Filterknoten hinzu und wählen Sie Cholesterol als Ziel aus.

Abbildung 11-6
Kurze Zeichenkettendetails im Feld “Cholesterol”

▶ Fügen Sie einen Logistikknoten zum Typknoten hinzu.

- Klicken Sie im Logistikknoten auf die Registerkarte “Modell” und wählen Sie die Prozedur Binomial aus.
- Sie können nun den binomialen Logistikknoten ausführen und ein Modell generieren, ohne dass eine Fehlermeldung angezeigt wird.

Abbildung 11-7
Auswählen von Binomial als Prozedur



Dieses Beispiel zeigt nur einen Teil eines Streams. Wenn Sie weitere Informationen über die Stream-Typen benötigen, bei denen eine Umkodierung langer Zeichenketten erforderlich sein kann, stehen Ihnen folgende Beispiele zur Verfügung:

- Knoten “Automatischer Klassifizierer” Für weitere Informationen siehe Thema Modellieren der Kundenreaktion (Automatischer Klassifizierer) in Kapitel 5 auf S. 44.
- Knoten “Binomiale logistische Regression”. Für weitere Informationen siehe Thema Kundenabwanderung bei Telekommunikationsunternehmen (binomiale logistische Regression) in Kapitel 14 auf S. 164.

Weitere Informationen zur Verwendung von IBM® SPSS® Modeler, wie beispielsweise das Benutzerhandbuch, die Knotenreferenz und das Algorithmushandbuch, stehen Ihnen im Verzeichnis *\Documentation* des Installationsdatenträgers zur Verfügung.

Teil III: Modellierungsbeispiele

Kapitel
12

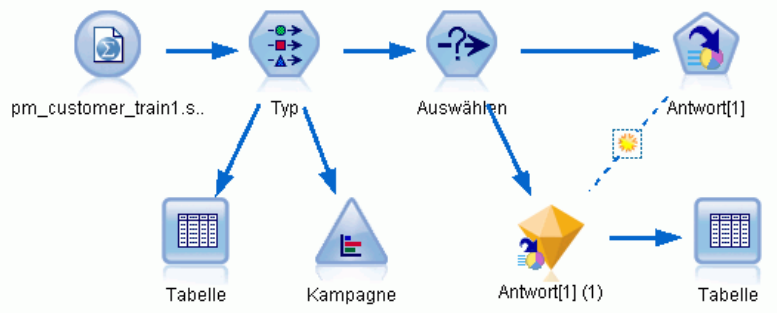
Modellieren der Kundenreaktion (Entscheidungsliste)

Der Entscheidungslisten-Algorithmus generiert Regeln, die eine höhere oder niedrigere Wahrscheinlichkeit eines bestimmten binären Ergebnisses (vom Typ “ja oder nein”) anzeigen. Die Verwendung von Entscheidungslistenmodellen ist im Customer Relationship Management, beispielsweise im Callcenter oder in Marketinganwendungen, weit verbreitet.

Dieses Beispiel beruht auf einem fiktiven Unternehmen, das in zukünftigen Marketingkampagnen profitablere Ergebnisse erzielen möchte, indem jedem Kunden ein speziell für ihn geeignetes Angebot unterbreitet wird. Insbesondere wird in dem Beispiel ein Entscheidungslistenmodell verwendet, mit dem auf der Grundlage früherer Werbeaktionen die Eigenschaften der Kunden ermittelt werden, die mit der größten Wahrscheinlichkeit positiv reagieren werden, und auf der Grundlage der Ergebnisse eine Mailingliste generiert wird.

Entscheidungslistenmodelle eignen sich besonders gut für die interaktive Modellierung, da Sie damit Parameter im Modell anpassen und sofort die Ergebnisse anzeigen können. Ein alternativer Ansatz, mit dem Sie automatisch eine Anzahl verschiedener Modelle erstellen und die Ergebnisse in Ränge einteilen können, kann mithilfe des Knotens “Automatischer Klassifizierer” erstellt werden.

Abbildung 12-1
Beispiel-Stream für Entscheidungsliste

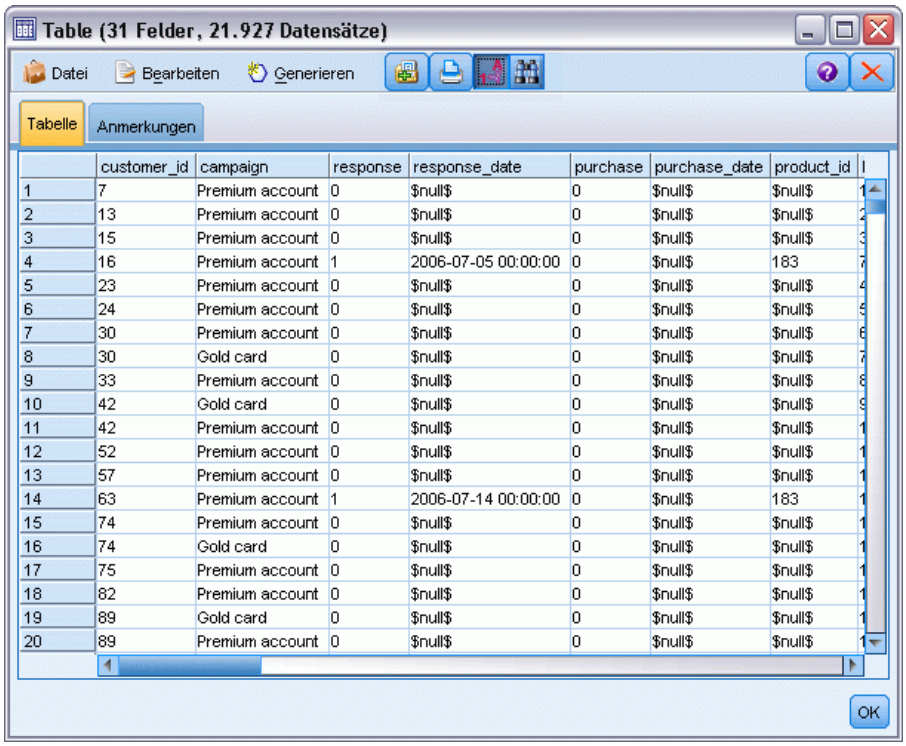


In diesem Beispiel wird ein Stream namens *pm_decisionlist.str* verwendet, der Bezug nimmt auf die Datendatei *pm_customer_train1.sav*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe “IBM® SPSS® Modeler” im Windows-Startmenü aufgerufen werden. Die Datei *pm_decisionlist.str* befindet sich im Verzeichnis *streams*.

Historische Daten

Die Datei *pm_customer_train1.sav* enthält historische Daten, die die Aufzeichnungen über die Angebote enthält, die bestimmten Kunden in früheren Kampagnen unterbreitet wurden, entsprechend dem Wert im Feld *campaign* (Kampagne). Die größte Anzahl an Datensätzen entfallen auf die Kampagne *Premium account* (Premium-Account).

Abbildung 12-2
Daten zu früheren Werbeaktionen



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold card	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold card	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold card	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold card	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

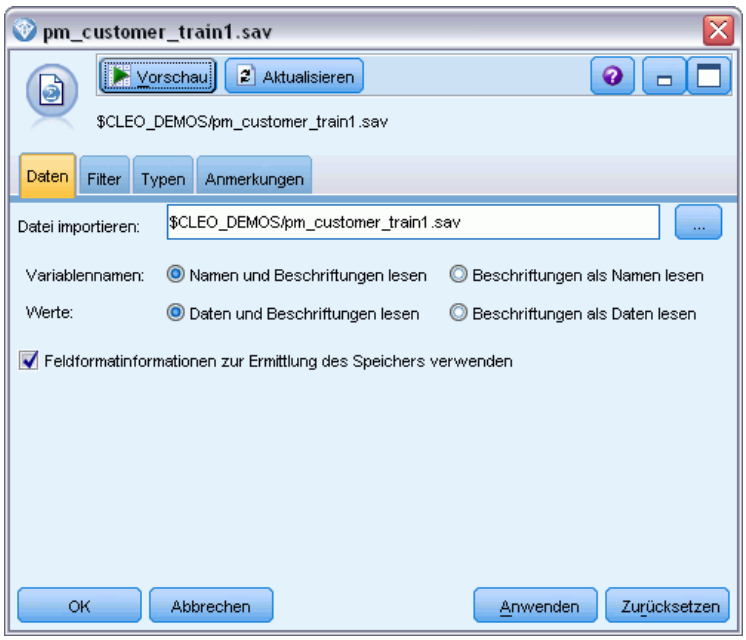
Die Werte des Felds *campaign* (Kampagne) sind in den Daten tatsächlich als ganze Zahlen kodiert. Die Labels sind im Typknoten definiert (Beispiel: *2 = Premium account*). Sie können die Anzeige der Wertelabels in der Tabelle mithilfe der Symbolleiste ein- bzw. ausblenden.

Die Datei enthält außerdem eine Reihe von Feldern mit demografischen Informationen und Finanzdaten zu den einzelnen Kunden, die zum Erstellen bzw. "Trainieren" eines Modells verwendet werden können, das die Antwortquoten für verschiedene Gruppen auf der Grundlage bestimmter Merkmale vorhersagt.

Erstellen des Streams

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf die Datei *pm_customer_train1.sav* im Ordner *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist. (Sie können $CLEO_DEMOS/ im Dateipfad als Verknüpfung zur Referenzierung dieses Ordners angeben.)

Abbildung 12-3
Einlesen der Daten



- Fügen Sie einen Typknoten hinzu und wählen Sie *response* (Antwort) als Zielfeld (Rolle = Ziel) aus. Setzen Sie das Messniveau für dieses Feld auf Flag.

Abbildung 12-4
Festlegen von Messniveau und Rolle

- Setzen Sie die Rolle für folgende Felder auf Keine. *customer_id* (Kunden-ID), *campaign* (Kampagne), *response_date* (Antwortdatum), *purchase* (Einkauf), *purchase_date* (Einkaufsdatum), *product_id* (Produkt-ID), *Rowid* (Zeilen-ID), and *X_random* (X-Zufall). Diese Felder dienen jeweils in den Daten bestimmten Zwecken, werden jedoch nicht zum Erstellen des tatsächlichen Modells verwendet.

- Klicken Sie auf die Schaltfläche Werte lesen im Typknoten, um sicherzustellen, dass die Werte instanziiert werden.

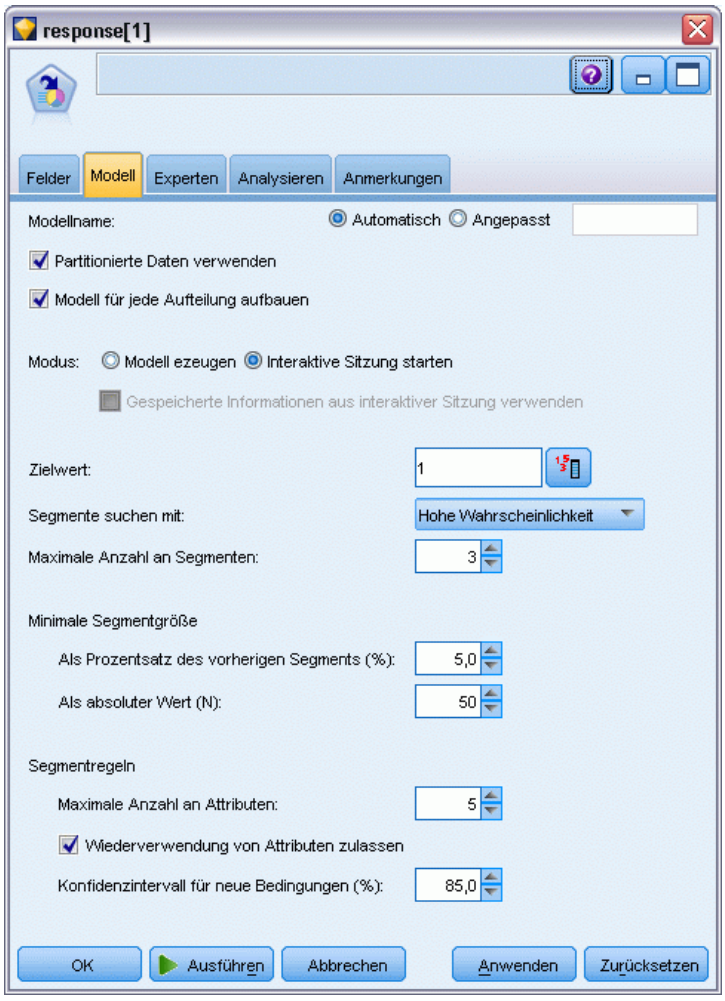
Die Daten enthalten Informationen zu vier verschiedenen Kampagnen, Sie konzentrieren die Analyse jedoch jeweils nur auf eine Kampagne. Da die größte Anzahl an Datensätzen auf die Premium-Kampagne entfällt (in den Daten kodiert als *campaign=2*), können Sie einen Auswahlknoten verwenden, um nur die betreffenden Datensätze in den Stream aufzunehmen.

Abbildung 12-5
Auswählen von Datensätzen für eine einzelne Kampagne

Erstellen des Modells

- Fügen Sie einen Entscheidungslistenknoten zum Stream hinzu. Setzen Sie auf der Registerkarte "Modell" den Zielwert auf 1, um das Ergebnis anzuzeigen, nach dem gesucht werden soll. In diesem Fall suchen Sie nach Kunden, die auf ein früheres Angebot mit *Ja* geantwortet haben.

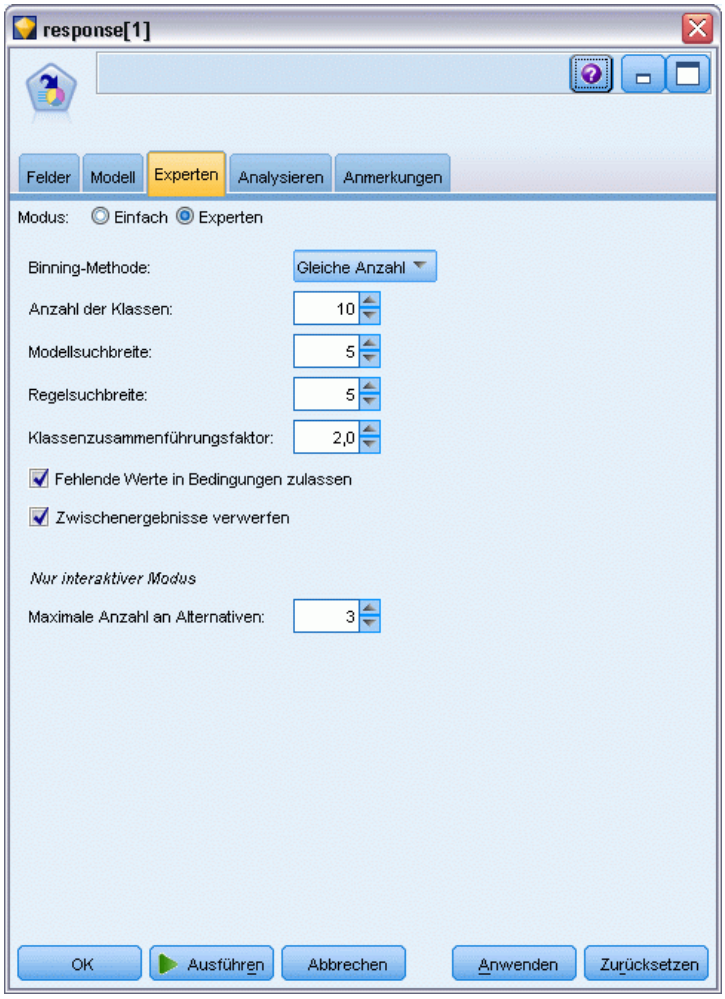
Abbildung 12-6
Entscheidungslistenknoten – Registerkarte "Modell"



- Wählen Sie Interaktive Sitzung starten.
- Um das Modell für dieses Beispiel einfach zu halten, geben Sie als maximale Anzahl an Segmenten den Wert 3 an.
- Ändern Sie das Konfidenzintervall für neue Bedingungen auf 85 %.

▶ Setzen Sie auf der Registerkarte “Experten” den Wert von Modus auf Experten.

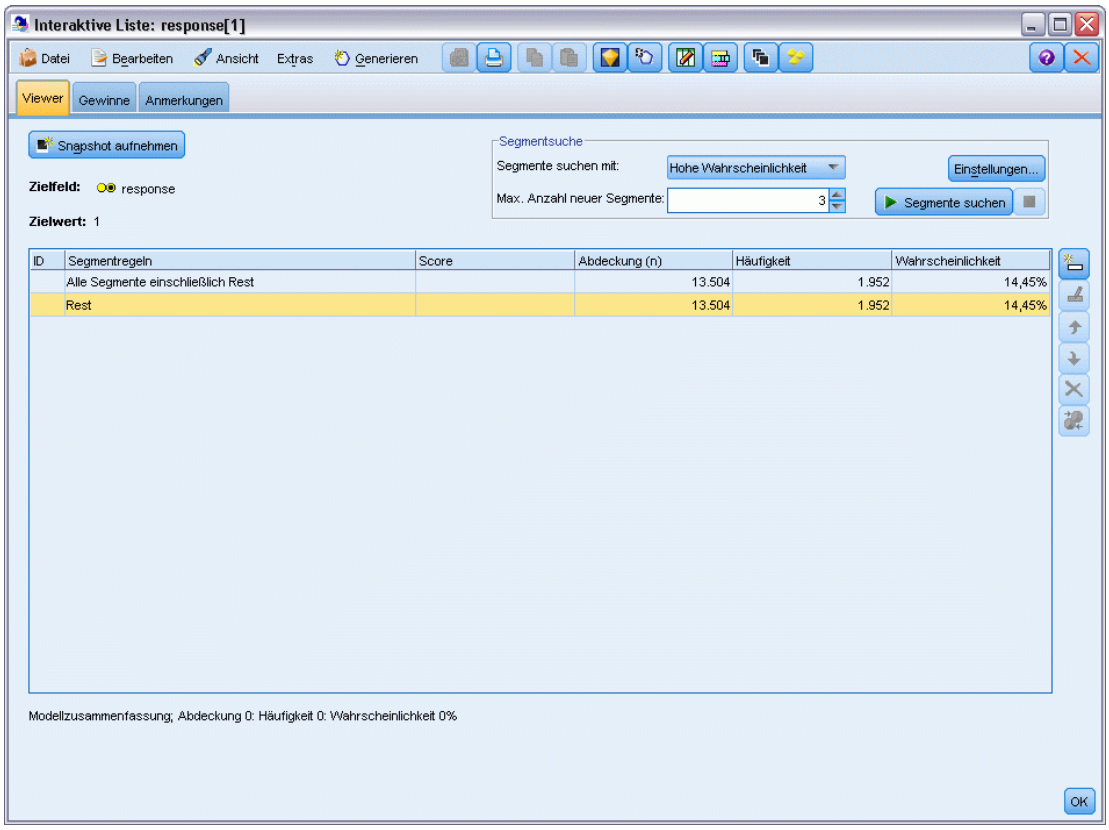
Abbildung 12-7
Entscheidungslistenknoten – Registerkarte “Experten”



▶ Erhöhen Sie die Maximale Anzahl an Alternativen auf 3. Diese Option kann zusammen mit der Einstellung Interaktive Sitzung starten verwendet werden, die Sie auf der Registerkarte “Modell” ausgewählt haben.

▶ Klicken Sie auf Ausführen, um den Viewer “Interaktive Liste” anzuzeigen.

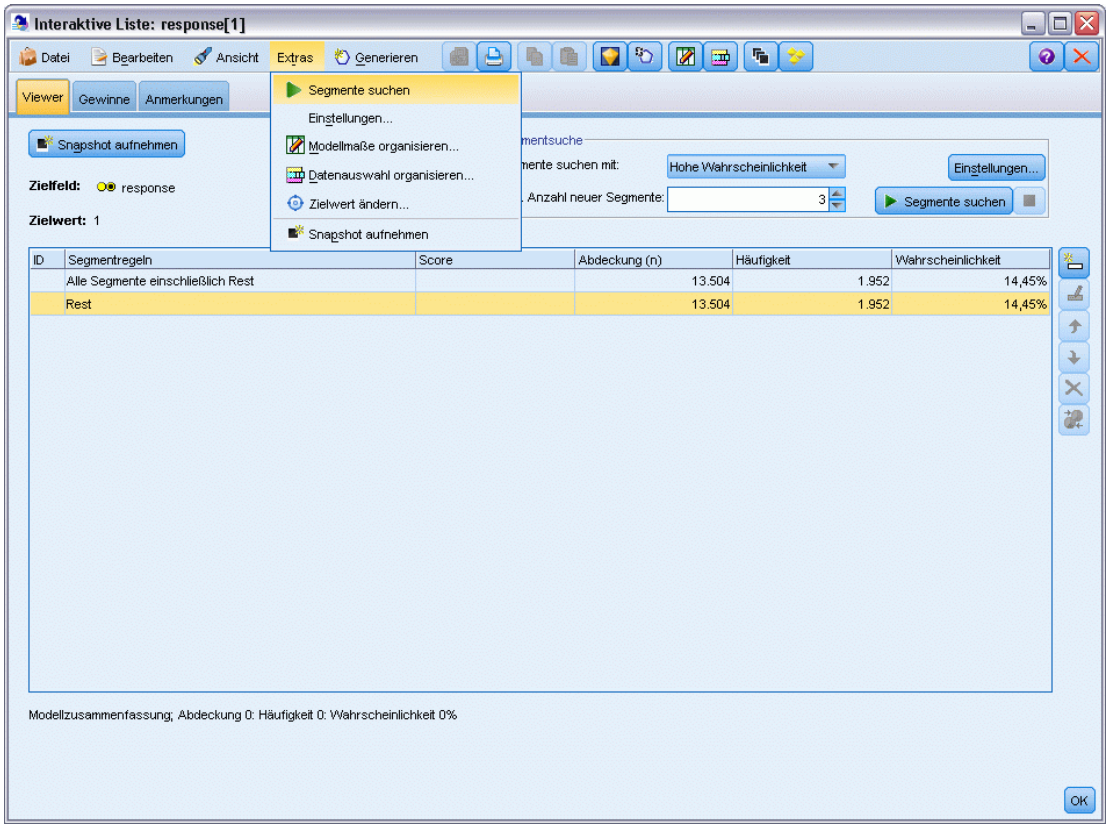
Abbildung 12-8
Viewer "Interaktive Liste"



Da bisher keine Segmente definiert wurden, entfallen alle Datensätze auf den Rest. Von den 13.504 Datensätzen in der Stichprobe weisen 1.952 die Antwort *Ja* auf, was einer Gesamttrefferquote von 14,45 % entspricht. Sie möchten diese Quote verbessern, indem Sie Kundensegmente ermitteln, für die die Wahrscheinlichkeit einer positiven Antwort höher (bzw. niedriger) liegt.

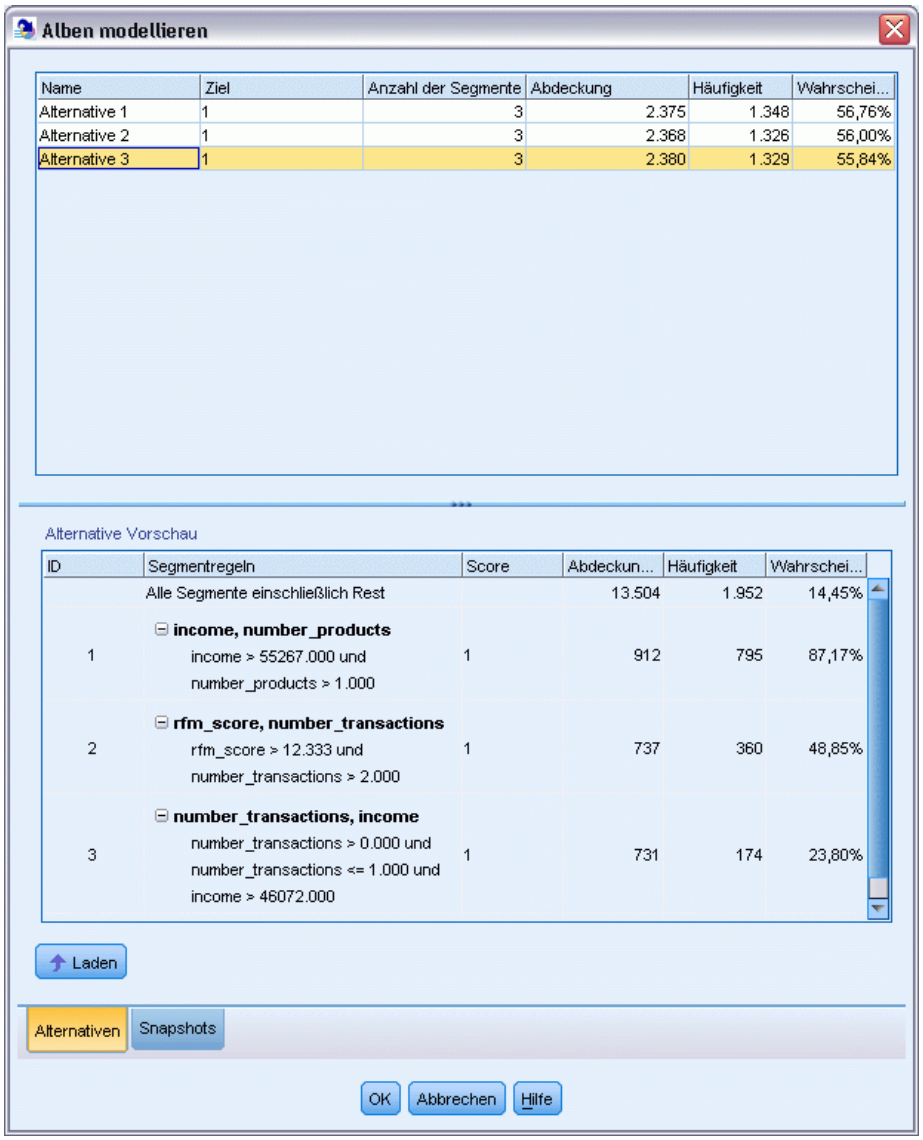
- Wählen Sie im Viewer “Interaktive Liste” folgende Optionsfolge aus den Menüs aus:
  Werkzeuge > Segmente suchen

Abbildung 12-9
Viewer “Interaktive Liste”



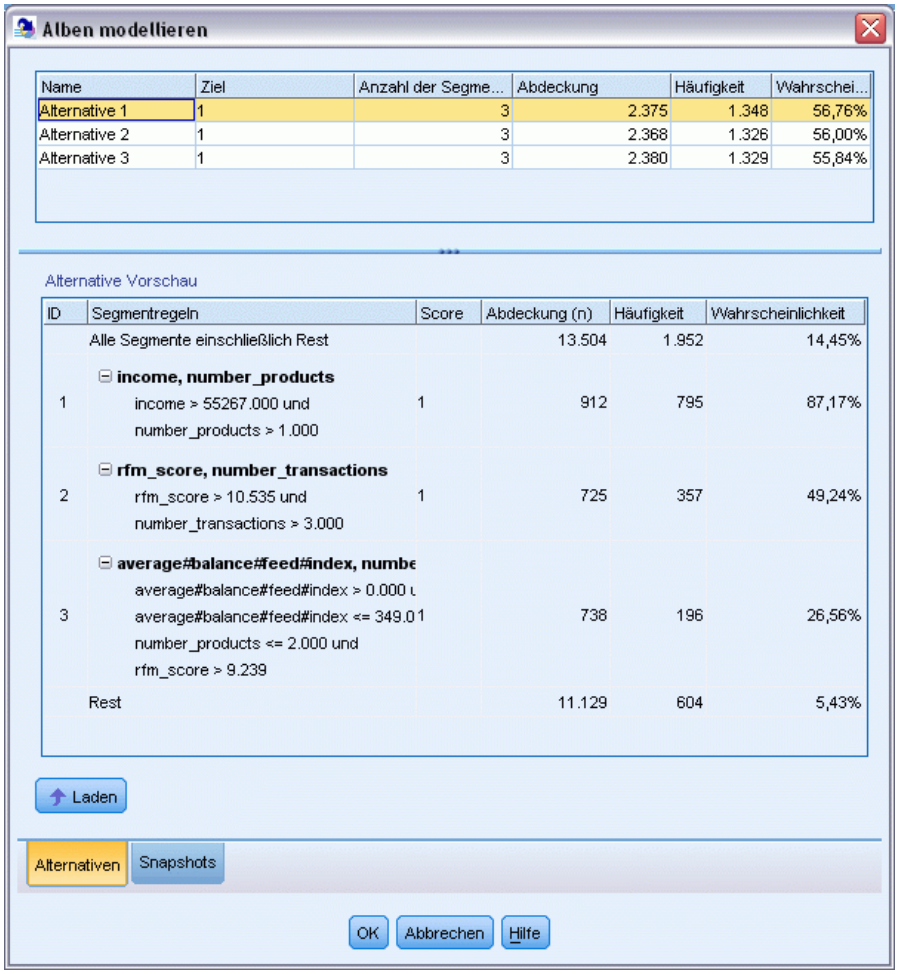
Mit diesem Tool wird die Standard-Mining-Aufgabe auf der Grundlage der im Entscheidungslistenknoten angegebenen Einstellungen ausgeführt. Die ausgeführte Aufgabe ergibt drei alternative Modelle, die auf der Registerkarte “Alternativen” im Dialogfeld “Alben modellieren” aufgeführt sind.

Abbildung 12-10
Verfügbare alternative Modelle



▶ Wählen Sie die erste Alternative in der Liste aus. Die entsprechenden Details werden im Bereich "Alternative Vorschau" angezeigt.

Abbildung 12-11
Ausgewähltes alternatives Modell



Im Fenster "Alternative Vorschau" können Sie schnell eine beliebige Anzahl an Alternativen durchsuchen, ohne das Arbeitsmodell ändern zu müssen, wodurch Sie ohne großen Aufwand mit verschiedenen Ansätzen experimentieren können.

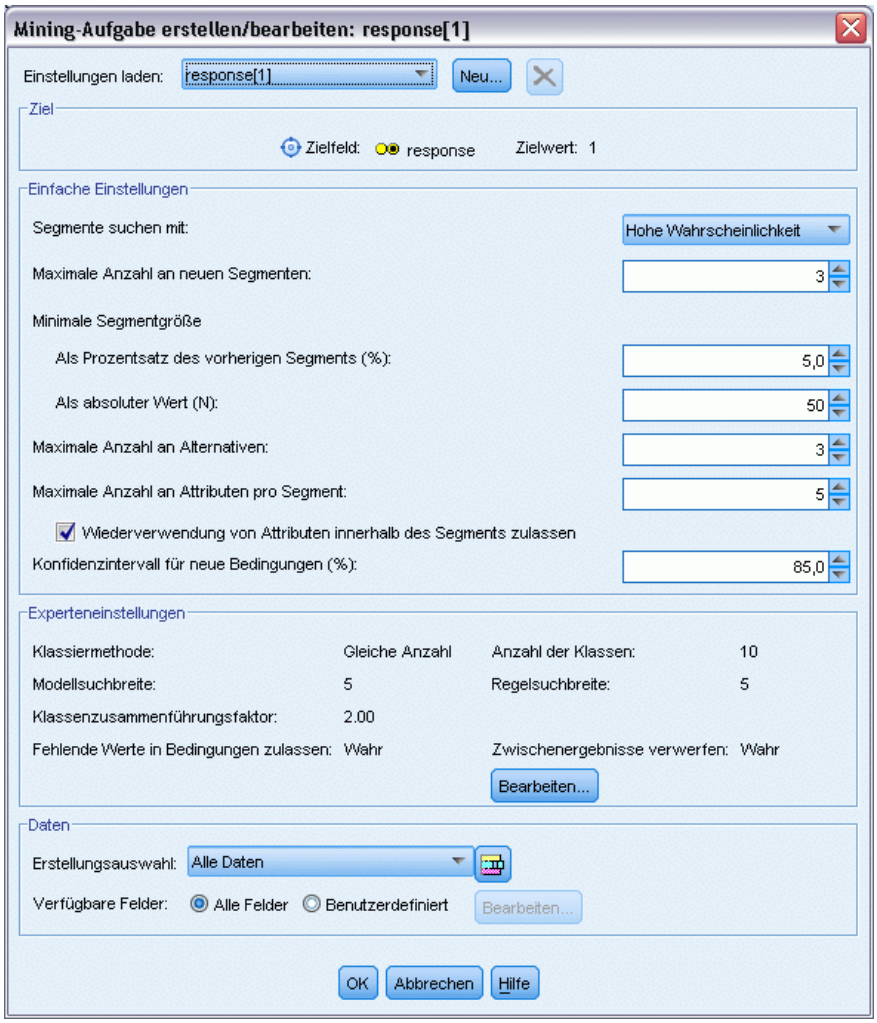
Anmerkung: Um das Modell besser sehen zu können, können Sie den Bereich "Alternative Vorschau" innerhalb des Dialogfelds, wie hier gezeigt, vergrößern. Sie können dies durch Ziehen der Bereichsgrenze erreichen.

Durch Verwendung von Regeln, die auf Prädiktoren wie Einkommen, Anzahl der Transaktionen pro Monat und RFM-Score basieren, ermittelt das Modell Segmente, bei denen die Antwortquote höher liegt als insgesamt für die Stichprobe. Bei einer Kombination der Segmente legt das Modell nahe, dass sich die Trefferquote steigern ließe, nämlich auf 56,76 %. Das Modell deckt jedoch nur einen kleinen Teil der Gesamtstichprobe ab und belässt über 11.000 Datensätze — darunter

mehrere Hundert Treffer— im Rest. Wir streben ein Modell an, das eine größere Anzahl dieser Treffer erfasst, aber weiterhin die Segmente mit niedrigen Trefferquoten ausschließt.

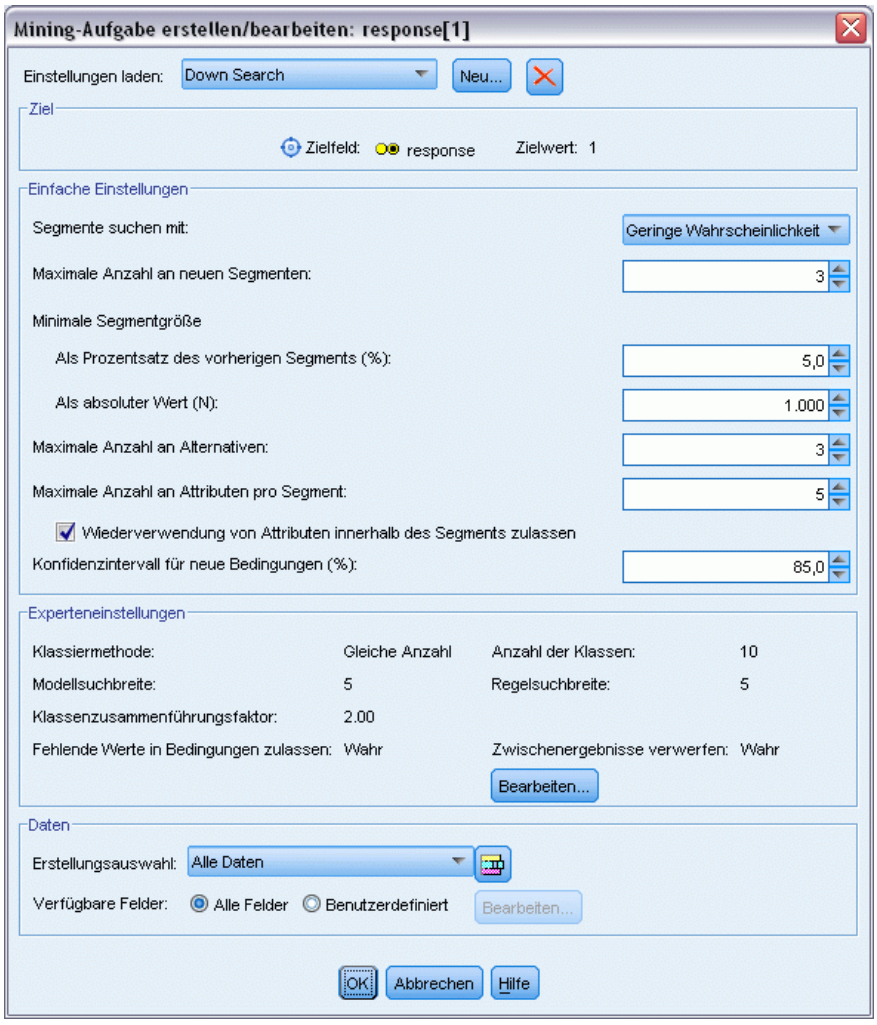
- Um einen anderen Modellierungsansatz auszuprobieren, wählen Sie folgende Optionsfolge aus den Menüs aus:
  Werkzeuge > Einstellungen

Abbildung 12-12
Dialogfeld "Mining-Aufgabe erstellen/bearbeiten"



- Klicken Sie auf die Schaltfläche Neu (rechts oben), um eine zweite Mining-Aufgabe hinzuzufügen, und geben Sie im Dialogfeld "Neue Einstellungen" als Namen für die Aufgabe *Abwärtssuche* ein.

Abbildung 12-13
Dialogfeld "Mining-Aufgabe erstellen/bearbeiten"

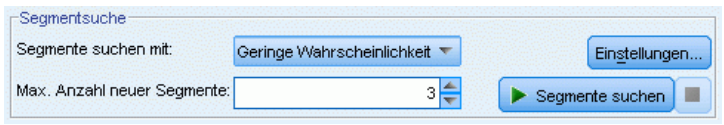


- Ändern Sie die Suchrichtung für die Aufgabe in Geringe Wahrscheinlichkeit. Dadurch sucht der Algorithmus anstatt nach den Segmenten mit den höchsten Antwortquoten nach den Segmenten mit den *niedrigsten* Antwortquoten.

- Erhöhen Sie die minimale Segmentgröße auf 1.000. Klicken Sie auf OK, um in den Viewer "Interaktive Liste" zurückzukehren.

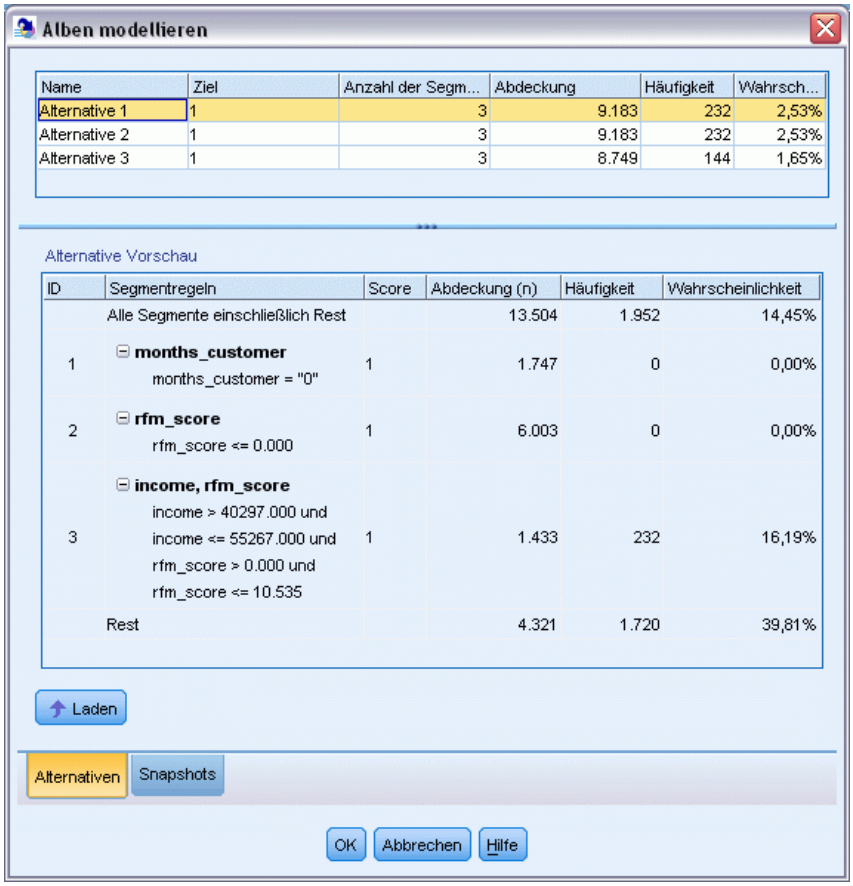
- Stellen Sie im Viewer "Interaktive Liste" sicher, dass der Bereich *Segmentsuche* die Details der neuen Aufgabe anzeigt, und klicken Sie auf Segmente suchen.

Abbildung 12-14
Segmente in einer neuen Mining-Aufgabe suchen



Die Aufgabe ergibt eine neue Menge von Alternativen, die in der Registerkarte “Alternativen” im Dialogfeld “Alben modellieren” angezeigt werden und die auf dieselbe Weise als Vorschau angezeigt werden können wie die vorherigen Ergebnisse.

Abbildung 12-15
Modellergebnisse für Abwärtssuche

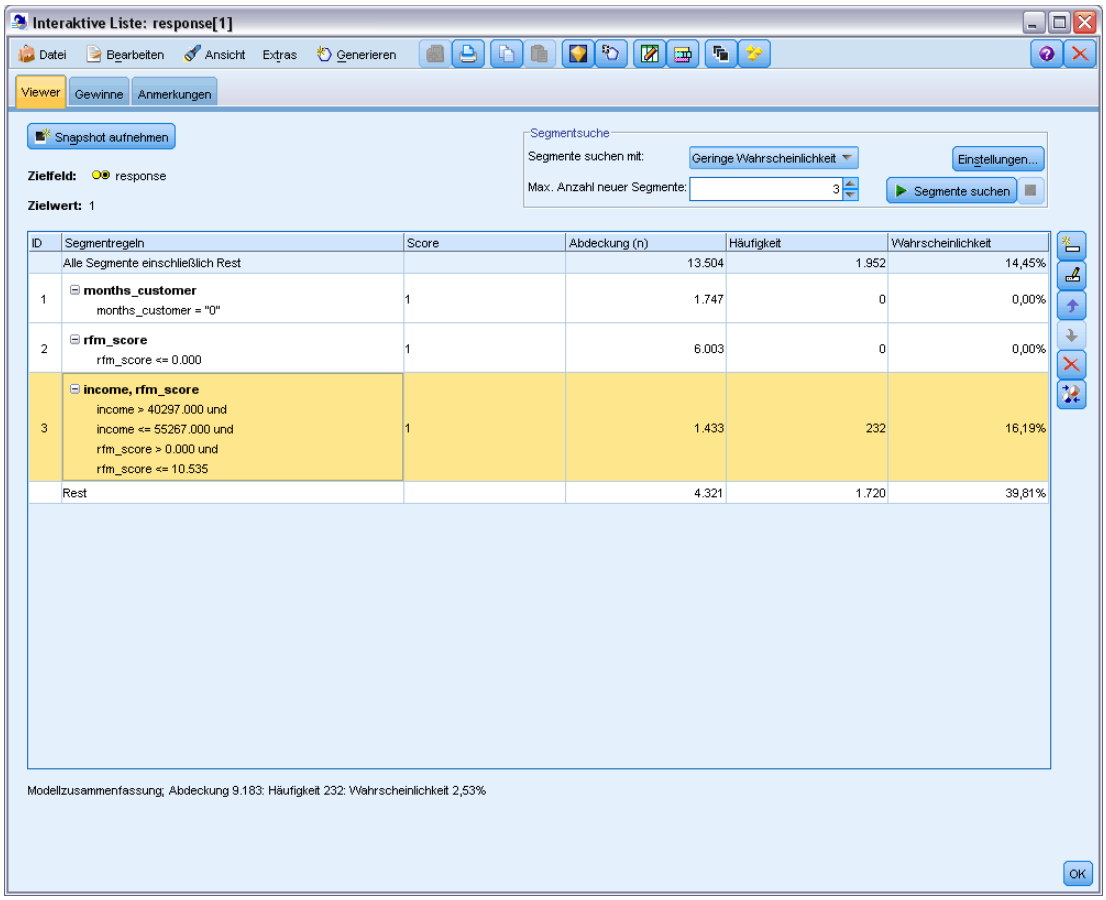


Diesmal ermitteln die einzelnen Modelle Segmente mit niedriger Antwortwahrscheinlichkeit und nicht mit hoher. Wenn wir die erste Alternative betrachten, sehen wir, dass einfach durch Ausführung dieser Segmente die Trefferquote für den Rest auf 39,81 % steigt. Dieser Wert liegt unter dem Wert des zuvor betrachteten Modells; diesmal wurde jedoch eine größere Abdeckung (also mehr Treffer insgesamt) erzielt.

Durch Kombination der beiden Ansätze – zuerst eine Suche des Typs “Geringe Wahrscheinlichkeit” zum Ausschluss irrelevanter Datensätze, gefolgt von Suche des Typs “Hohe Wahrscheinlichkeit” – können Sie dieses Ergebnis eventuell verbessern.

- Klicken Sie auf Laden, um dieses Modell (die erste gefundene Alternative in der Abwärtssuche) als Arbeitsmodell festzulegen, und klicken Sie auf OK, um das Dialogfeld “Alben modellieren” zu schließen.

Abbildung 12-16
Ausschließen von Segmenten



- Klicken Sie mit der rechten Maustaste auf jedes der ersten beiden Segmente und wählen Sie die Option Segment ausschließen. Zusammen erfassen diese Segmente fast 8.000 Datensätzen, die insgesamt 0 Treffer aufweisen. Es ist also sinnvoll, sie aus zukünftigen Angeboten auszuschließen. (Ausgeschlossene Segmente erhalten den Score null, um dies anzuzeigen.)

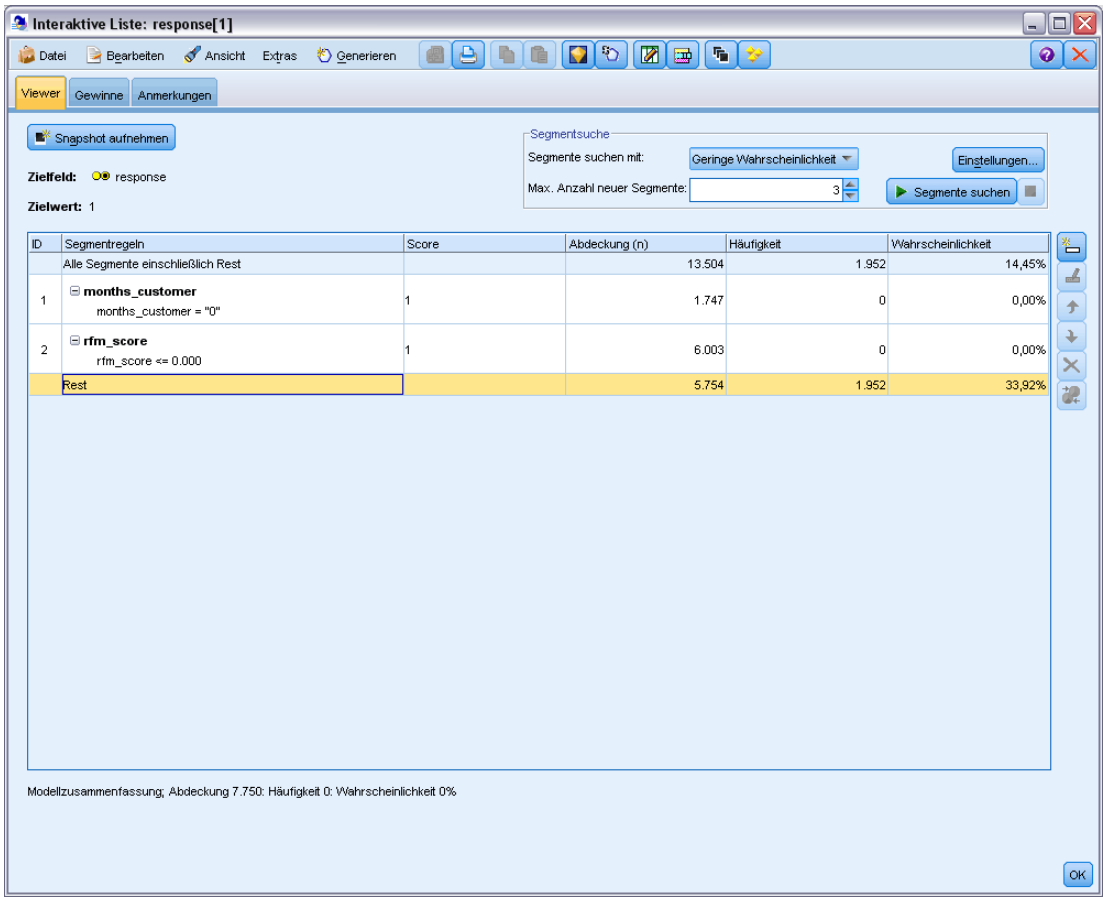
- Klicken Sie mit der rechten Maustaste auf das dritte Segment und wählen Sie die Option Segment löschen aus. Mit 16,19 %, unterscheidet sich die Trefferquote für dieses Segment nicht wesentlich von der Basis-Trefferquote von 14,45 %. Es fügt also nicht genug Informationen hinzu, die seine Beibehaltung rechtfertigen würden.

  Hinweis: Das Löschen eines Segments ist nicht dasselbe wie der Ausschluss eines Segments. Beim Ausschluss eines Segments wird einfach die Art und Weise geändert, wie das Segment gescort wird, wohingegen es beim Löschen vollständig aus dem Modell entfernt wird.

  Nach Ausschluss der Segmente mit den niedrigsten Trefferquoten suchen wir nun im Rest nach Segmenten mit hoher Trefferquote.

- Klicken Sie in der Tabelle auf die Rest-Zeile, um diese auszuwählen, sodass die nächste Mining-Aufgabe nur auf den Rest angewendet wird.

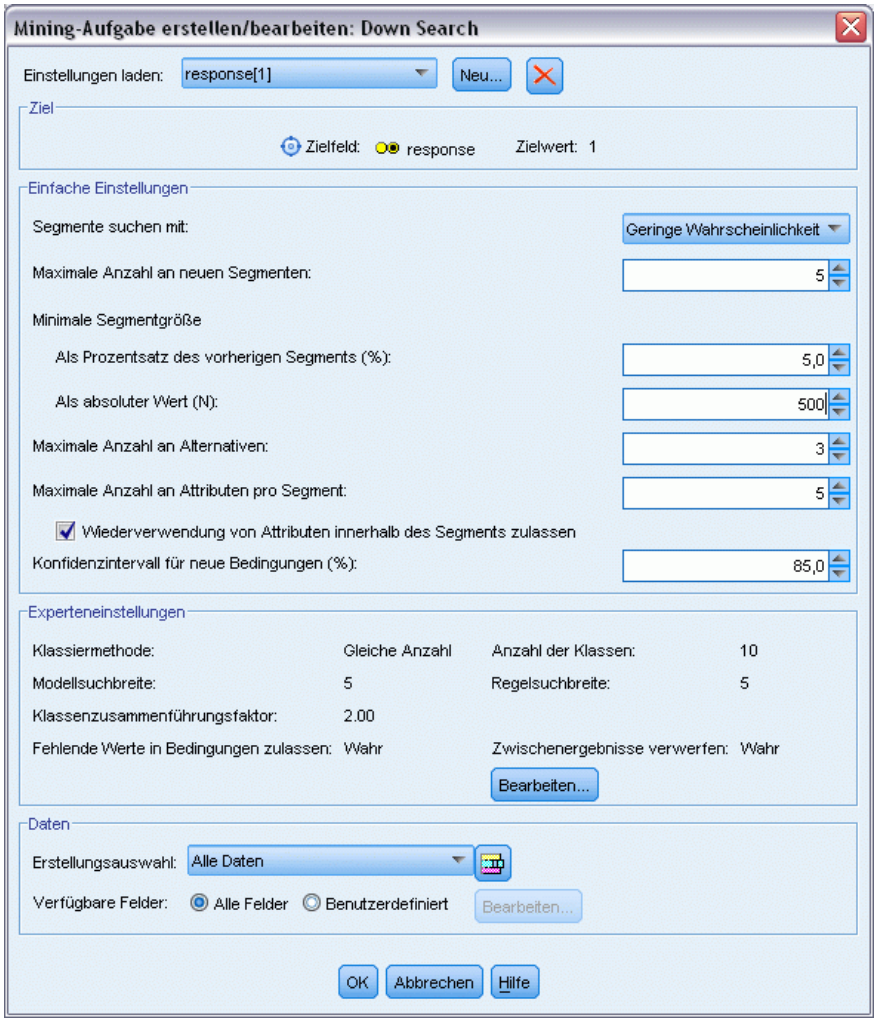
Abbildung 12-17
Auswählen von Segmenten



- Klicken Sie bei ausgewähltem Rest auf Einstellungen, um das Dialogfeld “Mining-Aufgaben erstellen/bearbeiten” erneut zu öffnen.
- Wählen Sie am oberen Rand unter Einstellungen laden die Standard-Mining-Aufgabe aus: response[1].
- Bearbeiten Sie die Daten unter Einfache Einstellungen, um die Anzahl der neuen Segmente auf 5 und die minimale Segmentgröße auf 500 zu erhöhen.

▶ Klicken Sie auf OK, um in den Viewer “Interaktive Liste” zurückzukehren.

Abbildung 12-18
Auswählen der Standard-Mining-Aufgabe

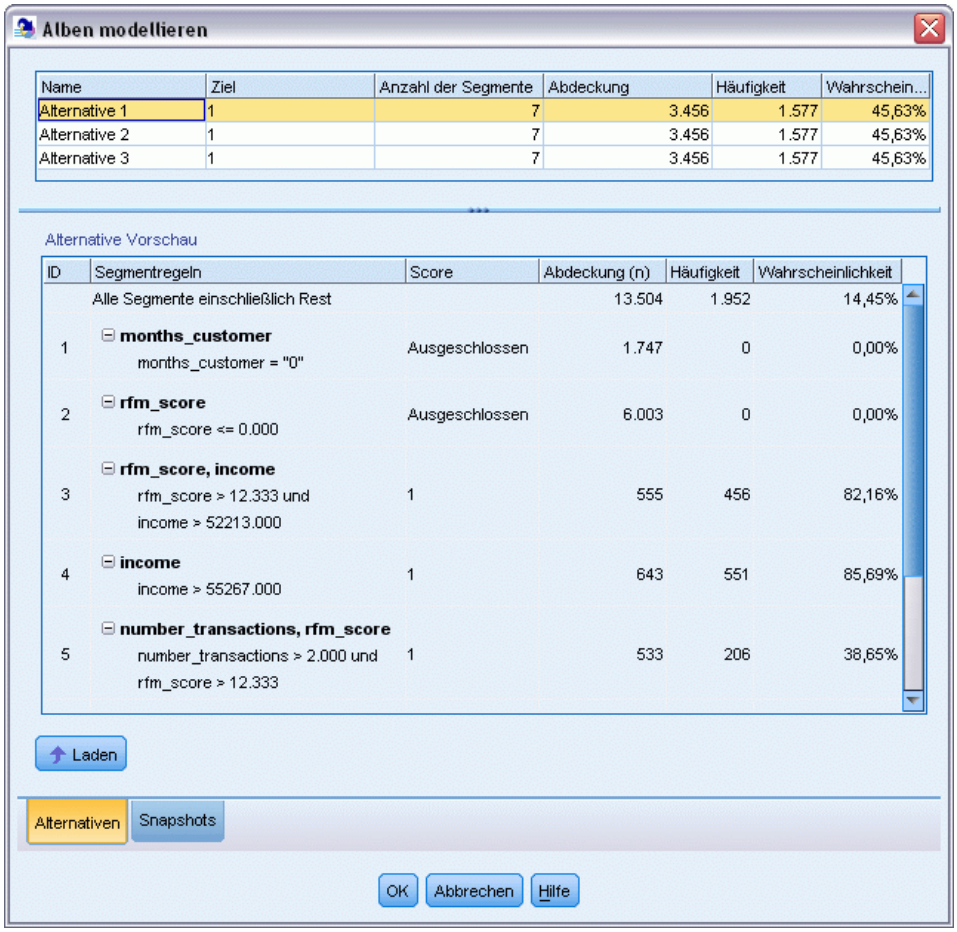


▶ Klicken Sie auf Segmente suchen.

Dadurch erhalten Sie eine weitere Menge an alternativen Modellen. Durch Einspeisung der Ergebnisse einer Mining-Aufgabe in eine andere enthalten diese letzten Modelle eine Mischung aus Segmenten mit hohen und niedrigen Trefferquoten. Segmente mit niedrigen Antwortquoten werden ausgeschlossen, sie werden also als null gescort, wohingegen die eingeschlossenen Segmente als 1 gescort werden. Die Gesamtstatistik spiegelt diese Ausschlüsse wider: Das erste

alternative Modell weist eine Trefferquote von 45,63 % und eine höhere Abdeckung (1.577 Treffer bei 3.456 Datensätzen) auf als alle vorherigen Modelle.

Abbildung 12-19
Alternativen für kombiniertes Modell

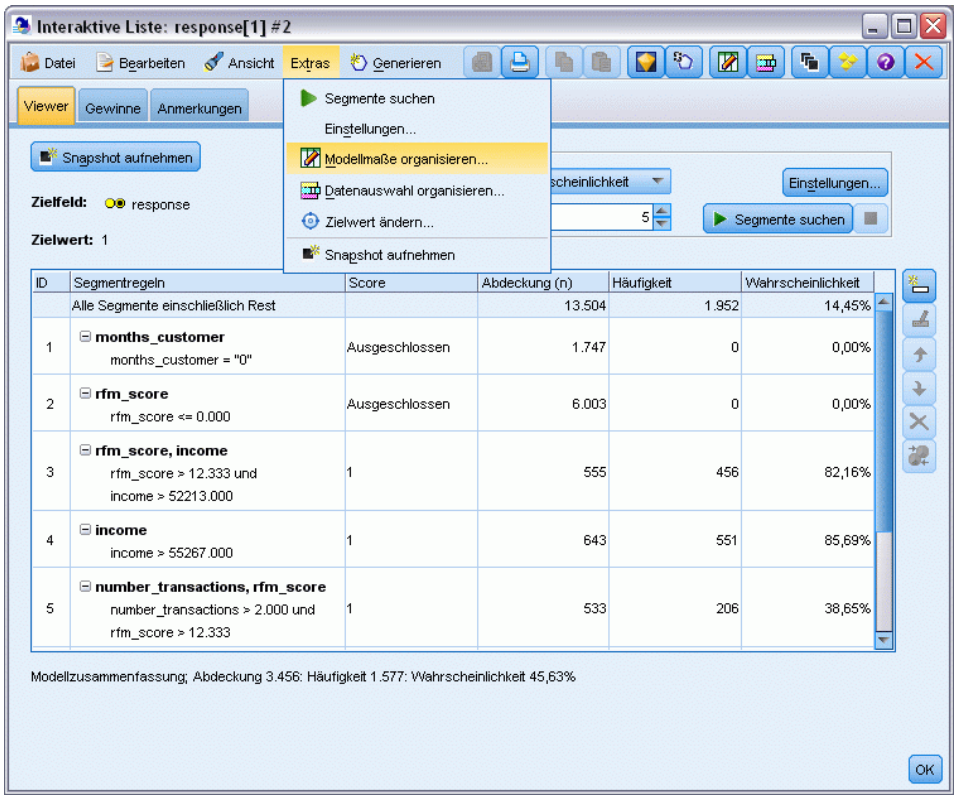


- Zeigen Sie die Vorschau der ersten Alternative an und wählen Sie anschließend Laden, um dieses Modell als Arbeitsmodell zu verwenden.

Berechnen von benutzerdefinierten Maßen mithilfe von Excel

▶ Um weitere Einblicken in die Funktionsweise des Modells in der Praxis zu gewinnen, wählen Sie im Menü “Extras” die Option Modellmaße organisieren aus.

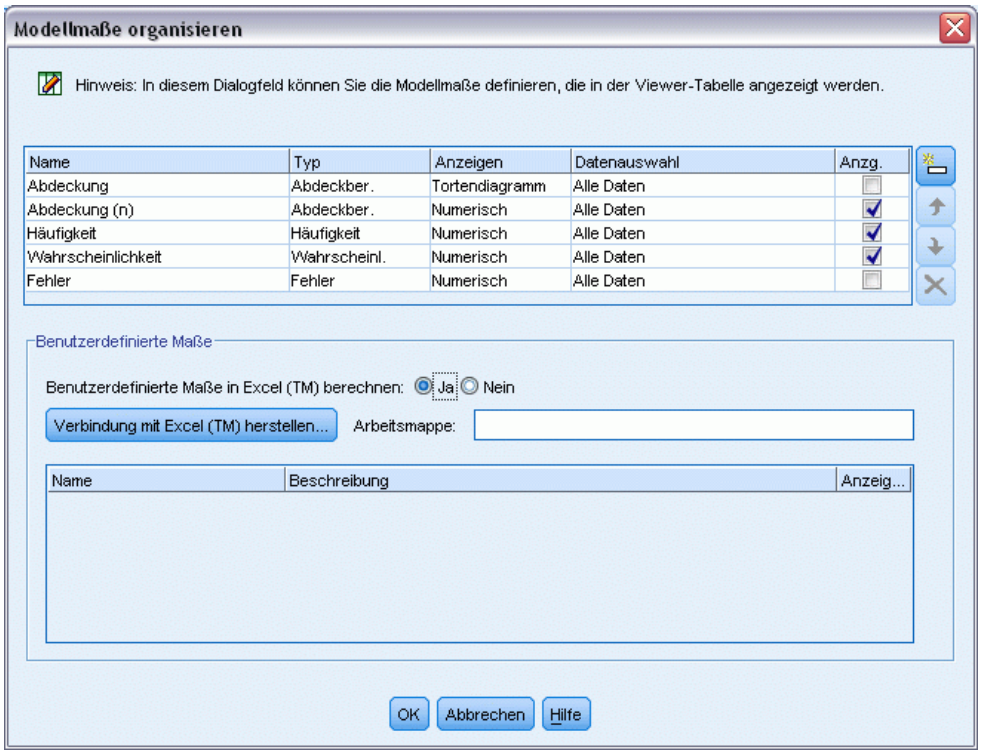
Abbildung 12-20
Organisieren von Modellmaßen



Im Dialogfeld “Modellmaße organisieren” können Sie die Maße (bzw. Spalten) auswählen, die im Viewer “Interaktive Liste” angezeigt werden sollen. Außerdem können Sie angeben, ob die Maße anhand aller Datensätze oder anhand einer ausgewählten Untergruppe berechnet werden

sollen, und Sie können auswählen, dass nach Möglichkeit ein Kreisdiagramm und keine Zahl angezeigt werden soll.

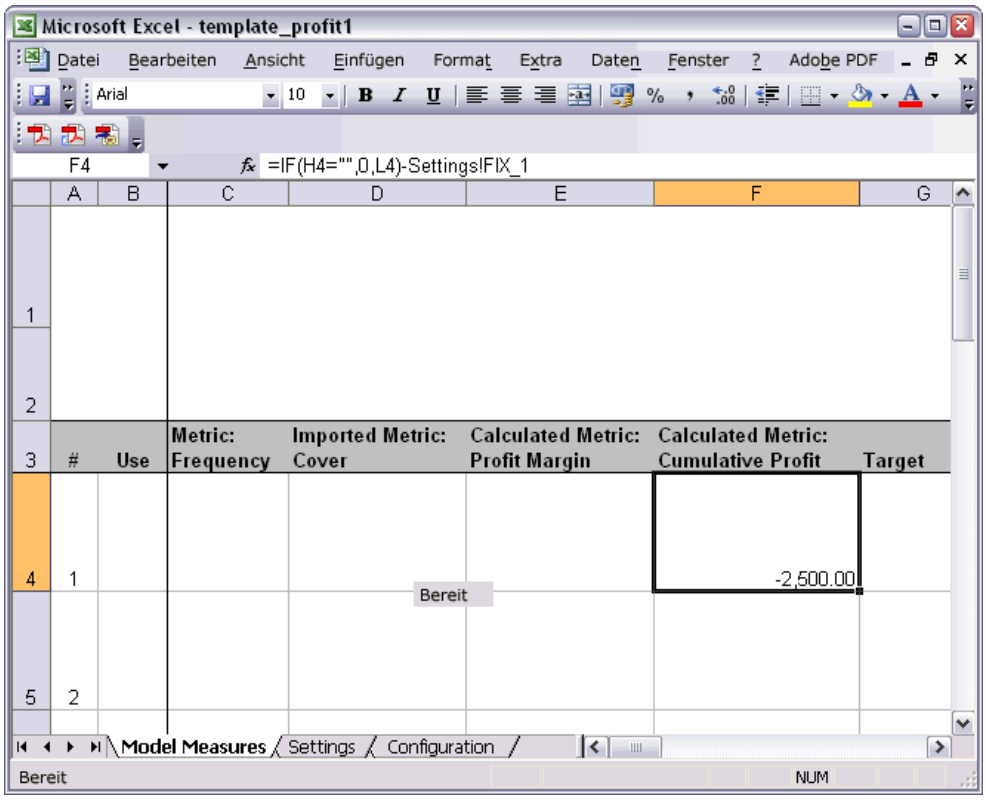
Abbildung 12-21
Dialogfeld "Modellmaße organisieren"



Wenn bei Ihnen Microsoft Excel installiert ist, können Sie außerdem eine Verknüpfung zu einer Excel-Vorlage herstellen, die dann benutzerdefinierte Maße berechnet und zur interaktiven Anzeige hinzufügt.

- Setzen Sie Benutzerdefinierte Maße in Excel (TM) berechnen im Dialogfeld "Modellmaße organisieren" auf Ja.
- Klicken Sie auf die Schaltfläche Verbindung mit Excel (TM) herstellen.
- Wählen Sie die Arbeitsmappe *template_profit.xlt* aus, die sich im Unterordner *streams* des Ordners *Demos* Ihrer IBM® SPSS® Modeler-Installation befindet, und klicken Sie auf Öffnen, um die Kalkulationstabelle zu öffnen.

Abbildung 12-22
Excel-Arbeitsblatt "Modellmaße"



Die Excel-Vorlage enthält drei Arbeitsblätter:

- Model Measures (Modellmaße) zeigt Modellmaße an, die aus dem Modell importiert wurden, und berechnet benutzerdefinierte Maße, die dann wieder in das Modell exportiert werden können.
- Settings (Einstellungen) enthält Parameter für die Berechnung von benutzerdefinierten Maßen.
- Configuration (Konfiguration) legt die Maße fest, die aus dem Modell importiert und in das Modell exportiert werden sollen.

Folgende Metriken werden wieder in das Modell exportiert:

- **Profit Margin (Profitspanne).** Der Nettoerlös aus dem Segment
- **Cumulative Profit (Kumulierter Profit).** Der Gesamtprofit aus der Kampagne

Durch folgende Formeln definiert:

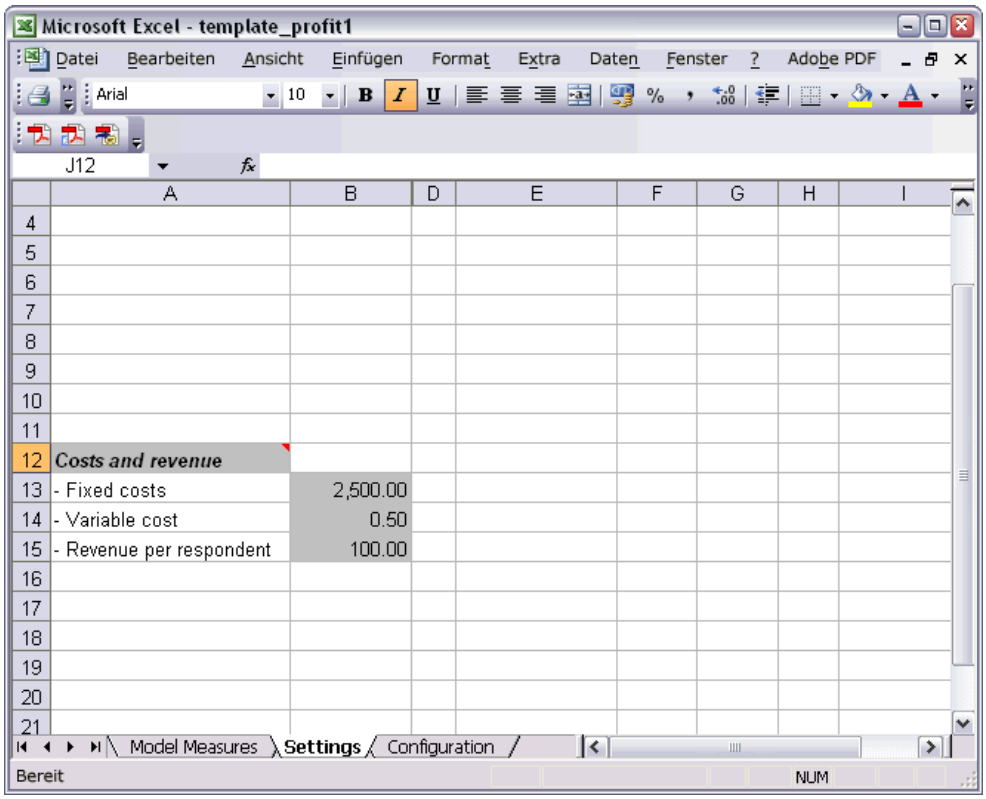
Profit Margin (Profitspanne) = Requency (Häufigkeit) * Revenue per respondent (Erlös pro Teilnehmer) - Cover (Abdeckung) * Variable cost (

Cumulative Profit (Kumulierter Profit) = Total Profit Margin (Gesamtprofitspanne) - Fixed cost (Feste Kosten)

Beachten Sie, dass Häufigkeit und Abdeckung aus dem Modell importiert werden.

Die Parameter für Kosten und Erlös werden vom Benutzer im Arbeitsblatt “Settings” (Einstellungen) angegeben.

Abbildung 12-23
Excel-Arbeitsblatt “Einstellungen”



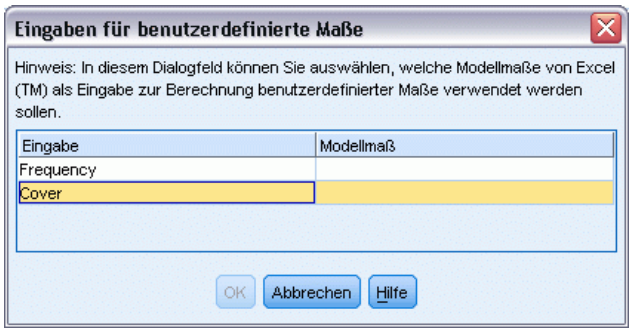
Fixed cost (Feste Kosten) sind die Einrichtungskosten für die Kampagne, beispielsweise Entwurf und Planung.

Variable cost (Variable Kosten) sind die Kosten für die Unterbreitung des Angebots für die einzelnen Kunden, also beispielsweise Umschläge und Briefmarken.

Revenue per respondent (Erlös pro Teilnehmer) ist der Nettoerlös für einen Kunden, der auf das Angebot reagiert.

- Um die Verknüpfung zurück zum Modell abzuschließen, wechseln Sie mithilfe der Windows-Taskleiste (oder durch Drücken von Alt-Tab) zurück zum Viewer “Interaktive Liste”.

Abbildung 12-24
Auswählen von Eingaben für benutzerdefinierte Maße



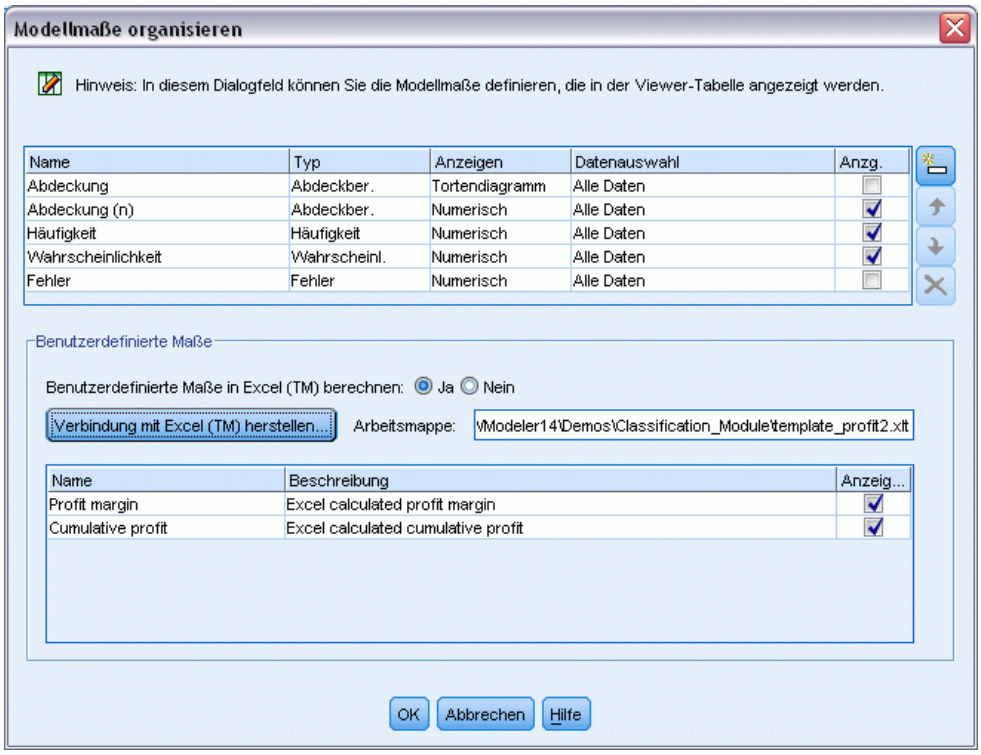
Das Dialogfeld “Eingaben für benutzerdefinierte Maße” wird angezeigt. Hier können Sie Eingaben aus dem Modell bestimmten in der Vorlage definierten Parametern zuordnen. In der linken Spalte sind die verfügbaren Maße aufgelistet und in der rechten Spalte werden diese Maße den im Arbeitsblatt “Configuration” (Konfiguration) definierten Tabellenkalkulationsparametern zugeordnet.

- Wählen Sie in der Spalte Modellmaße die Einträge Frequency und Cover (n) für die entsprechenden Eingaben aus und klicken Sie auf OK.

Im vorliegenden Fall entsprechen die Parameternamen in der Vorlage – “Frequency” (Häufigkeit) und “Cover (n)” (Abdeckung) – zufällig den Eingaben; es könnten jedoch auch andere Namen verwendet werden.

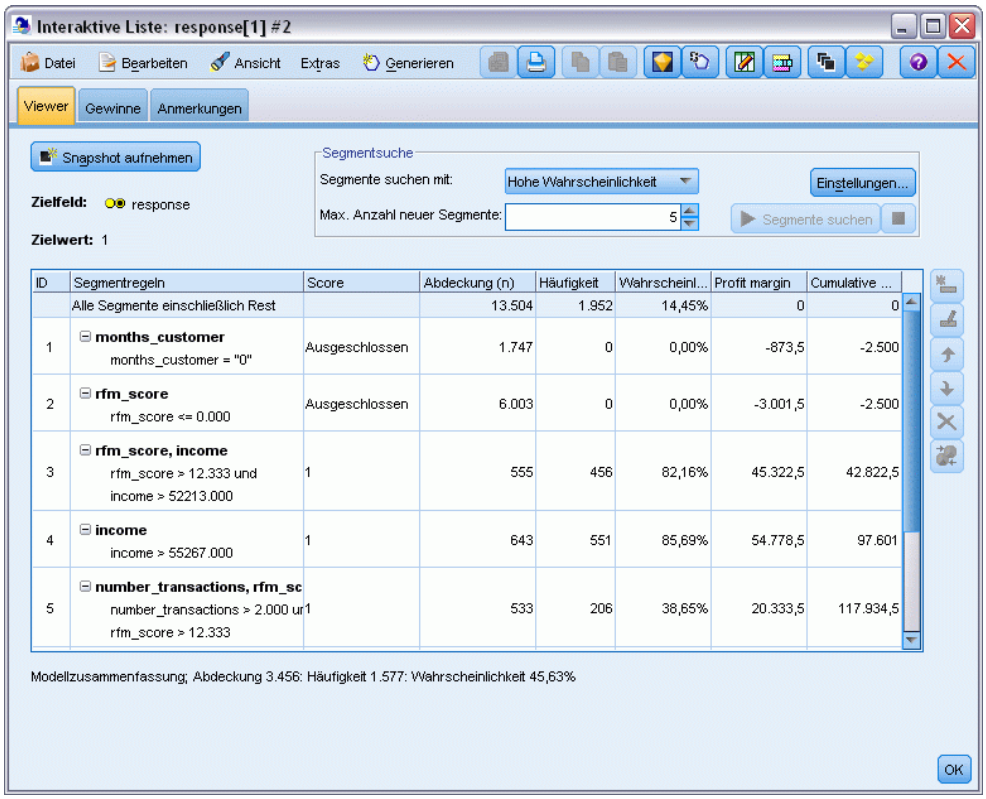
- Klicken Sie im Dialogfeld “Modellmaße organisieren” auf OK, um die Anzeige des Viewers “Interaktive Liste” zu aktualisieren.

Abbildung 12-25
Dialogfeld “Modellmaße organisieren” mit benutzerdefinierten Maßen aus Excel



Die neuen Maße werden nun als neue Spalten im Fenster hinzugefügt und bei jeder Aktualisierung des Modells neu berechnet.

Abbildung 12-26
Benutzerdefinierte Maße aus Excel im Viewer "Interaktive Liste"



Durch Bearbeiten der Excel-Vorlage können beliebig viele benutzerdefinierte Maße erstellt werden.

Ändern der Excel-Vorlage

IBM® SPSS® Modeler wird zwar mit einer Excel-Standardvorlage ausgeliefert, die im Viewer "Interaktive Liste" verwendet werden kann, jedoch kann es sinnvoll sein, die Einstellungen zu ändern oder eigene Einstellungen hinzuzufügen. Beispielsweise kann es sein, dass die Kosten in der Vorlage für Ihr Unternehmen nicht zutreffen und korrigiert werden müssen.

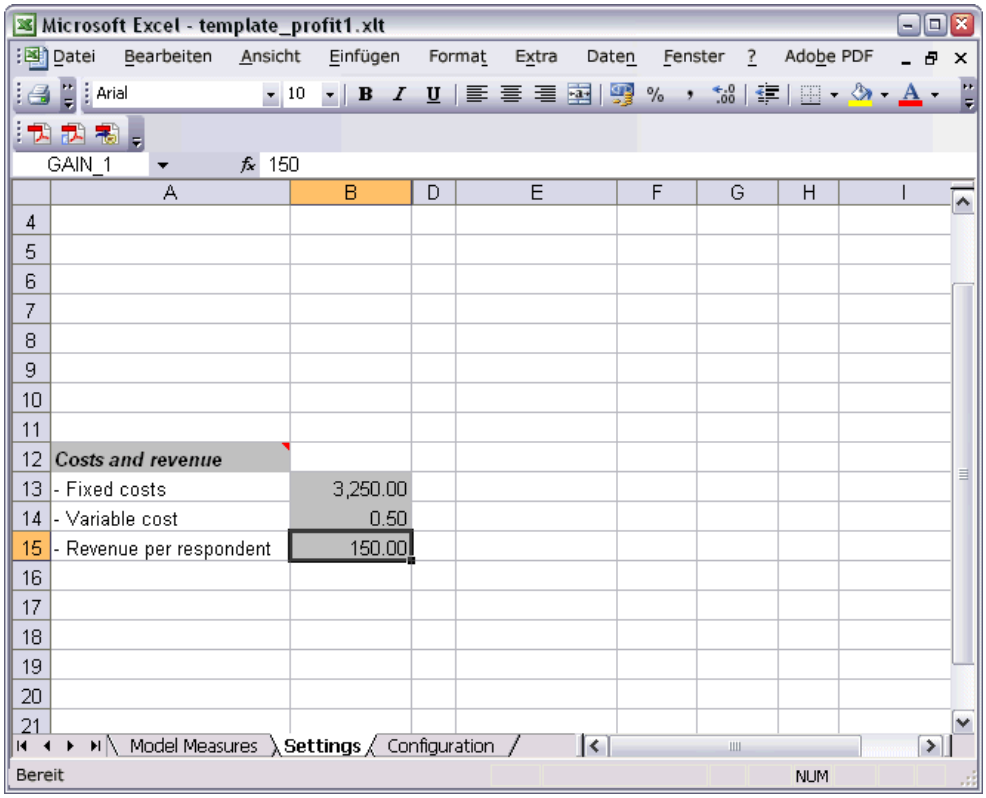
Hinweis: Wenn Sie eine bestehende Vorlage ändern oder eine eigene Vorlage erstellen, müssen Sie die Datei unbedingt mit der Dateierweiterung *.xlt* von Excel 2003 speichern.

So können Sie die Standardvorlage mit neuen Details zu Kosten und Erlösen und den Viewer "Interaktive Liste" mit den neuen Werten aktualisieren:

- Wählen Sie im Viewer "Interaktive Liste" im Menü "Extras" die Option Modellmaße organisieren aus.
- Klicken Sie im Dialogfeld "Modellmaße organisieren" auf Verbindung mit Excel™ herstellen.

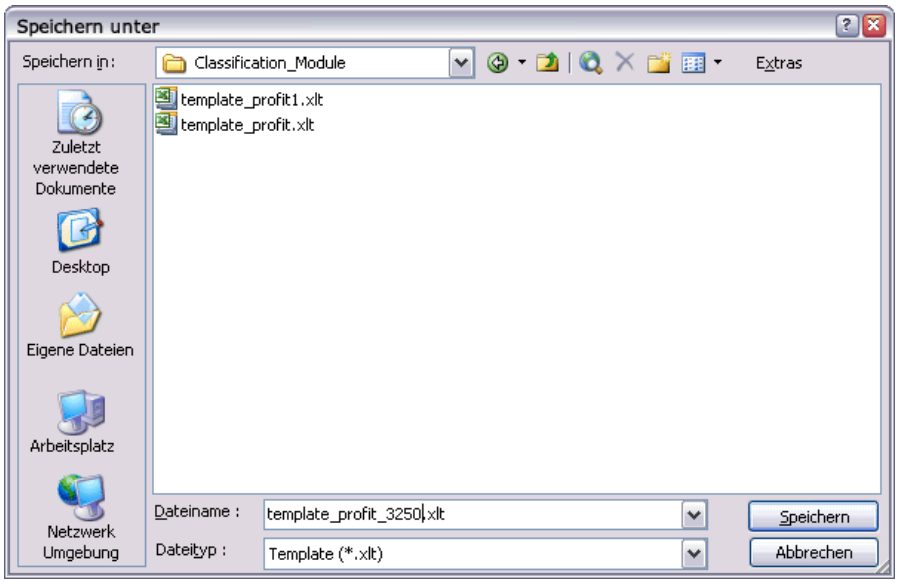
- Wählen Sie die Arbeitsmappe *template_profit.xlt* und klicken Sie auf Öffnen, um das Arbeitsblatt anzuzeigen.
- Wählen Sie das Arbeitsblatt “Einstellungen” aus.
- Ändern Sie den Wert für Fixed costs (Feste Kosten) in 3.250,00 und den Wert für Revenue per respondent (Erlös pro Teilnehmer) in 150,00.

Abbildung 12-27
Geänderte Werte im Excel-Arbeitsblatt “Einstellungen”



▶ Speichern Sie die geänderte Vorlage mit einem eindeutigen, aussagekräftigen Dateinamen. Achten Sie darauf, dass die Datei die Erweiterung *.xlt* von Excel 2003 trägt.

Abbildung 12-28
Speichern der geänderten Excel-Vorlage



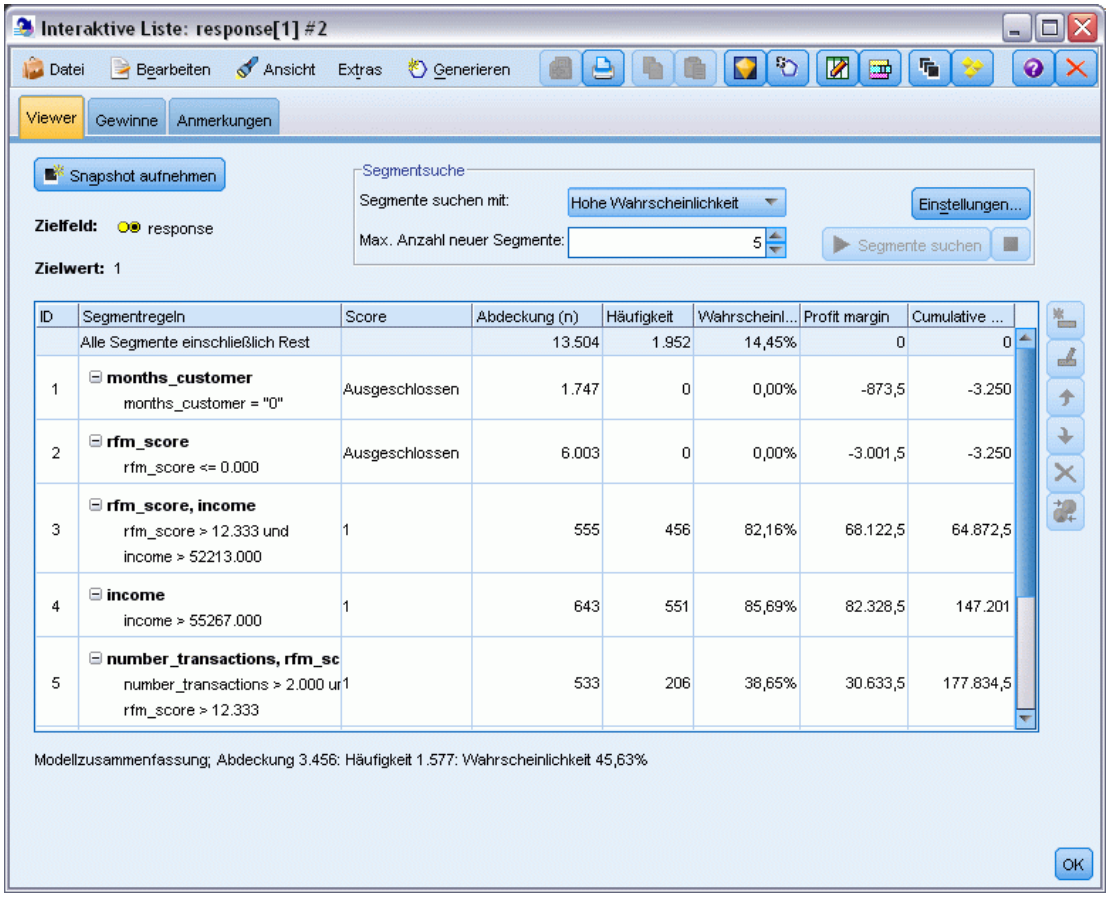
▶ Wechseln Sie mithilfe der Windows-Taskleiste (oder durch Drücken von Alt+Tab) zurück zum Viewer “Interaktive Liste”.

Wählen Sie im Dialogfeld “Eingaben für benutzerdefinierte Maße” die anzuzeigenden Maße aus und klicken Sie auf OK.

▶ Klicken Sie im Dialogfeld “Modellmaße organisieren” auf OK, um die Anzeige des Viewers “Interaktive Liste” zu aktualisieren.

In diesem Beispiel wurde natürlich nur eine einfache Methode zur Änderung der Excel-Vorlage gezeigt. Es sind weitere Änderungen möglich, mit denen Daten automatisch von dem bzw. an den Viewer "Interaktive Liste" übertragen werden können oder mit denen innerhalb von Excel gearbeitet werden kann, um andere Ausgaben, wie beispielsweise Diagramme zu erstellen.

Abbildung 12-29
Geänderte benutzerdefinierte Maße aus Excel im Viewer "Interaktive Liste"



Speichern der Ergebnisse

Um ein Modell während der interaktiven Sitzung für die spätere Verwendung zu speichern, können Sie einen Snapshot des Modells aufnehmen. Dieser wird dann auf der Registerkarte "Snapshots" aufgeführt. Sie können jederzeit während der interaktiven Sitzung zu jedem beliebigen gespeicherten Snapshot zurückkehren.

Auf diese Weise können Sie mit weiteren Mining-Aufgaben experimentieren, um nach zusätzlichen Segmenten zu suchen. Außerdem können Sie bestehende Segmente bearbeiten, benutzerdefinierte Segmente auf der Grundlage Ihrer eigenen Geschäftsregeln einfügen, Datenauswahlmöglichkeiten erstellen, um das Modell für bestimmte Gruppen zu optimieren, und das Modell auf andere Weise anpassen. Schließlich können Sie jedes Segment nach Bedarf explizit ein- bzw. ausschließen, um anzugeben, wie die einzelnen Segmente gescort werden.

Wenn Sie mit den Ergebnissen zufrieden sind, können Sie mit dem Menü “Generieren” ein Modell erzeugen, das zu Streams hinzugefügt oder zu Scoring-Zwecken verwendet werden kann.

Alternativ können Sie den aktuellen Status Ihrer interaktiven Sitzung für einen späteren Zeitpunkt speichern, indem Sie im Modell “Datei” die Option Modellierungsknoten aktualisieren auswählen. Dadurch wird der Modellierungsknoten der Entscheidungsliste mit den aktuellen Einstellungen aktualisiert, darunter Mining-Aufgaben, Modell-Snapshots, Datenauswahl und benutzerdefinierte Maße. Bei der nächsten Ausführung des Streams müssen Sie lediglich sicherstellen, dass im Modellierungsknoten der Entscheidungsliste die Option Gespeicherte Informationen aus interaktiver Sitzung verwenden ausgewählt ist, um den aktuellen Status der Sitzung wiederherzustellen. Für weitere Informationen siehe Thema Entscheidungsliste in Kapitel 9 in *IBM SPSS Modeler 14.2-Modellierungsknoten*.

Kapitel
13

Klassifizieren von Kunden im Telekommunikationsbereich (multinomiale logistische Regression)

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 Zielfelds ein kategoriales verwendet wird.

Nehmen wir beispielsweise an, dass ein Telekommunikationsanbieter seinen Kundenstamm nach Servicenutzungsmustern in vier Gruppen unterteilt hat. Wenn demografische Daten zum Vorhersagen der Gruppenzugehörigkeit verwendet werden können, sind angepasste Angebote für die einzelnen potenziellen Kunden möglich.

In diesem Beispiel wird ein Stream namens *telco_custcat.str* verwendet, der Bezug nimmt auf die Datendatei *telco.sav*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *telco_custcat.str* befindet sich im Verzeichnis *streams*.

Dieses Beispiel konzentriert sich auf die Verwendung von demografischen Daten zur Vorhersage von Nutzungsmustern. Das Zielfeld *custcat* weist vier mögliche Werte auf, die den vier Kundengruppen entsprechen:

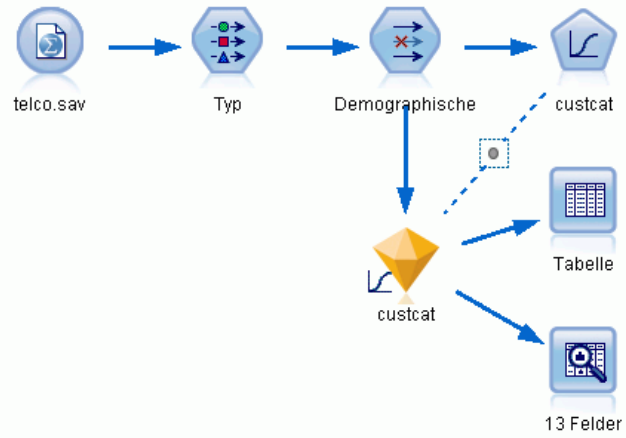
Wert	Label
1	Basic Service (Basis-Service)
2	E-Service
3	Plus Service (Plus-Service)
4	Total Service (Umfassender Service)

Da das Ziel mehrere Kategorien aufweist, wird ein multinomiales Modell verwendet. Bei einem Ziel mit zwei verschiedenen Kategorien, wie "Ja/Nein", "Wahr/Falsch", "Abwanderung/Keine Abwanderung" könnte stattdessen ein binomiales Modell erstellt werden. Für weitere Informationen siehe Thema Kundenabwanderung bei Telekommunikationsunternehmen (binomiale logistische Regression) in Kapitel 14 auf S. 164.

Erstellen des Streams

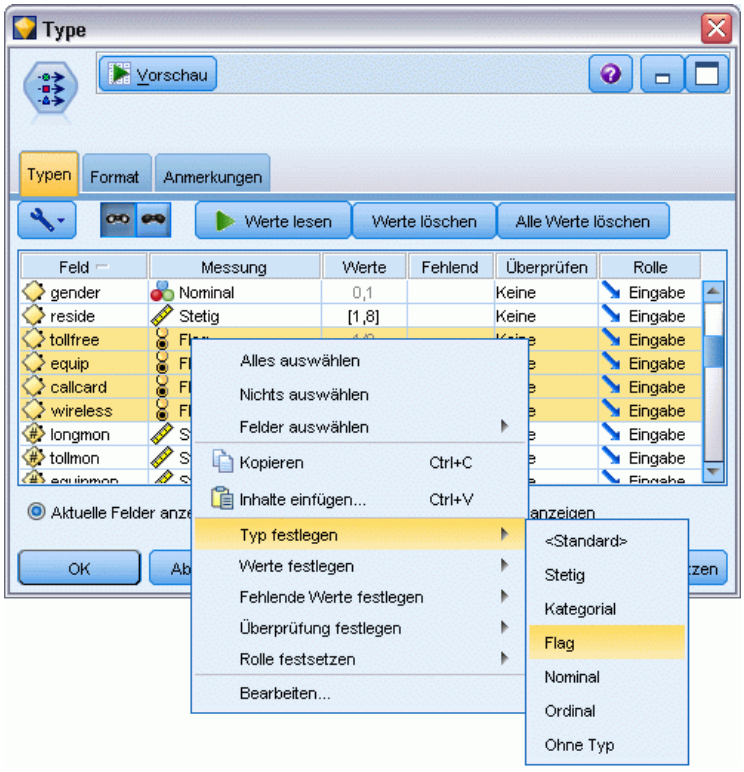
- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

Abbildung 13-1
Beispiel-Stream zur Klassifizierung von Kunden mithilfe der multinomialen logistischen Regression



- Fügen Sie einen Typknoten hinzu und klicken Sie auf Werte lesen. Achten Sie dabei darauf, dass alle Messniveaus korrekt festgelegt werden. Beispielsweise können die meisten Felder mit den Werten 1 und 0 als Flags betrachtet werden.

Abbildung 13-2
Festlegen des Messniveaus für mehrere Felder



Tipp: Um die Eigenschaften für mehrere Felder mit ähnlichen Werten (z. B. 0/1) zu ändern, klicken Sie auf die Überschrift der Spalte *Werte* (um die Felder nach ihrem Wert zu sortieren) und halten Sie anschließend die Umschalttaste gedrückt, während Sie mit der Maus oder den Pfeiltasten alle Felder auswählen, die geändert werden sollen. Anschließend können Sie mit der rechten Maustaste auf die Auswahl klicken, um das Messniveau oder andere Attribute der ausgewählten Felder zu ändern.

Beachten Sie, dass *Geschlecht* treffender als Feld mit einem Set von zwei Werten betrachtet wird denn als Flag. Belassen Sie also seinen Wert für "Messniveau" bei Nominal.

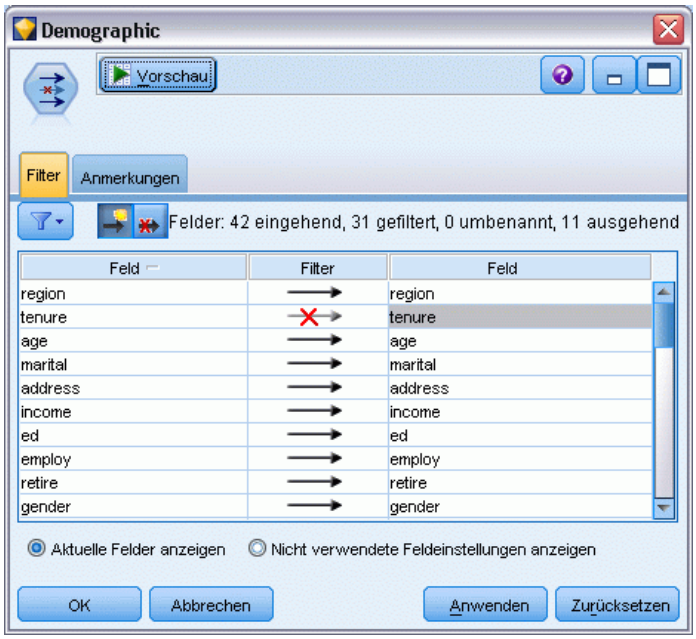
- Ändern Sie die Rolle für das Feld *custcat* in Ziel. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

Abbildung 13-3
Festlegen der Feldrolle

Da sich dieses Beispiel auf demografische Daten konzentriert, sollten Sie einen Filterknoten verwenden, mit dem nur die relevanten Felder (*region* (Region), *age* (Alter) *marital* (Familienstand), *address* (Adresse), *income* (Einkommen), *ed* (Bildung), *employ* (Beschäftigung),

retire (Ruhestand), *gender* (Geschlecht), *reside* (Wohnsitz) und *custcat* (Benutzerdef. Kategorie)) eingeschlossen werden. Die anderen Felder können für diese Analyse ausgeschlossen werden.

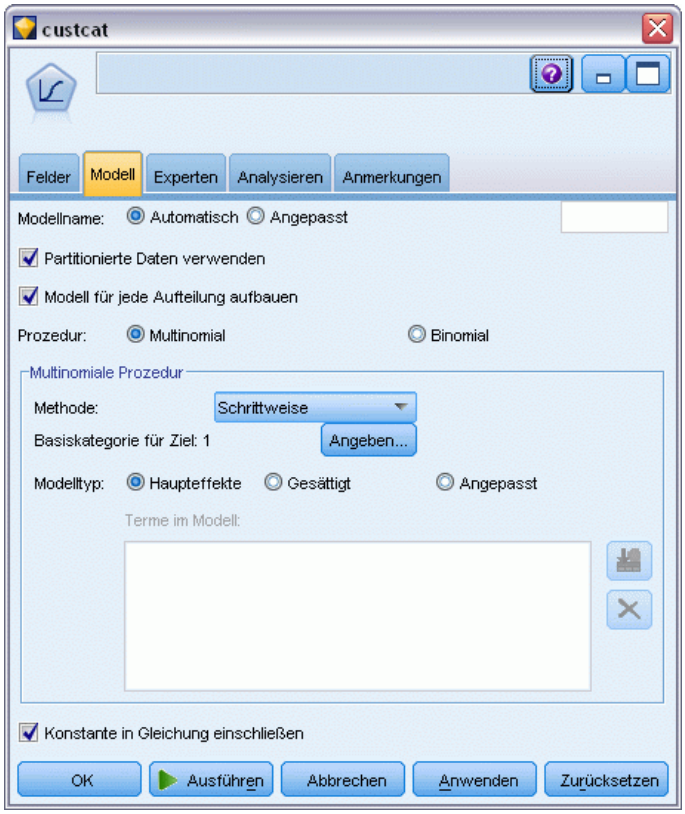
Abbildung 13-4
Filtern nach demografischen Feldern



(Alternativ können Sie die Rolle für diese Felder in Keine ändern, anstatt sie auszuschließen, oder die gewünschten Felder im Modellierungsknoten auswählen.)

- Klicken Sie im Logistikknoten auf die Registerkarte Modell und wählen Sie die Methode Schrittweise aus. Wählen Sie außerdem die Optionen Multinomial, Haupteffekte und Konstante in Gleichung einschließen aus.

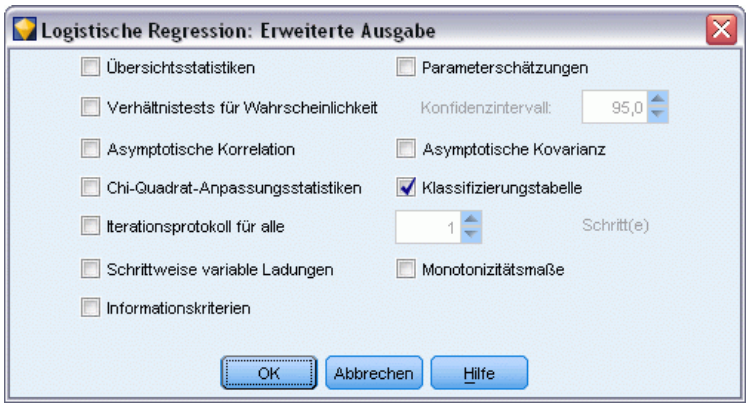
Abbildung 13-5
Auswählen Modelloptionen



Behalten Sie 1 als Basiskategorie für das Ziel bei. Das Modell vergleicht andere Kunden mit den Kunden, die den Basis-Service abonniert haben.

- Wählen Sie auf der Registerkarte “Experten” den Modus Experten aus und wählen Sie die Option Ausgabe; wählen Sie dann auf der Registerkarte “Erweiterte Ausgabe” die Option Klassifikationstabelle.

Abbildung 13-6
Auswahl der Ausgabeoptionen



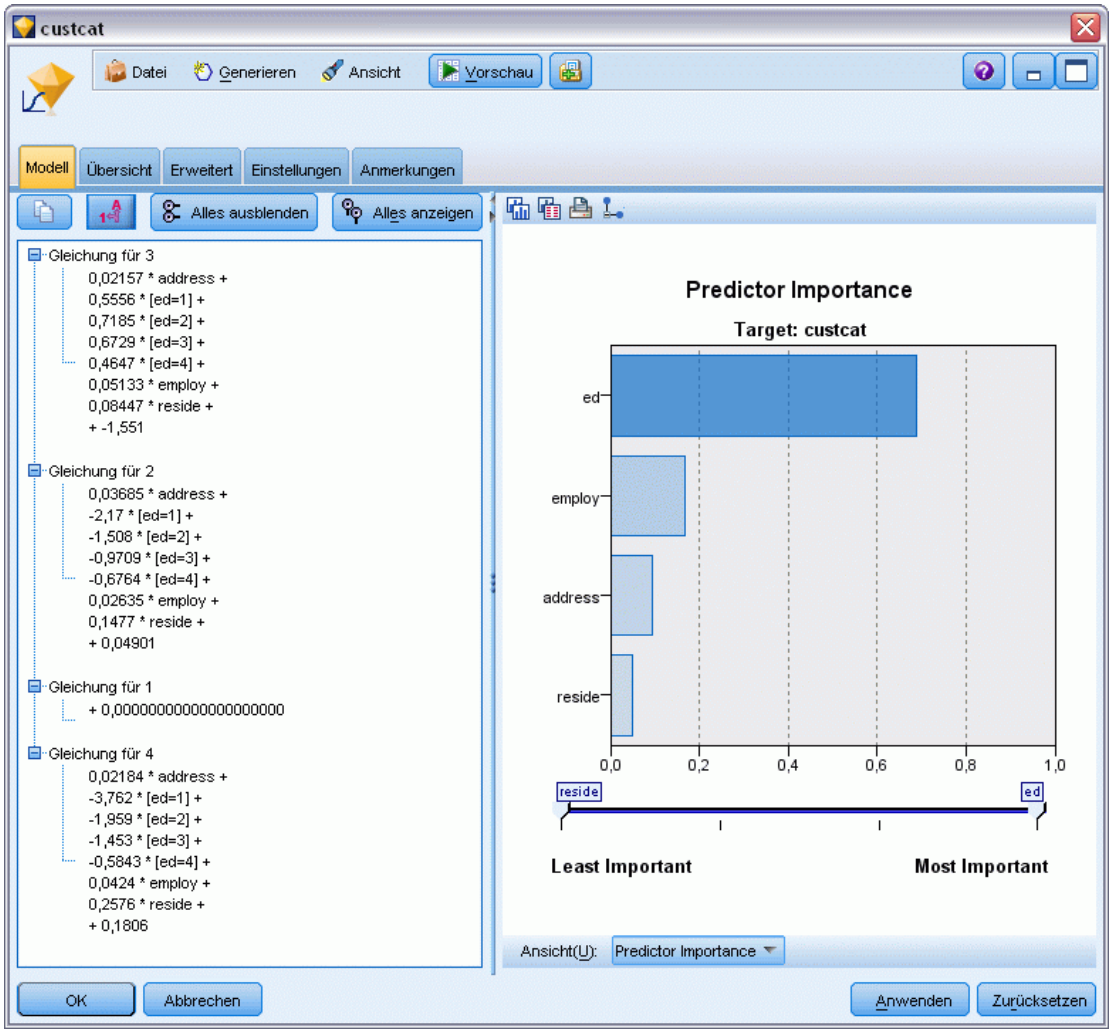
Durchsuchen des Modells

- Führen Sie den Knoten aus, um das Modell zu generieren; dieses wird zur Modellpalette in der rechten oberen Ecke hinzugefügt. Um die zugehörigen Details anzuzeigen, können Sie mit der rechten Maustaste auf den generierten Modellknoten klicken und Durchsuchen auswählen.

  Die Registerkarte “Modell” zeigt die Gleichungen an, die zur Zuweisung von Datensätzen zu den einzelnen Kategorien des Zielfelds verwendet werden. Es gibt vier mögliche Kategorien, eine davon ist die Basiskategorie, für die hier keine Gleichungsdetails angezeigt werden. Für die

übrigen drei Gleichungen werden Details angezeigt. Dabei steht Kategorie 3 für “Plus Service” (Plus-Service) usw.

Abbildung 13-7
Durchsuchen der Modellergebnisse



Auf der Registerkarte “Übersicht” werden (unter anderem) die Ziele und die Eingaben (Prädiktorfelder) angezeigt, die vom Modell verwendet werden. Beachten Sie, dass diese Felder tatsächlich anhand der Methode “Schrittweise” ausgewählt wurden und nicht anhand der vollständigen Liste, die zur Erwägung vorgelegt wurde.

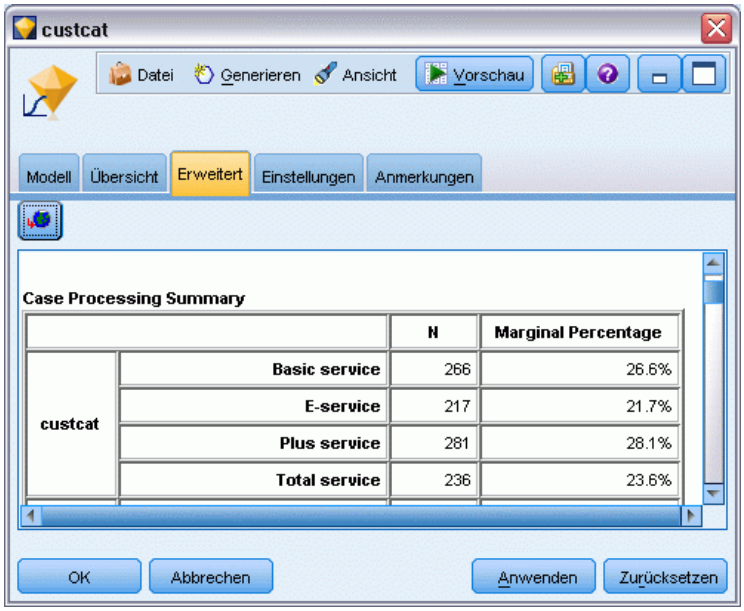
Abbildung 13-8
Modellübersicht mit Ziel- und Eingabefeldern

Die auf der Registerkarte “Erweitert” gezeigten Elemente hängen von den Optionen ab, die auf der Registerkarte “Erweiterte Ausgabe” im Modellierungsknoten ausgewählt wurden.

Ein Element, das immer angezeigt wird, ist die Fallverarbeitungsübersicht, die den Prozentsatz der Datensätze angibt, der jeweils auf die einzelnen Kategorien des Zielfelds entfällt. Auf diese Weise erhalten Sie ein Nullmodell, das Sie als Vergleichsgrundlage verwenden können.

Wenn kein Modell erstellt wurde, das Prädiktoren verwendet, wäre die naheliegendste Vorgehensweise, alle Kunden der am häufigsten vorkommenden Gruppe, also der Gruppe “Plus Service” (Plus-Service) zuzuweisen.

Abbildung 13-9
Zusammenfassung der Fallverarbeitung

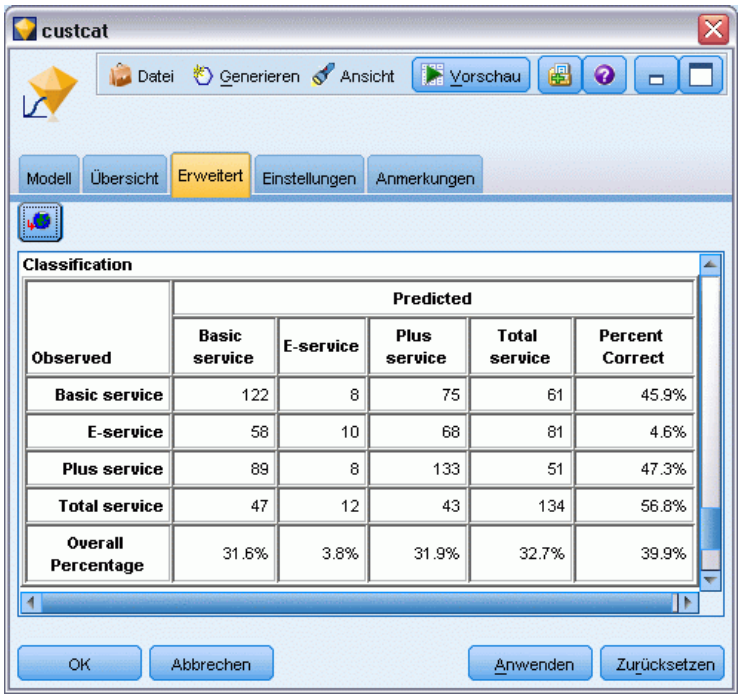


Auf der Grundlage der Trainingsdaten gilt: Wenn Sie alle Kunden dem Nullmodell zuweisen, erhalten Sie in 281 von 1.000 Fällen, also in 28,1 % der Fälle das richtige Ergebnis. Die Registerkarte “Erweitert” enthält weitere Informationen, mit denen Sie die Vorhersagen des Modells untersuchen können. Anschließend können Sie die Vorhersagen mit den Ergebnissen des Nullmodells vergleichen, um zu beurteilen, wie gut das Modell mit den vorliegenden Daten funktioniert.

Unten auf der Registerkarte “Erweitert” zeigt die Klassifikationstabelle die Ergebnisse für das Modell an, die in 39,9 % der Fälle korrekt sind.

Besonders erfolgreich ist das Modell bei der Ermittlung der Kunden, die sich für “Total Service” (Umfassender Service, Kategorie 4) entscheiden, es ist jedoch sehr schlecht bei der Ermittlung der E-Service-Kunden (Kategorie 2) Wenn Sie eine höhere Genauigkeit für Kunden in Kategorie 2 wünschen, müssen Sie einen anderen Prädiktor finden, mit dem sie besser ermittelt werden können.

Abbildung 13-10
Klassifikationsmatrix



Observed	Predicted				
	Basic service	E-service	Plus service	Total service	Percent Correct
Basic service	122	8	75	61	45.9%
E-service	58	10	68	81	4.6%
Plus service	89	8	133	51	47.3%
Total service	47	12	43	134	56.8%
Overall Percentage	31.6%	3.8%	31.9%	32.7%	39.9%

Je nachdem, was Sie vorhersagen möchten, kann das Modell für Ihre Zwecke auch hervorragend geeignet sein. Wenn Sie beispielsweise keinen Wert darauf legen, Kunden in Kategorie 2 zu ermitteln, kann das Modell für Sie genau genug sein. Dies kann dann der Fall sein, wenn es sich bei E-Service um ein Lockangebot handelt, das wenig Profit bringt.

Wenn die höchste Kapitalrendite (ROI) beispielsweise von Kunden herrührt, die in Kategorie 3 oder 4 fallen, bietet Ihnen das Modell möglicherweise die Informationen, die Sie benötigen.

Um einzuschätzen, wie gut das Modell tatsächlich an die Daten angepasst ist, stehen im Dialogfeld “Erweiterte Ausgabe” bei der Modellerstellung eine Reihe von Diagnosewerkzeugen zur Verfügung. Für weitere Informationen siehe Thema Logistik-Modell-Nugget – Erweiterte Ausgabe in Kapitel 10 in *IBM SPSS Modeler 14.2-Modellierungsknoten*.Erläuterungen der mathematischen Grundlagen für die in IBM® SPSS® Modeler verwendeten Modellierungsmethoden finden Sie im *SPSS Modeler-Algorithmushandbuch*, das sich im Verzeichnis *Documentation* des Installationsdatenträgers befindet.

Beachten Sie außerdem, dass diese Ergebnisse nur auf den Trainingsdaten beruhen. Um einzuschätzen, wie gut sich das Modell für andere Daten in der Praxis verallgemeinern lässt, könnten Sie mit einem Partitionsknoten eine Teilmenge der Datensätze für Test- und Validierungszwecke zurückhalten. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in *IBM SPSS Modeler 14.2- Quellen- , Prozess- und Ausgabeknoten*.

Kapitel
14

Kundenabwanderung bei Telekommunikationsunternehmen (binomiale logistische Regression)

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 Zielfelds ein kategoriales verwendet wird.

In diesem Beispiel wird ein Stream namens *telco_churn.str* verwendet, der Bezug nimmt auf die Datendatei *telco.sav*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *telco_churn.str* befindet sich im Verzeichnis *streams*.

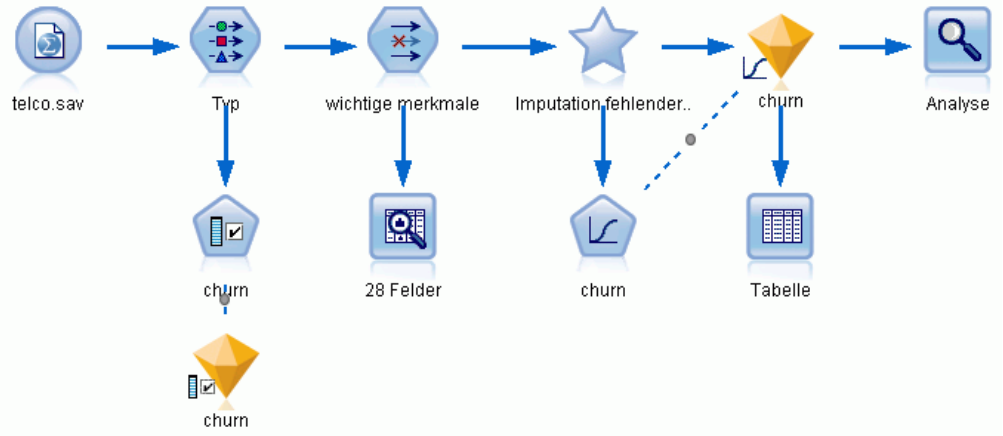
Hier ein Beispiel: Ein Telekommunikationsanbieter ist besorgt über die Anzahl an Kunden, die er an Mitbewerber verliert. Wenn Daten über die Servicenutzung verwendet werden können, um zu prognostizieren, welche Kunden mit hoher Wahrscheinlichkeit zu einem anderen Anbieter wechseln, können die Angebote entsprechend angepasst werden, um so viele Kunden wie möglich zu halten.

Dieses Beispiel konzentriert sich auf die Verwendung von Nutzungsdaten zur Vorhersage des Kundenverlusts (Abwanderung). Da das Ziel zwei verschiedene Kategorien aufweist, wird ein binomiales Modell verwendet. Bei einem Ziel mit mehreren Kategorien könnte stattdessen ein multinomiales Modell erstellt werden. Für weitere Informationen siehe Thema Klassifizieren von Kunden im Telekommunikationsbereich (multinomiale logistische Regression) in Kapitel 13 auf S. 154.

Erstellen des Streams

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

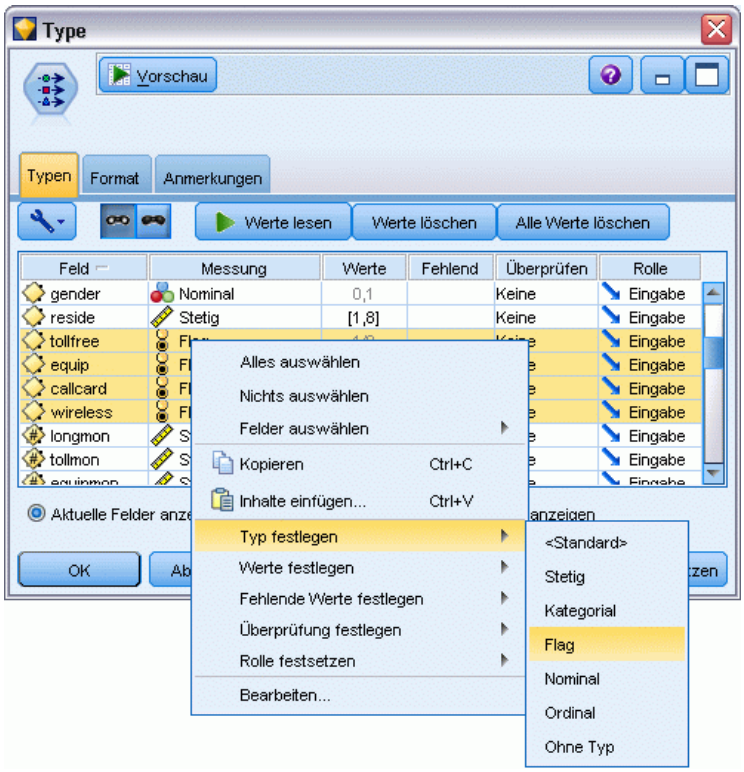
Abbildung 14-1
Beispiel-Stream zur Klassifizierung von Kunden mithilfe der binomialen logistischen Regression



- Fügen Sie einen Typknoten zur Definition von Feldern hinzu. Achten Sie dabei darauf, dass alle Messniveaus korrekt festgelegt werden. So können beispielsweise die meisten Felder mit den

Werten 0 und 1 als Flags betrachtet werden, manche Felder, wie beispielsweise das Geschlecht, sollten jedoch besser als nominales Feld mit zwei Werten betrachtet werden.

Abbildung 14-2
Festlegen des Messniveaus für mehrere Felder



Tipp: Um die Eigenschaften für mehrere Felder mit ähnlichen Werten (z. B. 0/1) zu ändern, klicken Sie auf die Überschrift der Spalte *Werte* (um die Felder nach ihrem Wert zu sortieren) und halten Sie anschließend die Umschalttaste gedrückt, während Sie mit der Maus oder den Pfeiltasten alle Felder auswählen, die geändert werden sollen. Anschließend können Sie mit der rechten Maustaste auf die Auswahl klicken, um das Messniveau oder andere Attribute der ausgewählten Felder zu ändern.

- Setzen Sie das Messniveau für das Feld *churn* (Abwanderung) auf Flag und setzen Sie die Rolle auf Ziel. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

Abbildung 14-3
Festlegen von Messniveau und Rolle für das Feld "churn"

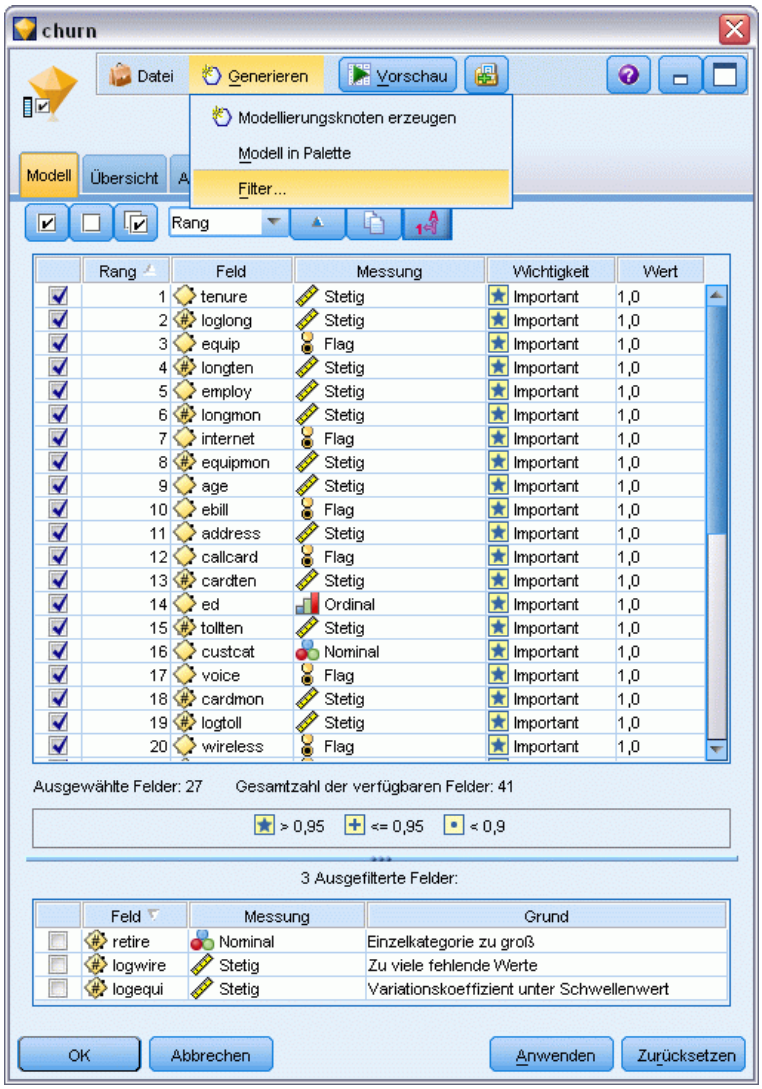
- Fügen Sie dem Typknoten einen Merkmalsauswahl-Modellierungsknoten hinzu.

  Mit einem Merkmalsauswahlknoten können Sie Prädiktoren bzw. Daten entfernen, die hinsichtlich der Beziehung zwischen Prädiktor und Ziel keine nützlichen Informationen hinzufügen.

- Führen Sie den Stream aus.

▶ Öffnen Sie das Ergebnis-Modell-Nugget und wählen Sie aus dem Menü Generieren die Option Filter, um einen Filterknoten zu erstellen.

Abbildung 14-4
Generieren eines Filterknotens aus einem Merkmalsauswahlknoten



Nicht alle Daten in der Datei *telco.sav* sind für die Vorhersage der Abwanderung von Nutzen. Verwenden Sie den Filter, um nur die Daten auszuwählen, die als wichtig für die Verwendung als Prädiktor erachtet werden.

▶ Wählen Sie im Dialogfeld “Filter generieren” die Option Alle Felder, die markiert sind als: Bedeutsam und klicken Sie auf OK.

- Verbinden Sie den generierten Filterknoten mit dem Typknoten.

Abbildung 14-5
Auswählen bedeutsamer Felder

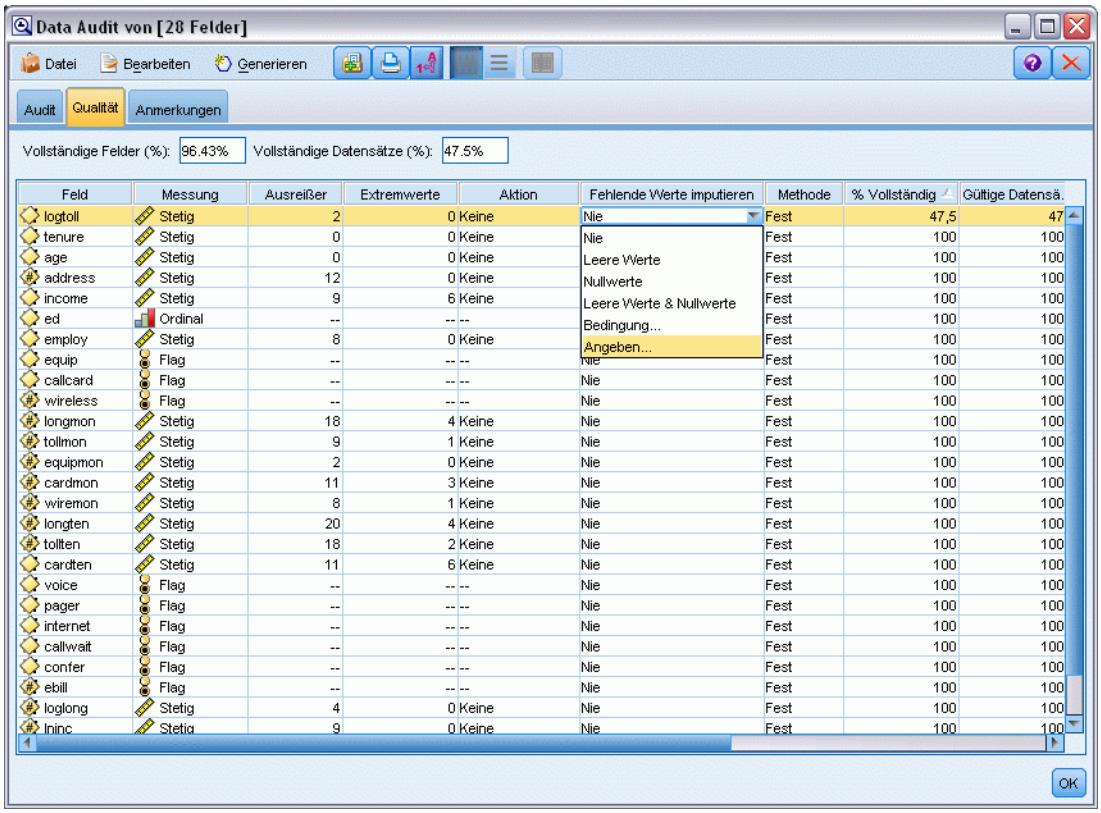
- Fügen Sie einen Data Audit-Knoten zum generierten Filterknoten hinzu.

  Öffnen Sie den Data Audit-Knoten und klicken Sie auf Ausführen.

- Klicken Sie auf der Registerkarte "Qualität" des Data Audit-Browsers auf die Spalte *% Vollständig*, um sie in aufsteigender numerischer Reihenfolge zu sortieren. Dadurch können Sie alle Felder ermitteln, die große Mengen fehlender Daten enthalten; in diesem Fall müssen Sie lediglich das Feld *logtoll* bearbeiten, das zu weniger als 50 % vollständig ist.

▶ Klicken Sie in der Spalte *Fehlende Werte imputieren* für *logtoll* auf Angeben.

Abbildung 14-6
Imputieren fehlender Werte für "logtoll"



▶ Wählen Sie für Imputieren, wenn die Option Leere Werte und Nullwerte. Wählen Sie für Festgelegt als die Option Mittelwert und klicken Sie auf OK.

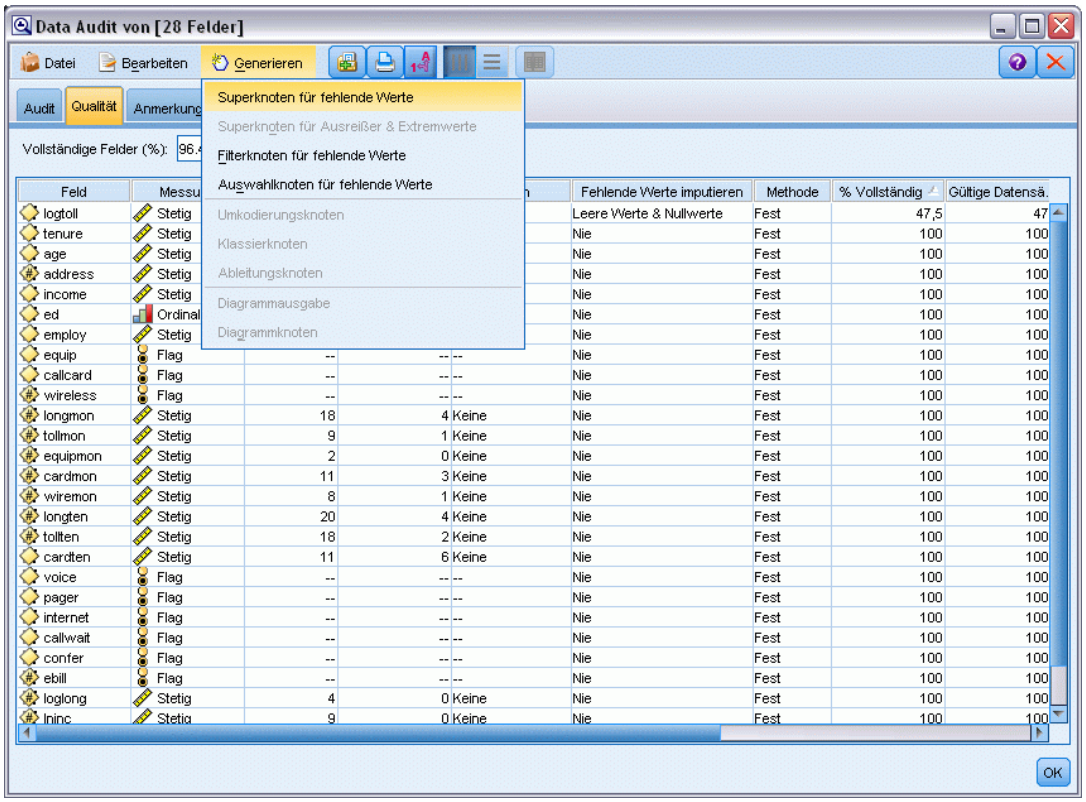
Die Auswahl von Mittelwert gewährleistet, dass die imputierten Werte keinen negativen Einfluss auf den Mittelwert aller Werte in den Daten insgesamt haben.

Abbildung 14-7
Auswahl der Imputationseinstellungen

- Generieren Sie im Data Audit-Browser auf der Registerkarte “Qualität” den Superknoten für fehlende Werte. Wählen Sie hierzu die folgenden Befehle aus den Menüs aus:

  Erzeugen > Superknoten für fehlende Werte

Abbildung 14-8
Generieren eines Superknotens für fehlende Werte

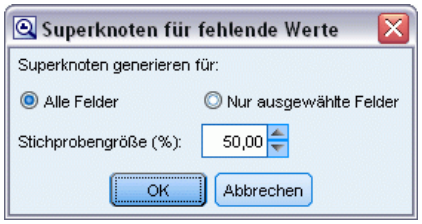


Erhöhen Sie im Dialogfeld “Superknoten für fehlende Werte” den Wert für Stichprobengröße auf 50 % und klicken Sie auf OK.

Der Superknoten wird im Stream-Zeichenbereich angezeigt und trägt den Titel: *Imputation fehlender Werte*.

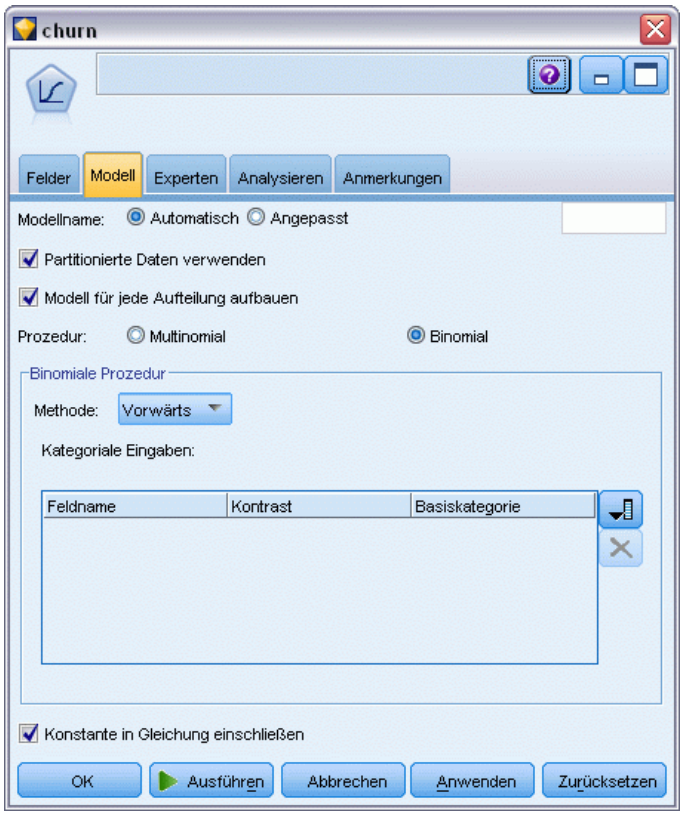
- Verbinden Sie den Superknoten mit dem Filterknoten.

Abbildung 14-9
Angeben des Stichprobenumfangs



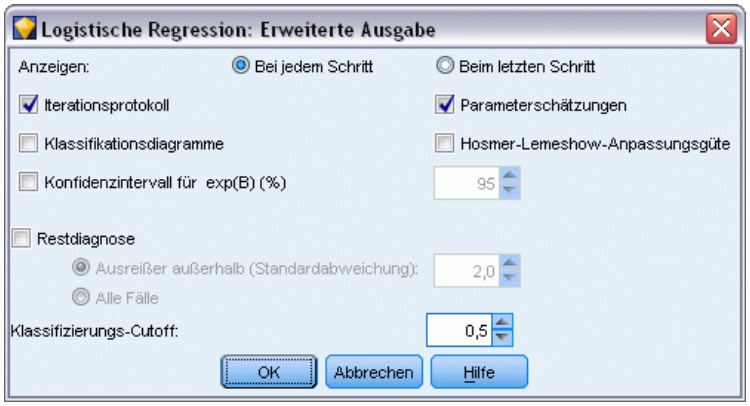
- Fügen Sie dem Superknoten einen Logistikknoten hinzu.
- Klicken Sie im Logistikknoten auf die Registerkarte “Modell” und wählen Sie die Prozedur Binomial aus. Wählen Sie im Bereich *Binomiale Prozedur* die Methode Vorwärts aus.

Abbildung 14-10
Auswählen Modelloptionen



- Wählen Sie auf der Registerkarte “Experten” den Modus Experten und klicken Sie dann auf Ausgabe. Das Dialogfeld “Erweiterte Ausgabe” wird angezeigt.
- Wählen Sie im Dialogfeld “Erweiterte Ausgabe” die Option Bei jedem Schritt als Typ für *Anzeige*. Wählen Sie Iterationsprotokoll und Parameterschätzungen und klicken Sie auf OK.

Abbildung 14-11
Auswahl der Ausgabeoptionen



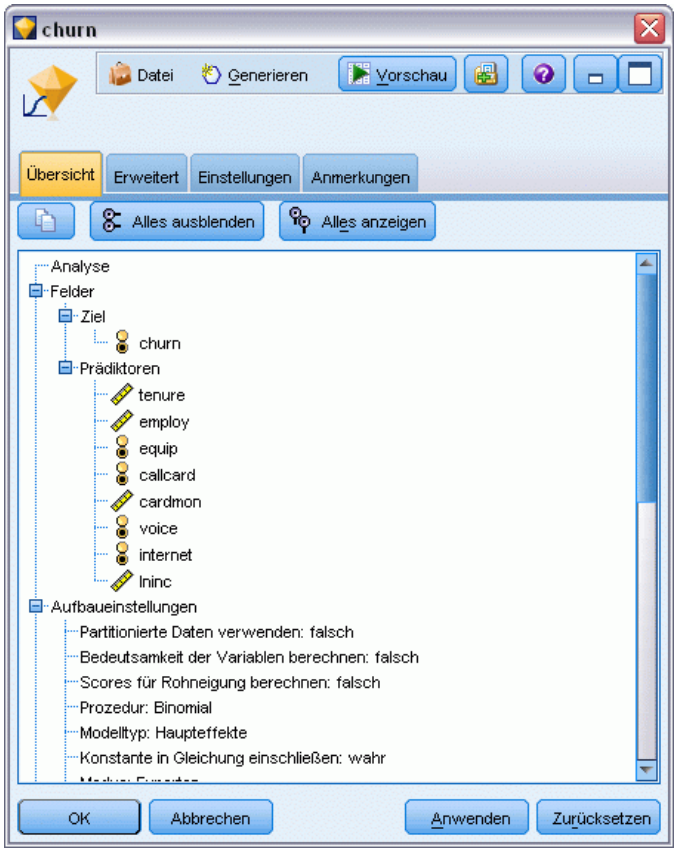
Durchsuchen des Modells

- Klicken Sie auf dem Logistikknoten auf Ausführen, um das Modell zu erstellen.

  Das generierte Modell-Nugget wird dem Stream-Zeichenbereich und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die zugehörigen Details anzuzeigen, klicken Sie mit der rechten Maustaste auf das Modell-Nugget und wählen Bearbeiten oder Durchsuchen.

Auf der Registerkarte “Übersicht” werden (unter anderem) die Ziele und die Eingaben (Prädiktorfelder) angezeigt, die vom Modell verwendet werden. Beachten Sie, dass diese Felder tatsächlich anhand der Methode “Vorwärts” ausgewählt wurden und nicht anhand der vollständigen Liste, die zur Erwägung vorgelegt wurde.

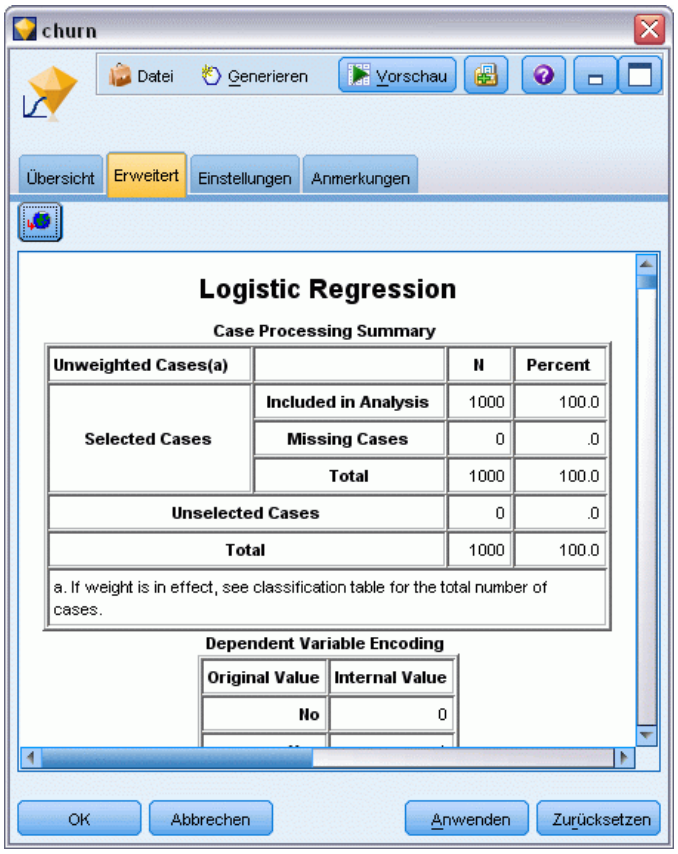
Abbildung 14-12
Modellübersicht mit Ziel- und Eingabefeldern



Die auf der Registerkarte “Erweitert” gezeigten Elemente hängen von den Optionen ab, die auf der Registerkarte “Erweiterte Ausgabe” im Logistikknoten ausgewählt wurden. Ein Element, das immer angezeigt wird, ist die Fallverarbeitungsübersicht, die die Anzahl und den Prozentsatz der in der Analyse einbezogenen Datensätze anzeigt. Außerdem wird ggf. die Anzahl der fehlenden

Fälle aufgeführt, bei denen ein oder mehrere Eingabefelder nicht verfügbar sind, und alle Fälle, die nicht ausgewählt wurden.

Abbildung 14-13
Zusammenfassung der Fallverarbeitung

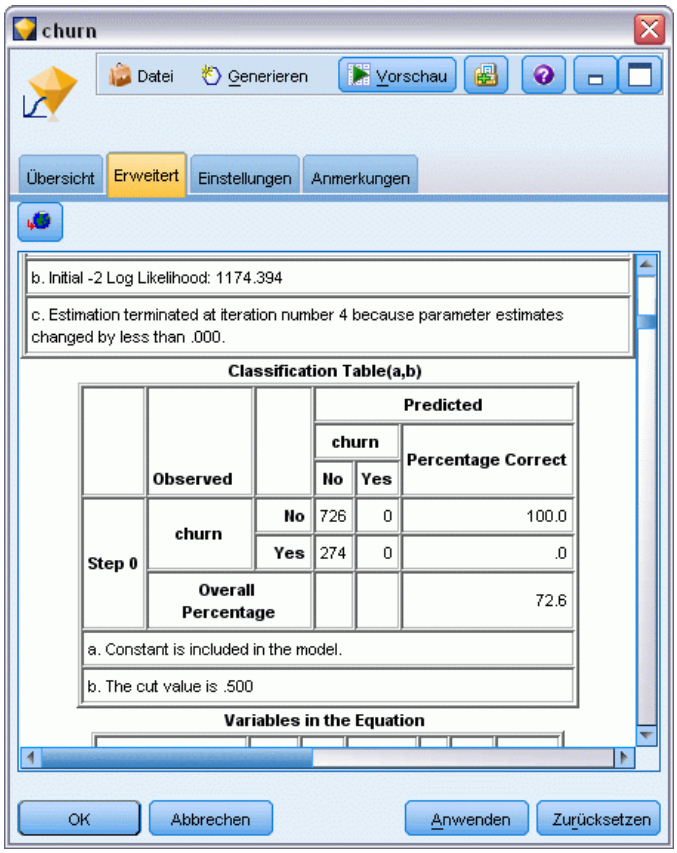


▶ Führen Sie in der Fallverarbeitungsübersicht einen Bildlauf nach unten durch, um die Klassifikationstabelle unter Block 0: Anfangsblock, anzuzeigen.

Die Methode “Schrittweise vorwärts” beginnt mit einem Nullmodell, also einem Modell ohne Prädiktoren, das als Grundlage für den Vergleich mit dem endgültig erstellten Modell verwendet werden kann. Das Nullmodell sagt laut Konvention alles als 0 voraus. Das Nullmodell weist somit eine Genauigkeit von 72,6 % auf, da die 726 Kunden, die nicht abgewandert sind, korrekt

vorausgesagt wurden. Die Kunden, die abwanderten, wurden jedoch überhaupt nicht richtig vorhergesagt.

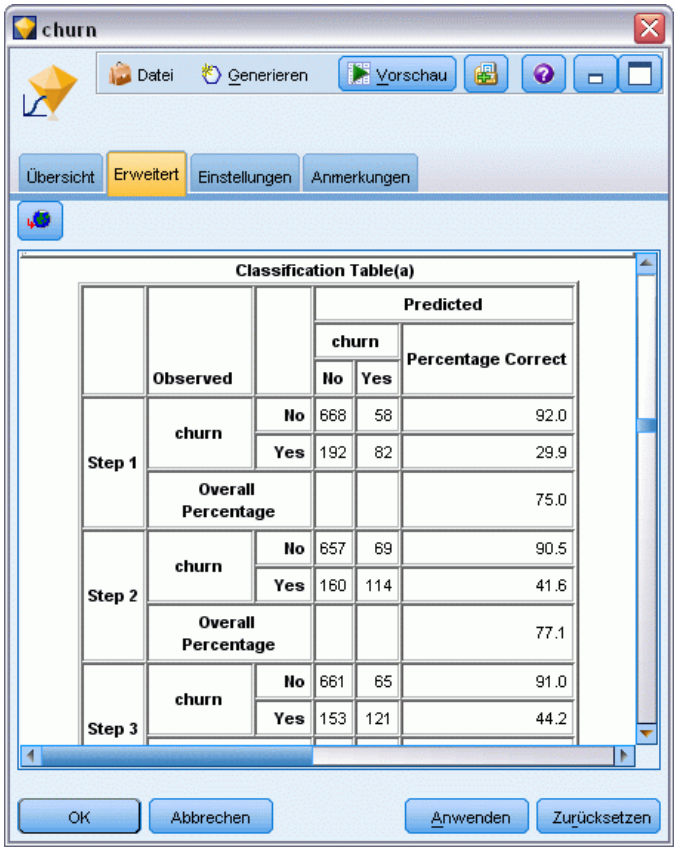
Abbildung 14-14
Anfangsklassifikationstabelle – Block 0



- Führen Sie nun einen Bildlauf nach unten durch, um die Klassifikationstabelle unter Block 1 anzuzeigen: Methode = Vorwärts schrittweise.

Diese Klassifikationstabelle zeigt die Ergebnisse für das Modell an, während in jedem Schritt ein Prädiktor hinzugefügt wird. Bereits im ersten Schritt – nach Verwendung eines einzigen Prädiktors – hat das Modell die Genauigkeit für die Abwanderungsvorhersage von 0,0 % auf 29,9 % gesteigert.

Abbildung 14-15
Klassifikationstabelle – Block 1



Classification Table(a)

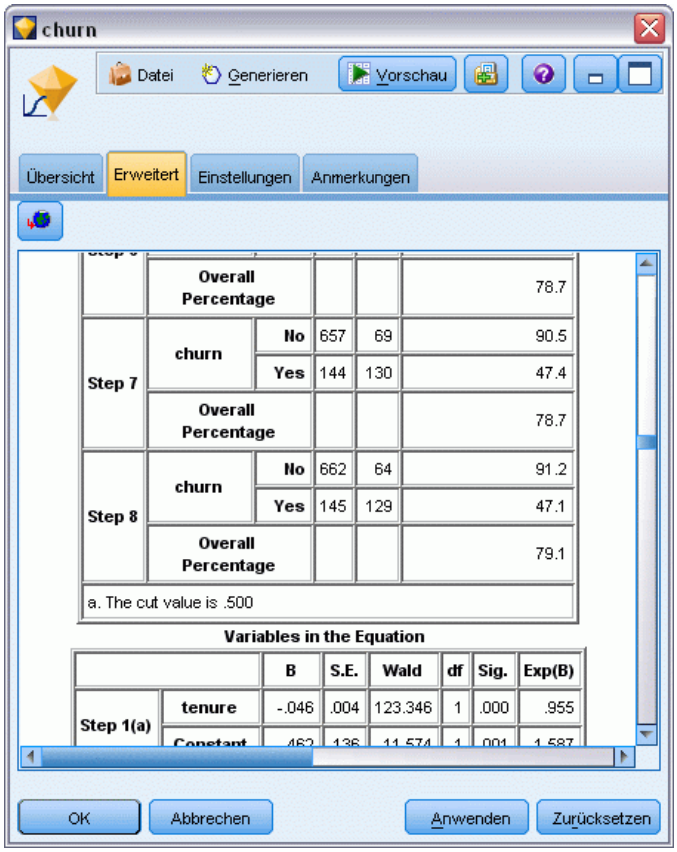
	Observed		Predicted		
			churn		Percentage Correct
			No	Yes	
Step 1	churn	No	668	58	92.0
		Yes	192	82	29.9
	Overall Percentage				75.0
Step 2	churn	No	657	69	90.5
		Yes	160	114	41.6
	Overall Percentage				77.1
Step 3	churn	No	661	65	91.0
		Yes	153	121	44.2

▶ Führen Sie einen Bildlauf zum Ende dieser Klassifiktionstabelle durch.

Die Klassifikationstabelle zeigt, dass der letzte Schritt Schritt 8 ist. In dieser Phase hat der Algorithmus ermittelt, dass keine weiteren Prädiktoren mehr in das Modell aufgenommen werden müssen. Die Genauigkeit für die Kunden, die nicht abwandern, ist zwar leicht gesunken (auf 91,2 %), im Gegenzug ist jedoch die Vorhersagegenauigkeit für die Kunden, die abwandern,

von den ursprünglichen 0 % auf 47,1 % gestiegen. Dies stellt eine signifikante Verbesserung gegenüber dem ursprünglichen Nullmodell dar, bei dem keine Prädiktoren verwendet wurden.

Abbildung 14-16
Klassifikationstabelle – Block 1



Beim Ziel, die Abwanderung zu reduzieren, wäre die Möglichkeit, sie um annähernd die Hälfte zu reduzieren, ein großer Schritt in Richtung der Aufrechterhaltung der Einkommensströme.

Hinweis: Dieses Beispiel zeigt auch, dass die Verwendung des Gesamtprozentsatzes als Richtschnur für die Genauigkeit eines Modells in einigen Fällen irreführend sein kann. Das ursprüngliche Nullmodell wies eine Gesamtgenauigkeit von 72,6 % auf, das endgültige vorhergesagte Modell weist eine Gesamtgenauigkeit von 79,1 % auf; wie wir jedoch gesehen haben, unterscheiden sich die beiden Modelle deutlich hinsichtlich der Genauigkeit in den einzelnen Kategorien.

Um einzuschätzen, wie gut das Modell tatsächlich an die Daten angepasst ist, stehen im Dialogfeld "Erweiterte Ausgabe" bei der Modellerstellung eine Reihe von Diagnosewerkzeugen zur Verfügung. Für weitere Informationen siehe Thema Logistik-Modell-Nugget – Erweiterte Ausgabe in Kapitel 10 in *IBM SPSS Modeler 14.2-Modellierungsknoten*.Erläuterungen der mathematischen Grundlagen für die in IBM® SPSS® Modeler verwendeten Modellierungsmethoden finden Sie im *SPSS Modeler-Algorithmushandbuch*, das sich im Verzeichnis *Documentation* des Installationsdatenträgers befindet.

Beachten Sie außerdem, dass diese Ergebnisse nur auf den Trainingsdaten beruhen. Um einzuschätzen, wie gut sich das Modell für andere Daten in der Praxis verallgemeinern lässt, könnten Sie mit einem Partitionsknoten eine Teilmenge der Datensätze für Test- und Validierungszwecke zurückhalten. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in *IBM SPSS Modeler 14.2- Quellen- , Prozess- und Ausgabeknoten*.

Kapitel
15

Vorhersage der Bandbreitennutzung (Zeitreihen)

Prognoseerstellung mit dem Zeitreihenknoten

Ein Analyst eines Breitband-Providers soll eine Prognose über die Vertragsabschlüsse mit Kunden erstellen, um die Nutzung der Bandbreite vorhersagen zu können. 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 von lokalen Märkten erstellen. Ein zweites Beispiel zeigt, wie Sie Datenquellen konvertieren können, wenn sie nicht im richtigen Format für die Eingabe in den Zeitreihenknoten vorliegen.

In diesen Beispielen wird ein Stream namens *broadband_create_models.str* verwendet, der Bezug nimmt auf die Datendatei *broadband_1.sav*. Die Dateien stehen im Ordner *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *broadband_create_models.str* befindet sich im Ordner *streams*.

Das letzte Beispiel zeigt, wie die gespeicherten Modelle auf ein aktualisiertes Daten-Set angewendet werden können, um die Vorhersagen um weitere drei Monate auszuweiten.

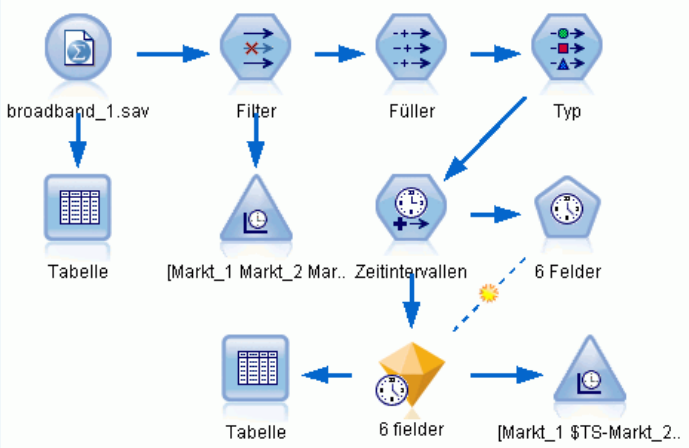
In SPSS Modeler können Sie mehrere Zeitreihenmodelle in einem einzelnen Vorgang erstellen. Die Quellendatei, die Sie verwenden, weist Zeitreihendaten für 85 verschiedene Märkte auf; der Einfachheit halber führen Sie die Modellierung jedoch nur für fünf dieser Märkte und für einen Gesamtwert für alle Märkte durch.

Die Datendatei *broadband_1.sav* enthält die monatlichen Nutzungsdaten für 85 lokale Märkte. Für dieses Beispiel werden nur die ersten fünf Zeitreihen verwendet. Für jede der fünf Zeitreihen sowie für die Gesamtmenge wird jeweils ein gesondertes Modell erstellt.

Außerdem enthält die Datei ein Datumsfeld, in dem für jeden Datensatz Monat und Jahr angegeben sind. Dieses Feld wird im Zeitintervallknoten zur Beschriftung der Datensätze verwendet. Das Datumsfeld wird als Zeichenkette in SPSS Modeler eingelesen. Um das Feld

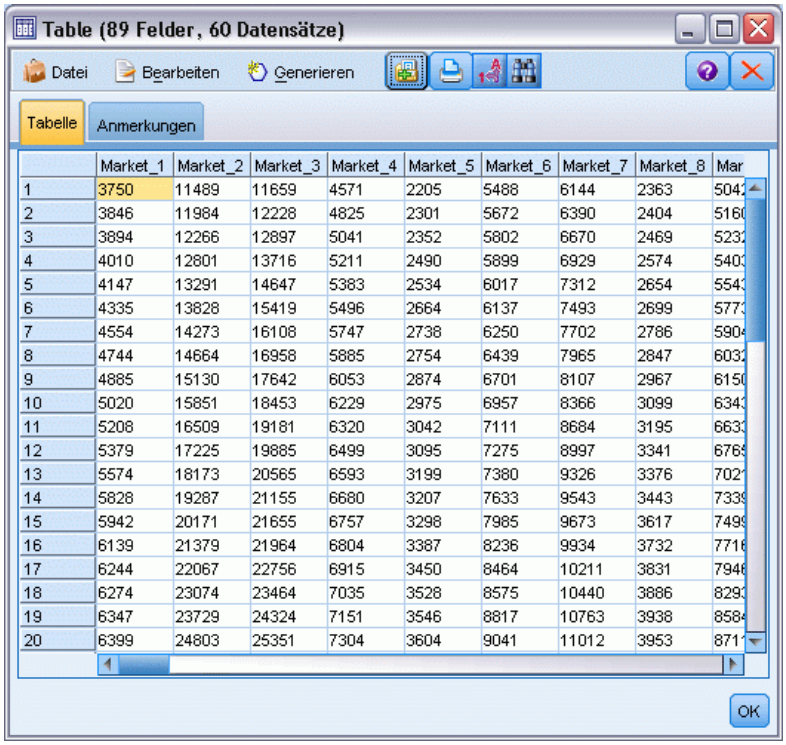
in SPSS Modeler verwenden zu können, müssen Sie jedoch den Speichertyp mithilfe eines Füllerknotens in das numerische Datumsformat konvertieren.

Abbildung 15-1
Beispiel-Stream zur Anzeige der Zeitreihenmodellierung



Für den Zeitreihenknoten ist es erforderlich, dass sich jede Zeitreihe in einer separaten Spalte befindet und dabei jeweils eine Zeile für jedes Intervall vorliegt. SPSS Modeler bietet Methoden, mit denen die Daten, falls erforderlich, in dieses Format umgewandelt werden können.

Abbildung 15-2
Monatliche Abonnementdaten für lokale Breitbandmärkte



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	504
2	3846	11984	12228	4825	2301	5672	6390	2404	516
3	3894	12266	12897	5041	2352	5802	6670	2469	523
4	4010	12801	13716	5211	2490	5899	6929	2574	540
5	4147	13291	14647	5383	2534	6017	7312	2654	554
6	4335	13828	15419	5496	2664	6137	7493	2699	577
7	4554	14273	16108	5747	2738	6250	7702	2786	590
8	4744	14664	16958	5885	2754	6439	7965	2847	603
9	4885	15130	17642	6053	2874	6701	8107	2967	615
10	5020	15851	18453	6229	2975	6957	8366	3099	634
11	5208	16509	19181	6320	3042	7111	8684	3195	663
12	5379	17225	19885	6499	3095	7275	8997	3341	676
13	5574	18173	20565	6593	3199	7380	9326	3376	702
14	5828	19287	21155	6680	3207	7633	9543	3443	733
15	5942	20171	21655	6757	3298	7985	9673	3617	749
16	6139	21379	21964	6804	3387	8236	9934	3732	771
17	6244	22067	22756	6915	3450	8464	10211	3831	794
18	6274	23074	23464	7035	3528	8575	10440	3886	829
19	6347	23729	24324	7151	3546	8817	10763	3938	858
20	6399	24803	25351	7304	3604	9041	11012	3953	871

Erstellen des Streams

- Erstellen Sie einen neuen Stream und fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *broadband_1.sav* verweist.

- Verwenden Sie einen Filterknoten, um die Felder *Market_6* bis *Market_85* und die Felder *MONTH_* und *YEAR_* zu entfernen und das Modell so zu vereinfachen.

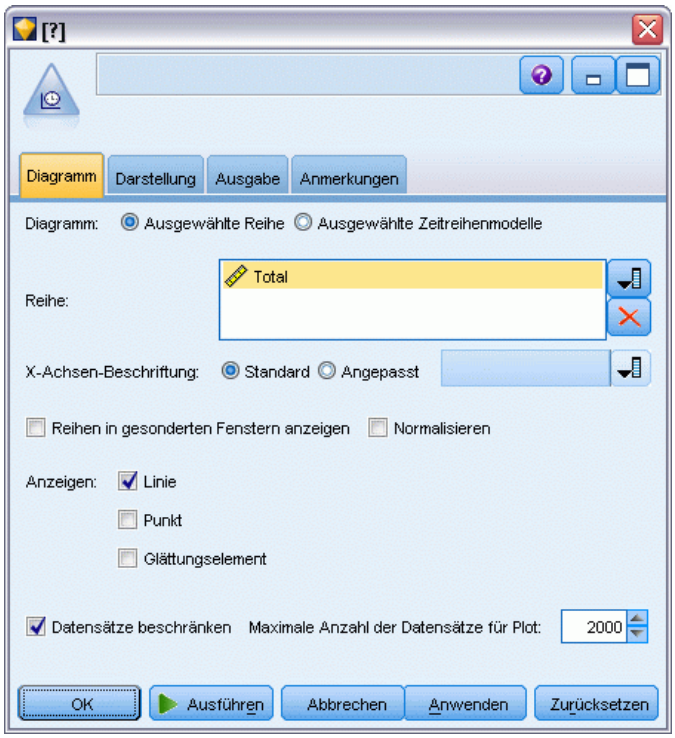
  Tipp: Um mehrere nebeneinander liegende Felder auf einmal auszuwählen, klicken Sie auf das Feld *Market_6*, halten Sie die linke Maustaste gedrückt und ziehen Sie die Maus nach unten bis zum Feld *Market_85*. Die ausgewählten Felder sind blau hervorgehoben. Um die anderen Felder hinzuzufügen, halten Sie die Strg-Taste gedrückt und klicken Sie auf die Felder *MONTH_* und *YEAR_*.

Abbildung 15-3
Vereinfachen des Modells

Untersuchen der Daten

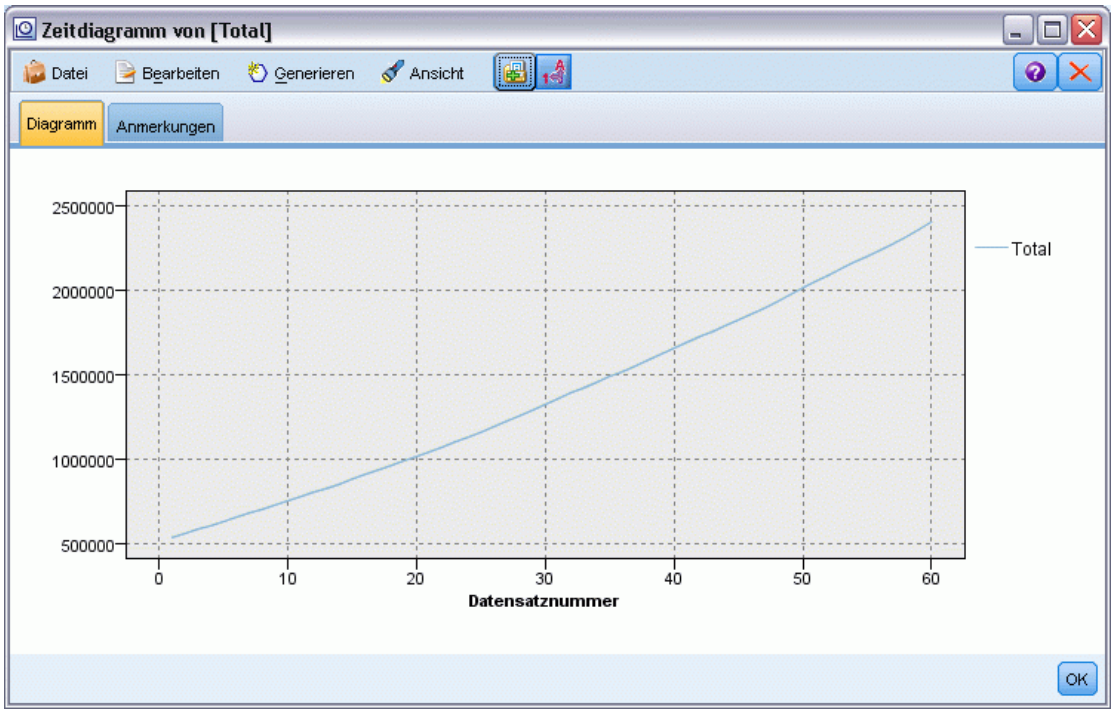
Es empfiehlt sich grundsätzlich, ein Gefühl für die Natur der Daten zu entwickeln, bevor Sie das Modell erstellen. Weisen die Daten saisonale Schwankungen auf? Der Expert Modeler kann zwar automatisch das beste saisonale oder nichtsaisonale Modell für jede Zeitreihe ermitteln, Sie können jedoch häufig schnellere Ergebnisse erzielen, indem Sie die Suche auf nichtsaisonale Modelle beschränken, wenn keine Saisonalität in den Daten vorliegt. Ohne eine Untersuchung der Daten für jeden der lokalen Märkte können wir uns ein grobes Bild darüber verschaffen, ob Saisonalität vorliegt oder nicht, indem wir die Gesamtzahl der Abonnenten in allen fünf Märkten plotten.

Abbildung 15-4
Plotten der Gesamtzahl an Abonnenten



- Gliedern Sie auf der Diagrammpalette einen Zeitdiagrammknoten an den Filterknoten an.
- Fügen Sie das Feld *Total* (Gesamt) zur Liste “Series” (Reihe) hinzu.
- Deaktivieren Sie die Kontrollkästchen Reihen in gesonderten Fenstern anzeigen und Normalisieren.
- Klicken Sie auf Ausführen.

Abbildung 15-5
Zeitdiagramm des Felds "Total" (Gesamt)

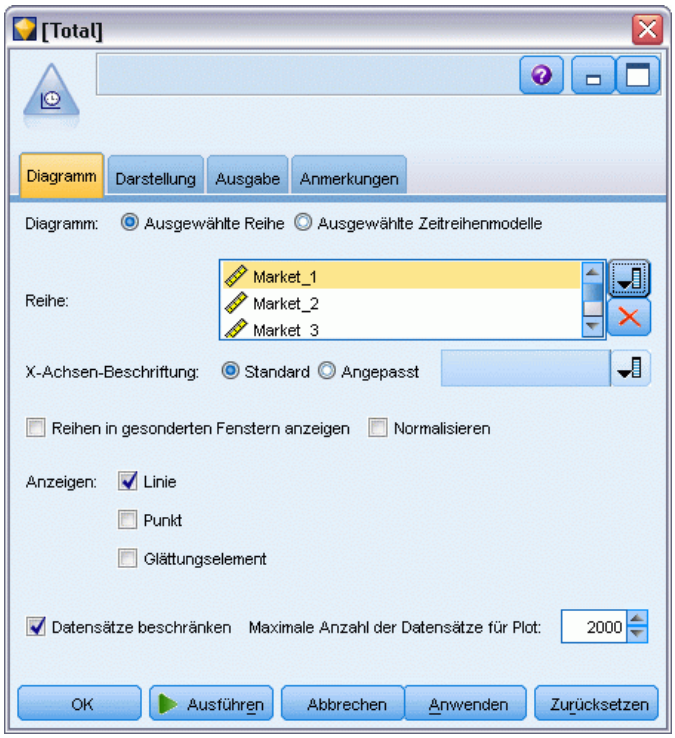


Die Zeitreihe zeigt einen sehr gleichmäßigen Aufwärtstrend ohne Anzeichen für saisonale Variationen. Möglicherweise weisen einzelne Zeitreihen Saisonalität auf, jedoch scheint die Saisonalität im Allgemeinen kein ausgeprägtes Merkmal der Daten zu sein.

Selbstverständlich müssen Sie jede der Zeitreihen untersuchen, bevor Sie saisonale Modelle ausschließen. Sie können dann die Zeitreihen aussondern, die Saisonalität aufweisen, und diese separat modellieren.

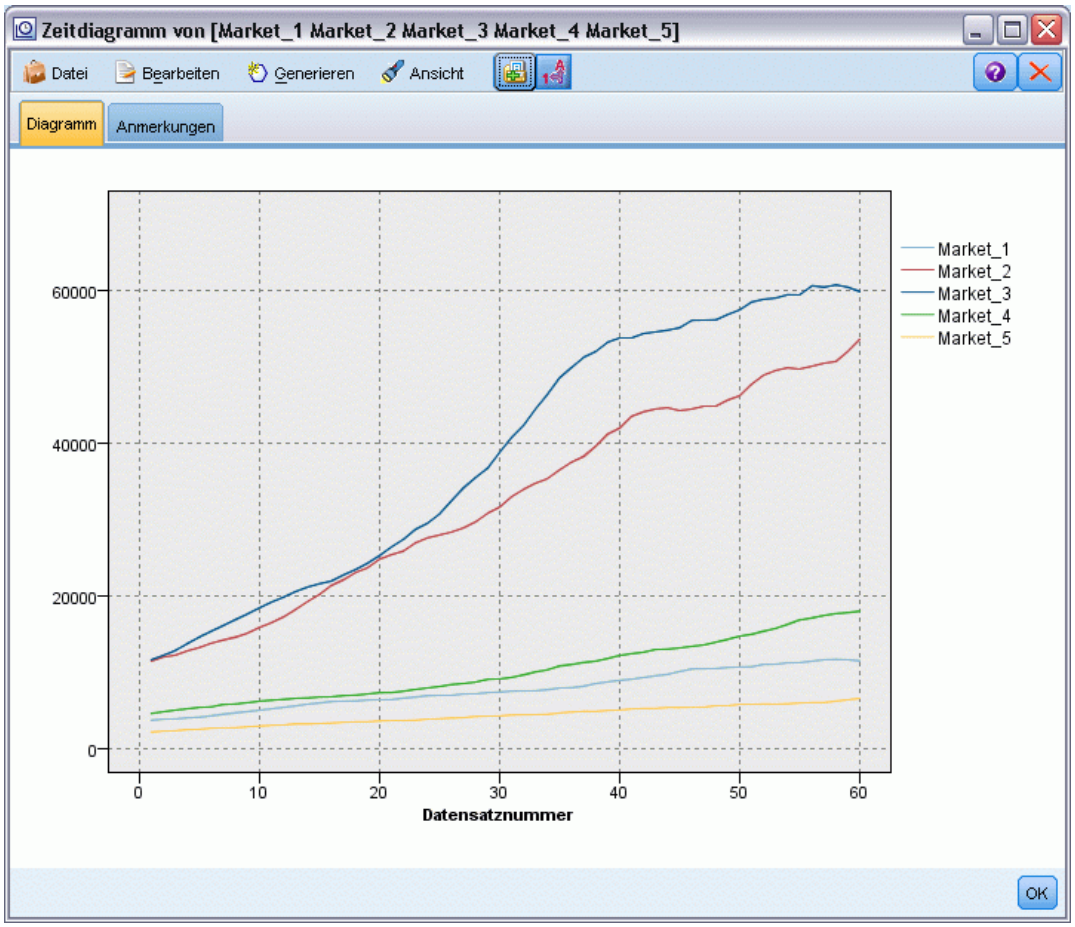
Mit IBM® SPSS® Modeler ist es einfach, mehrere Zeitreihen gemeinsam zu plotten.

Abbildung 15-6
Plotten mehrerer Zeitreihen



- Öffnen Sie den Zeitdiagrammknoten erneut.
- Entfernen Sie das Feld *Total* (Gesamt) aus der Liste “Series” (Reihe) (wählen Sie es aus und klicken Sie auf das rote X).
- Fügen Sie die Felder *Market_1* bis *Market_5* zur Liste hinzu.
- Klicken Sie auf Ausführen.

Abbildung 15-7
Zeitdiagramm mehrerer Felder



Die Untersuchung der einzelnen Märkte ergibt jeweils einen stetigen Aufwärtstrend. Einige Märkte sind zwar ein wenig unregelmäßiger als andere, es sind jedoch keine Anzeichen für Saisonalität zu beobachten.

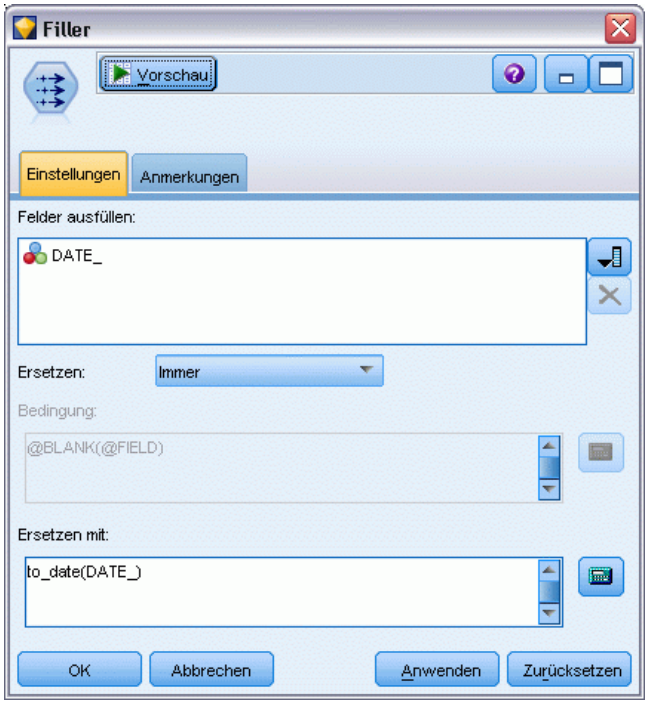
Definieren der Datumswerte

Nun müssen Sie den Speichertyp des Felds *DATE_* in das Datumsformat ändern.

- Gliedern Sie einen Füllerknoten an den Filterknoten an.
- Öffnen Sie den Füllerknoten und klicken Sie auf die Feldauswahlschaltfläche.
- Wählen Sie DATE_ aus und fügen Sie das Feld zu Felder ausfüllen hinzu.
- Setzen Sie die Bedingung Ersetzen auf Immer.

- Setzen Sie den Wert von Ersetzen mit auf to_date(DATE_).

Abbildung 15-8
Einrichten des Datenspeichertyps

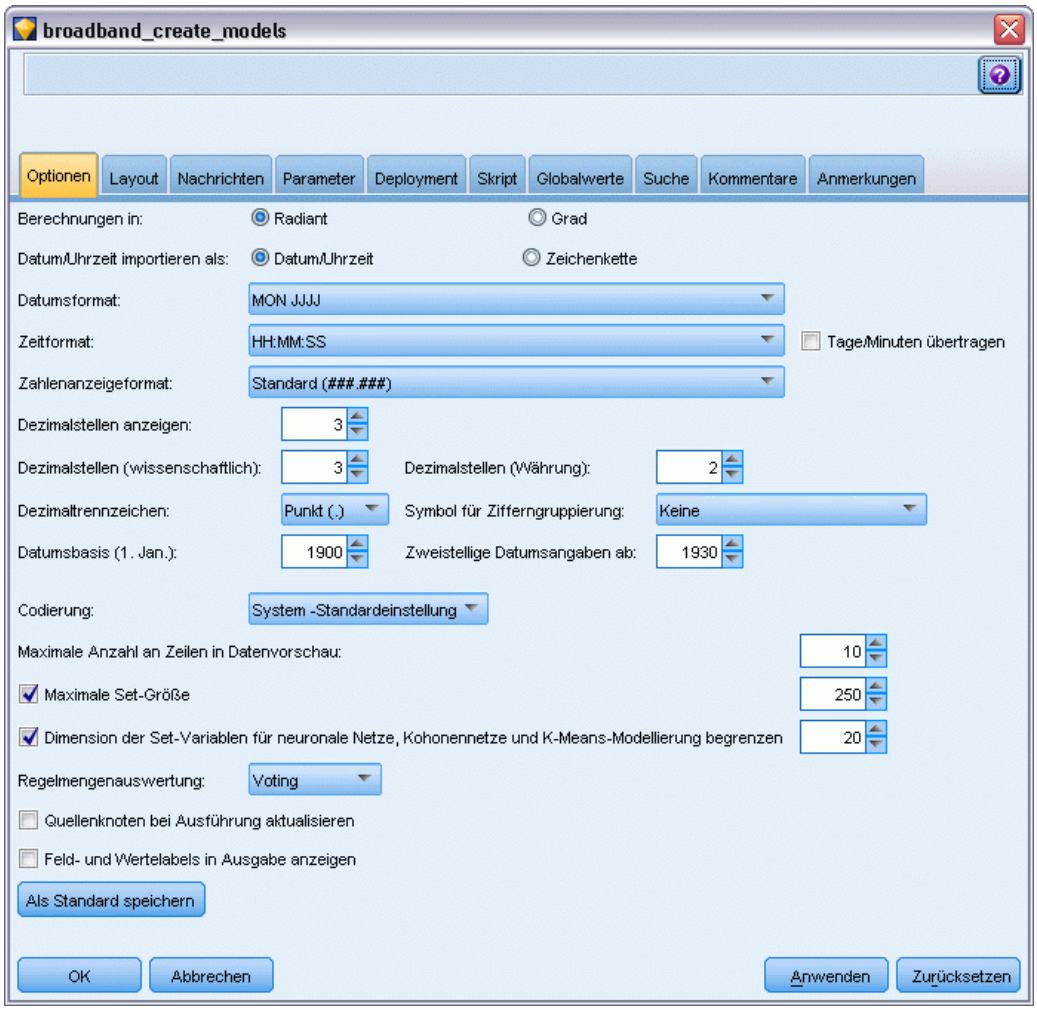


Ändern Sie das standardmäßige Datumsformat so, dass es mit dem Format des Datumsfelds übereinstimmt. Dies ist erforderlich, damit die Konvertierung des Datumsfelds wie erwartet funktioniert.

- Wählen Sie im Menü die Optionsfolge Tools > Stream-Eigenschaften > Optionen aus, um das Dialogfeld für die Stream-Optionen anzuzeigen.

▶ Setzen Sie den Standardwert für Datumsformat auf MON JJJJ .

Abbildung 15-9
Einrichten des Datumsformats



Definieren der Ziele

▶ Fügen Sie einen Typknoten hinzu und setzen Sie die Rolle für das Feld *DATE_*auf Keine. Setzen Sie die Rolle für alle anderen Felder (die Felder *Market_n* und das Feld *Total* (Gesamt)) auf Ziel.

- Klicken Sie auf die Schaltfläche Werte lesen, um die Spalte “Werte” auszufüllen.

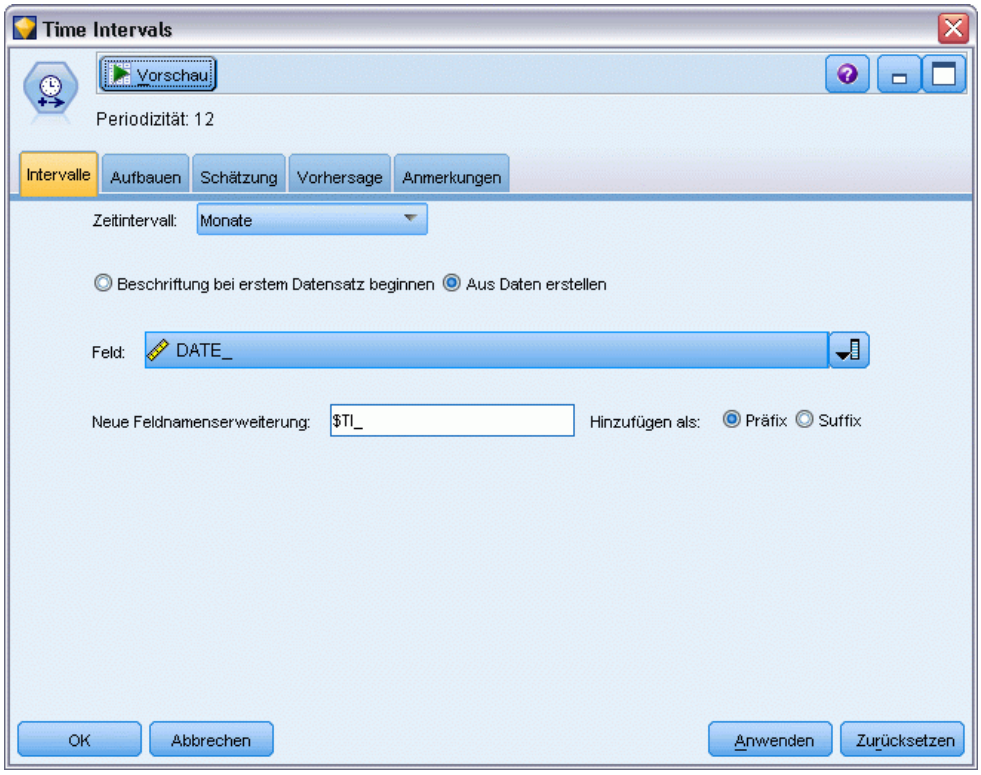
Abbildung 15-10
Festlegen der Rolle für mehrere Felder

Festlegen der Zeitintervalle

- Fügen Sie einen Zeitintervallknoten (aus der Palette “Feldfunktionen) hinzu.
- Wählen Sie auf der Registerkarte “Intervalle” als Zeitintervall die Option Monate.
- Wählen Sie die Option Aus Daten erstellen.

- Wählen Sie DATE_ als Erstellungsfeld.

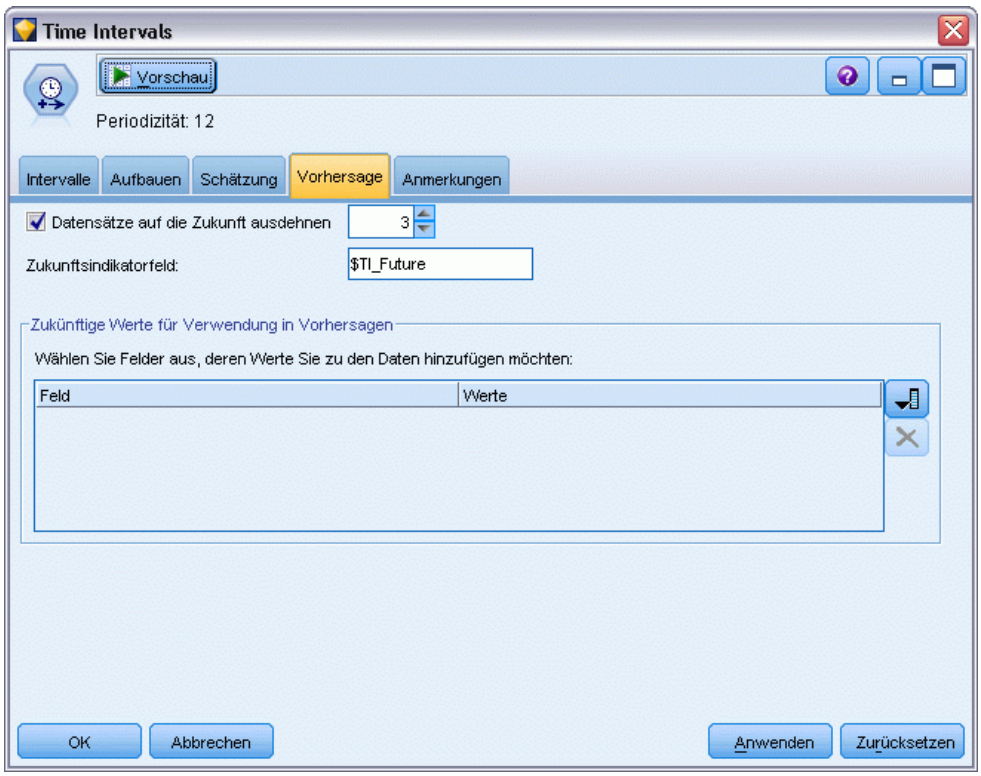
Abbildung 15-11
Festlegen des Zeitintervalls



- Wählen Sie auf der Registerkarte “Vorhersage” die Option Datensätze auf die Zukunft ausdehnen.
- Setzen Sie den Wert auf 3.

▶ Klicken Sie auf OK.

Abbildung 15-12
Festlegen der Vohersageperiode

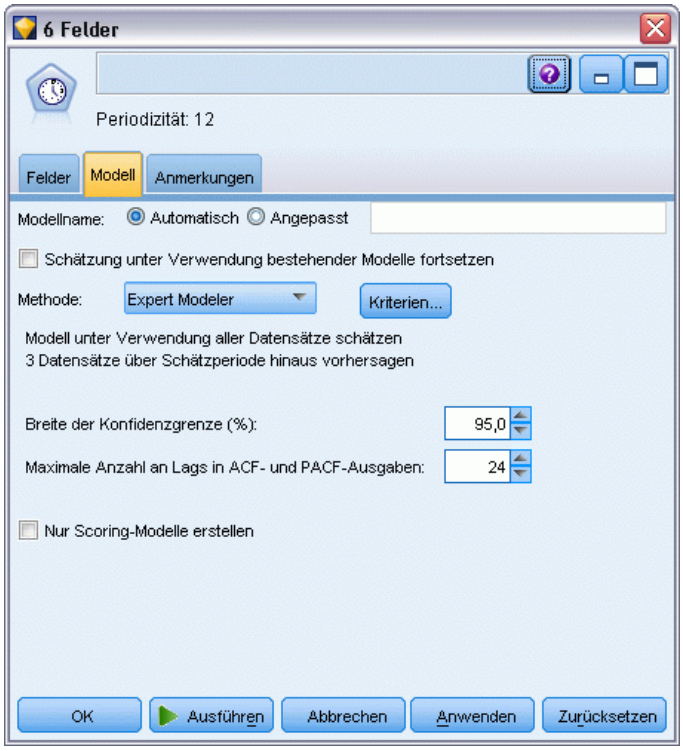


Erstellen des Modells

▶ Fügen Sie aus der Modellierungspalette einen Zeitreihenknoten zum Stream hinzu und verbinden Sie ihn mit dem Zeitintervallknoten.

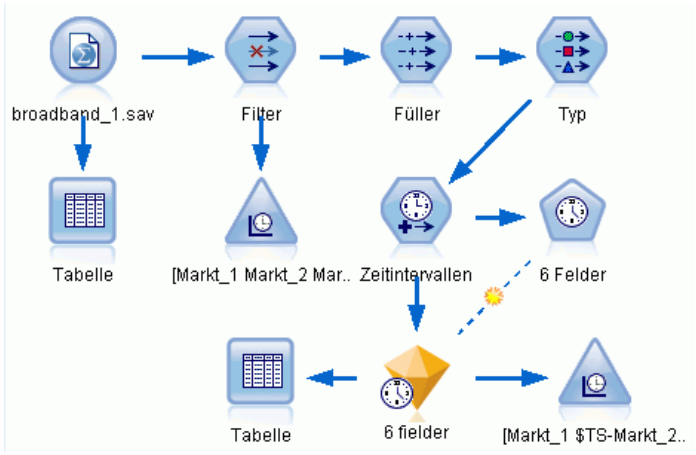
- Klicken Sie auf Ausführen und führen Sie den Zeitreihenknoten mit allen Standardeinstellungen aus. Dadurch kann der Expert Modeler das geeignetste Modell für die einzelnen Zeitreihen ermitteln.

Abbildung 15-13
Auswählen des Expert Modelers für Zeitreihen



- Verbinden Sie das Modell-Nugget vom Typ “Zeitreihe” mit dem Zeitintervallknoten.
- Verbinden Sie einen Tabellenknoten mit dem Zeitreihenmodell und klicken Sie auf Ausführen.

Abbildung 15-14
Beispiel-Stream zur Anzeige der Zeitreihenmodellierung



Es wurden nun drei neue Zeilen (61 bis 63) an die ursprünglichen Daten angehängt. Dabei handelt es sich um die Zeilen für die Vorhersageperiode, in diesem Fall Januar bis März 2004.

Jetzt liegen auch mehrere neue Spalten vor: Eine Reihe von *$TI_*-Spalten, die vom Zeitintervallknoten hinzugefügt wurden, und die vom Zeitreihenknoten hinzugefügten *$TS*--Spalten. Die Spalten enthalten folgende Angaben für die einzelnen Zeilen (d. h. für jedes Intervall in den Zeitreihendaten):

Column	Beschreibung
$TI_TimeIndex	Der Indexwert des Zeitintervalls für die betreffende Zeile.
$TI_TimeLabel	Die Beschriftung des Zeitintervalls für die betreffende Zeile.
$TI_Year	Die Indikatoren für Jahr und Monat für die generierten Daten in der betreffenden Zeile.
$TI_Month	
$TI_Count	Die Anzahl der Datensätze, die an der Bestimmung der neuen Daten für diese Zeile beteiligt sind.
$TI_Future	Gibt an, ob die betreffende Zeile Prognosedaten enthält.
$TS-*Spaltenname*	Die Daten des generierten Modells für die einzelnen Spalten der ursprünglichen Daten.
$TSLCI-*Spaltenname*	Der untere Wert des Konfidenzintervalls für die einzelnen Spalten der Daten des generierten Modells.
$TSUCI-*Spaltenname*	Der obere Wert des Konfidenzintervalls 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.

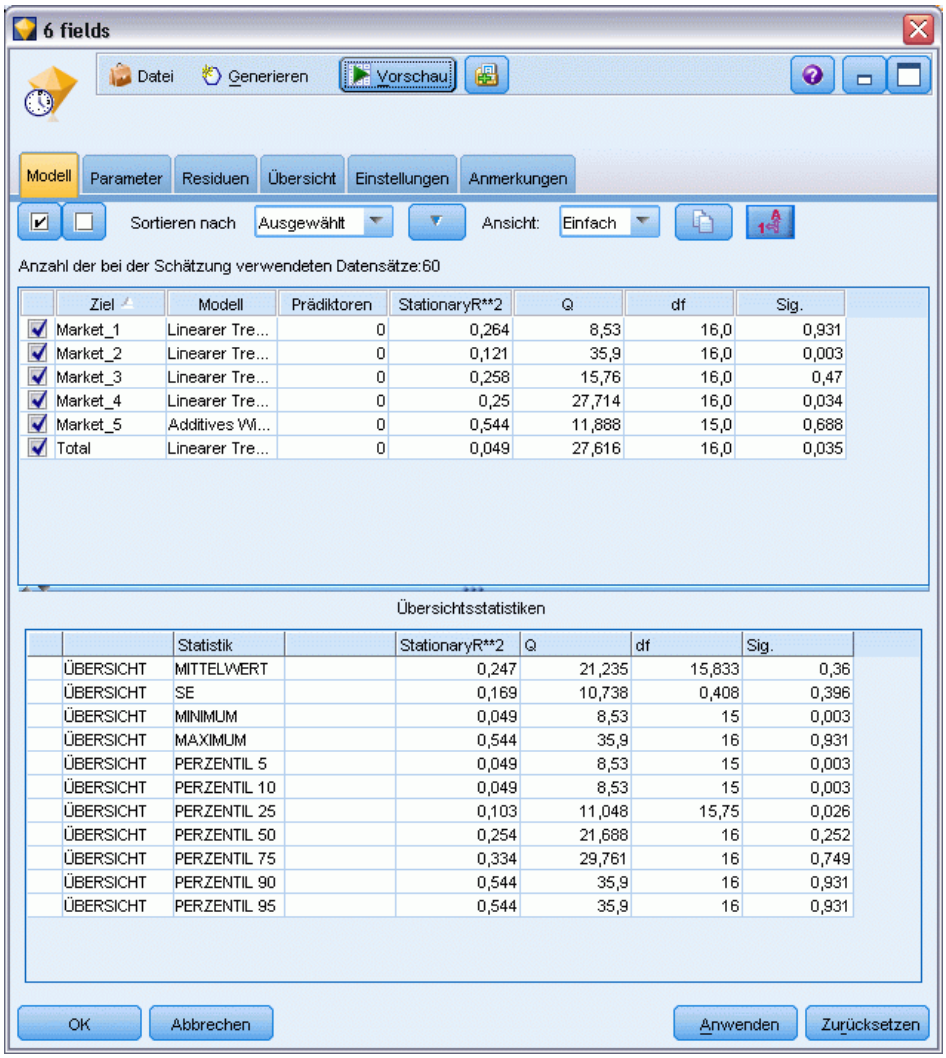
Die wichtigsten Spalten für die Vorhersageoperation sind die Spalten *$TS-Market_n*, *$TSLCI-Market_n* und *$TSUCI-Market_n*. Insbesondere enthalten diese Spalten in den Zeilen 61 bis 63 die Prognosedaten für die Benutzerabonnements und die Konfidenzintervalle für die einzelnen lokalen Märkte.

Untersuchen des Modells

- Doppelklicken Sie auf das Modell-Nugget vom Typ “Zeitreihe”, um Daten zu den für die einzelnen Märkte generierten Modellen anzuzeigen.

Beachten Sie, dass der Expert Modeler für Markt 5 einen anderen Modelltyp generiert hat als für die anderen Märkte.

Abbildung 15-15
Für die Märkte generierte Zeitreihenmodelle



Ziel	Modell	Prädiktoren	StationaryR**2	Q	df	Sig.
Market_1	Linearer Tre...	0	0,264	8,53	16,0	0,931
Market_2	Linearer Tre...	0	0,121	35,9	16,0	0,003
Market_3	Linearer Tre...	0	0,258	15,76	16,0	0,47
Market_4	Linearer Tre...	0	0,25	27,714	16,0	0,034
Market_5	Additives Wi...	0	0,544	11,888	15,0	0,688
Total	Linearer Tre...	0	0,049	27,616	16,0	0,035

	Statistik	StationaryR**2	Q	df	Sig.
ÜBERSICHT	MITTELWERT	0,247	21,235	15,833	0,36
ÜBERSICHT	SE	0,169	10,738	0,408	0,396
ÜBERSICHT	MINIMUM	0,049	8,53	15	0,003
ÜBERSICHT	MAXIMUM	0,544	35,9	16	0,931
ÜBERSICHT	PERZENTIL 5	0,049	8,53	15	0,003
ÜBERSICHT	PERZENTIL 10	0,049	8,53	15	0,003
ÜBERSICHT	PERZENTIL 25	0,103	11,048	15,75	0,026
ÜBERSICHT	PERZENTIL 50	0,254	21,688	16	0,252
ÜBERSICHT	PERZENTIL 75	0,334	29,761	16	0,749
ÜBERSICHT	PERZENTIL 90	0,544	35,9	16	0,931
ÜBERSICHT	PERZENTIL 95	0,544	35,9	16	0,931

In der Spalte “Prädiktoren” wird angezeigt, wie viele Felder als Prädiktoren für die einzelnen Ziele verwendet wurden – in diesem Fall: keine.

Die übrigen Spalten in dieser Ansicht zeigen verschiedene Maße für die Anpassungsgüte für die einzelnen Modelle. In der Spalte StationaryR**2 wird der Wert “*R*-Quadrat für stationären Teil” angezeigt. Diese Statistik bietet eine Schätzung dafür, welcher Anteil an der Gesamtvariation in der Zeitreihe durch das Modell erklärt wird. Je höher der Wert (bis maximal 1,0), desto besser ist die Anpassung des Modells.

Die Spalten Q, df und Sig. beziehen sich auf die Ljung-Box-Statistik, einen Test der Zufälligkeit der Restfehler im Model: je zufälliger die Fehler, desto besser ist das Modell vermutlich. Q ist die Ljung-Box-Statistik selbst, während df (Freiheitsgrade) die Anzahl an Modellparametern angibt, die bei der Schätzung eines bestimmten Ziels frei variieren können.

Die Spalte Sig. enthält den Signifikanzwert der Ljung-Box-Statistik, der ein weiteres Anzeichen dafür darstellt, ob das Modell korrekt angegeben wurde. Ein Signifikanzwert von unter 0,05 bedeutet, dass die Restfehler nicht zufällig sind, was darauf hinweist, dass es in der beobachteten Zeitreihe eine Struktur gibt, die sich nicht durch das Modell erklären lässt.

Unter Berücksichtigung der Werte "*R*-Quadrat für stationären Teil" und "Signifikanz" sind die Modelle, die der Expert Modeller für *Market_1*, *Market_3* und *Market_5* ausgewählt hat, recht brauchbar. Die Sig.-Werte für *Market_2* und *Market_4* liegen jeweils unter 0,05, was darauf hinweist, dass wohl einige Versuche mit besser passenden Modellen für diese Märkte erforderlich sind.

Die Übersichtswerte im unteren Teil der Anzeige bieten Informationen zur Verteilung der Statistiken quer durch alle Modelle. Beispielsweise beträgt der modellübergreifende Mittelwert für "*R*-Quadrat für stationären Teil" 0,247; das Minimum für diesen Wert beträgt 0,049 (im Modell *Total* (Gesamt)) und das Maximum 0,544 (Wert für*Market_5*).

SE gibt den modellübergreifenden Standardfehler für die einzelnen Statistiken an. Beispielsweise weist der modellübergreifende Standardfehler für "*R*-Quadrat für stationären Teil" den Wert 0,169 auf.

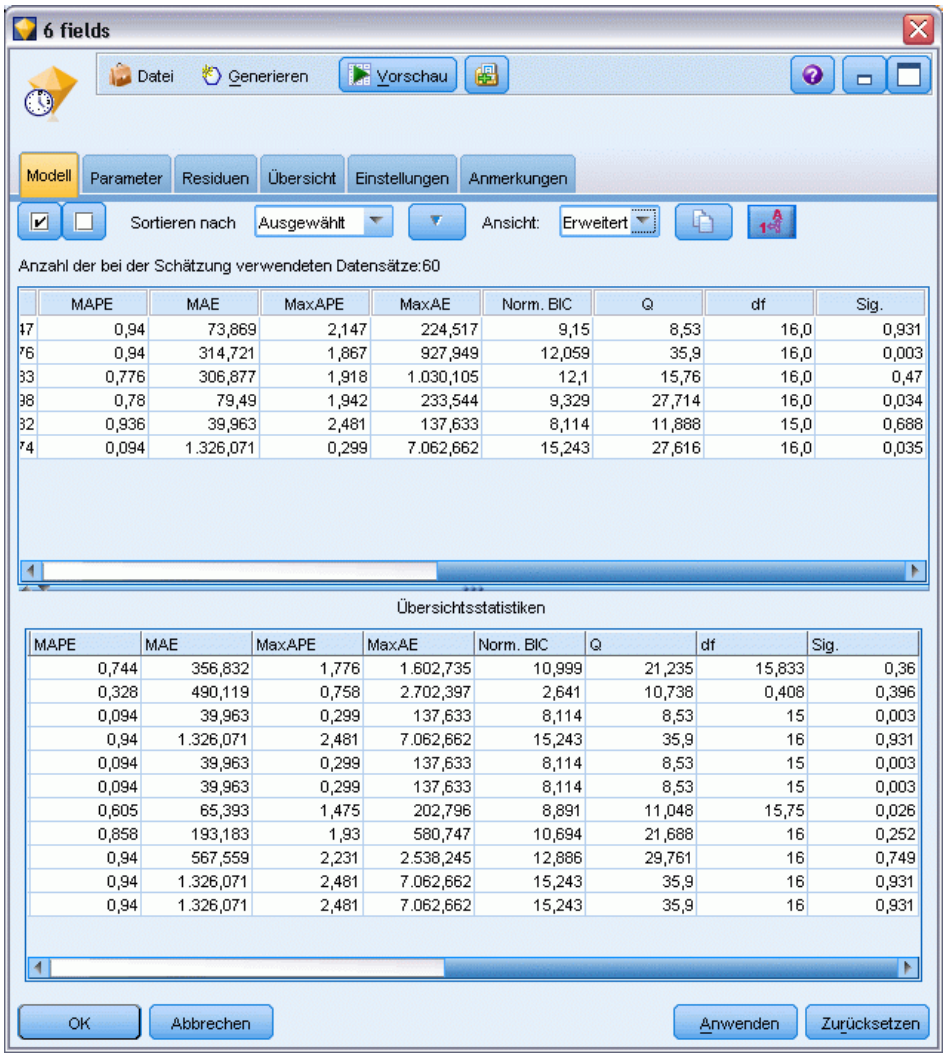
Der Übersichtsabschnitt enthält außerdem Perzentilwerte, die Informationen zu der Verteilung der modellübergreifenden Statistiken bieten. Das jeweilige Perzentil gibt den Prozentsatz der Modelle an, die einen Wert der Anpassungsstatistik aufweisen, der unter dem angegebenen Wert liegt.

So liegt beispielsweise nur bei 25 % der Modelle der Wert für "*R*-Quadrat für stationären Teil" unter 0,121.

▶ Klicken Sie auf die Dropdown-Liste "Ansicht" und wählen Sie die Option Erweitert.

Die Anzeige bietet eine Reihe von weiteren Maßen für die Anpassungsgüte. R**2 ist der *R*-Quadrat-Wert, eine Schätzung der Gesamtvariation in der Zeitreihe, die durch das Modell erklärt werden kann. Da der Maximalwert für diese statistische Funktion 1,0 beträgt, sind die vorliegenden Modelle in dieser Hinsicht sehr gut brauchbar.

Abbildung 15-16
Zeitreihenmodell – erweiterte Anzeige



RMSE steht für Root Mean Square Error (Wurzel der mittleren quadratischen Abweichung), ein Maß das angibt, wie stark die tatsächlichen Werte eine Zeitreihe von den vom Modell vorhergesagten Werten abweichen. Für dieses Maß werden dieselben Einheiten verwendet wie für die Zeitreihe selbst. Da es sich hierbei um ein Fehlermaß handelt, soll dieser Wert so niedrig wie möglich gehalten werden. Auf den ersten Blick hat es den Anschein, dass die Modelle für *Market_2* und *Market_3* zwar gemäß den bisherigen statistischen Funktionen durchaus brauchbar sind, aber doch weniger erfolgreich als die für die anderen drei Märkte.

Zu diesen zusätzlichen Maßen für die Anpassungsgüte gehören der mittlere absolute Fehler in Prozent (MAPE) sowie der zugehörige maximale Wert (MaxAPE). Der absolute Fehler in Prozent ist ein Maß dafür, wie stark eine Ziel-Zeitreihe von dem vom Modell vorhergesagten Niveau abweicht. Dieses Maß wird als Prozentwert angegeben. Wenn Sie den mittleren und maximalen Prozentsatz modellübergreifend untersuchen, erhalten Sie einen Hinweis auf die Unsicherheit in Ihren Vorhersagen.

Der MAPE-Wert zeigt, dass alle Modelle eine mittlere Unsicherheit von ungefähr 1 % aufweisen, was sehr niedrig ist. Der MaxAPE-Wert gibt den maximalen absoluten Fehler in Prozent an und kann zur Erstellung eines Worst-Case-Szenarios für Ihre Vorhersagen herangezogen werden. Er zeigt, dass der größte Fehler in Prozent für jedes Modell, grob gesagt, in den Bereich von 1,8 bis 2,5 % fällt, was ebenfalls sehr niedrig ist.

Der MAE-Wert (Mean Absolute Error, mittlerer absoluter Fehler) gibt den Mittelwert der absoluten Werte der Vorhersagefehler an. Wie beim RMSW-Wert wird dieser Wert in denselben Einheiten ausgedrückt wie die Zeitreihe selbst. MaxAE zeigt den größten Vorhersagefehler in denselben Einheiten und bietet ein Worst-Case-Szenario für die Vorhersagen.

So interessant diese absoluten Werte sein mögen, sind doch die Fehlerwerte in Prozent (MAPE und MaxAPE) in diesem Fall nützlicher, da die Ziel-Zeitreihen auf Abonnentenzahlen für unterschiedlich große Märkte beruhen.

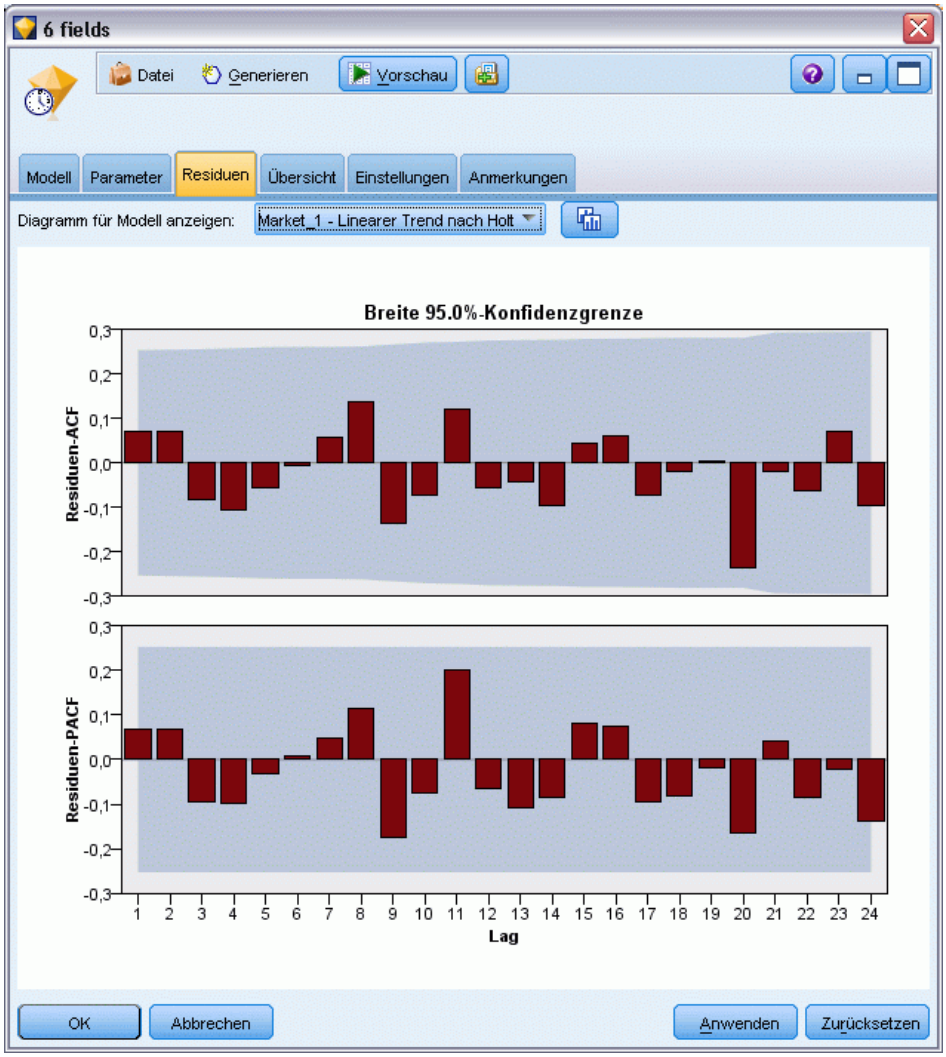
Stellen die Werte MAPE und MaxAPE einen Grad an Unsicherheit dar, der bei den Modellen akzeptabel ist? Sie sind sicherlich sehr niedrig. In einer solchen Situation kommt der Geschäftssinn ins Spiel, da das akzeptable Risiko von Problem zu Problem unterschiedlich ist. Wir nehmen an, dass die Statistiken für die Anpassungsgüte innerhalb akzeptabler Grenzen liegen und fahren mit einer Untersuchung der Restfehler fort.

Eine Untersuchung der Autokorrelationsfunktion (ACF) und der partiellen Autokorrelationsfunktion (PACF) für die Modellresiduen bietet quantitativere Einblicke in die Modelle als die bloße Betrachtung von Statistiken für die Anpassungsgüte.

Ein gut angegebenes Zeitreihenmodell erfasst die gesamte nichtzufällige Variation, einschließlich Saisonalität, Trend sowie zyklischen und sonstigen Faktoren, die von Bedeutung sind. Wenn dies der Fall ist, sollten etwaige Fehler nicht im Laufe der Zeit mit sich selbst korreliert sein (Autokorrelation). Eine signifikante Struktur in einer dieser Autokorrelationsfunktionen würde bedeuten, dass das zugrunde liegende Modell unvollständig ist.

- Klicken Sie auf die Registerkarte “Residuen”, um die Werte für die Autokorrelationsfunktion (ACF) und die partielle Autokorrelationsfunktion (PACF) für die Restfehler im Modell für den ersten der lokalen Märkte anzuzeigen.

Abbildung 15-17
ACF- und PACF-Werte für die Märkte



In diesen Plots wurden die ursprünglichen Werte der Fehlervariablen um bis zu 24 Zeitperioden verschoben und mit dem ursprünglichen Wert verglichen, um festzustellen, ob eine Korrelation im Laufe der Zeit vorliegt. Damit das Modell akzeptabel ist, sollte keiner der Balken im oberen Plot (ACF) über den schattierten Bereich hinausgehen, weder in positiver Richtung (nach oben) noch in negativer (nach unten).

Sollte dieser Fall eintreten, müssen Sie den unteren Plot (PACF) überprüfen, um zu sehen, ob die Struktur dort bestätigt wird. der PACF-Plot untersucht Korrelationen unter Kontrolle der Zeitreihenwerte an den dazwischenkommenden Zeitpunkten.

Die Werte für *Market_1* liegen alle im schattierten Bereich, sodass wir mit der Überprüfung der Werte für die anderen Märkte fortfahren können.

- Klicken Sie auf Diagramm für Modell anzeigen, um diese Werte für die anderen Märkte und die Gesamtwerte anzuzeigen.

  Die Werte für *Market_2* und *Market_4* bieten Anlass zu einer gewissen Besorgnis, denn sie bestätigen, was wir zuvor aus den zugehörigen Sig.-Werten vermutet haben. Irgendwann kommen wir nicht umhin, mit einigen verschiedenen Modellen für diese Märkte experimentieren, um zu sehen, ob sich eine bessere Anpassung erreichen lässt, für den Rest dieses Beispiels jedoch konzentrieren wir uns darauf, was wir noch aus dem Modell für *Market_1* entnehmen können.

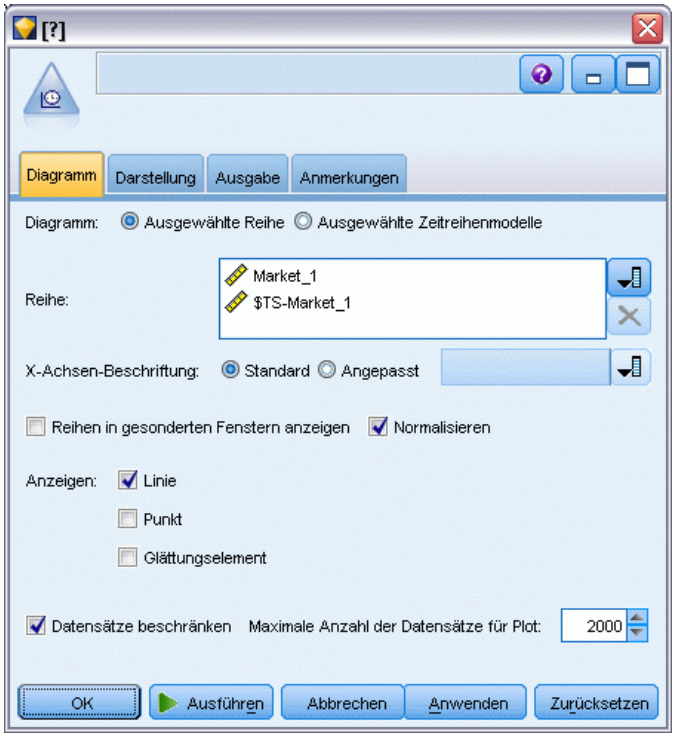
- Gliedern Sie auf der Diagrammpalette einen Zeitdiagrammknoten an das Modell-Nugget vom Typ "Zeitreihe" an.

- Deaktivieren Sie auf der Registerkarte "Plot" das Kontrollkästchen Reihen in gesonderten Fenstern anzeigen.

- Klicken Sie in der Liste Series (Reihe) auf die Feldauswahlschaltfläche, wählen Sie die Felder *Market_1* und *$TS-Market_1* aus und klicken Sie auf OK, um sie zur Liste hinzuzufügen.

- Klicken Sie auf Ausführen, um ein Liniendiagramm der Ist-Daten und der vorhergesagten Daten für die ersten der lokalen Märkte anzuzeigen.

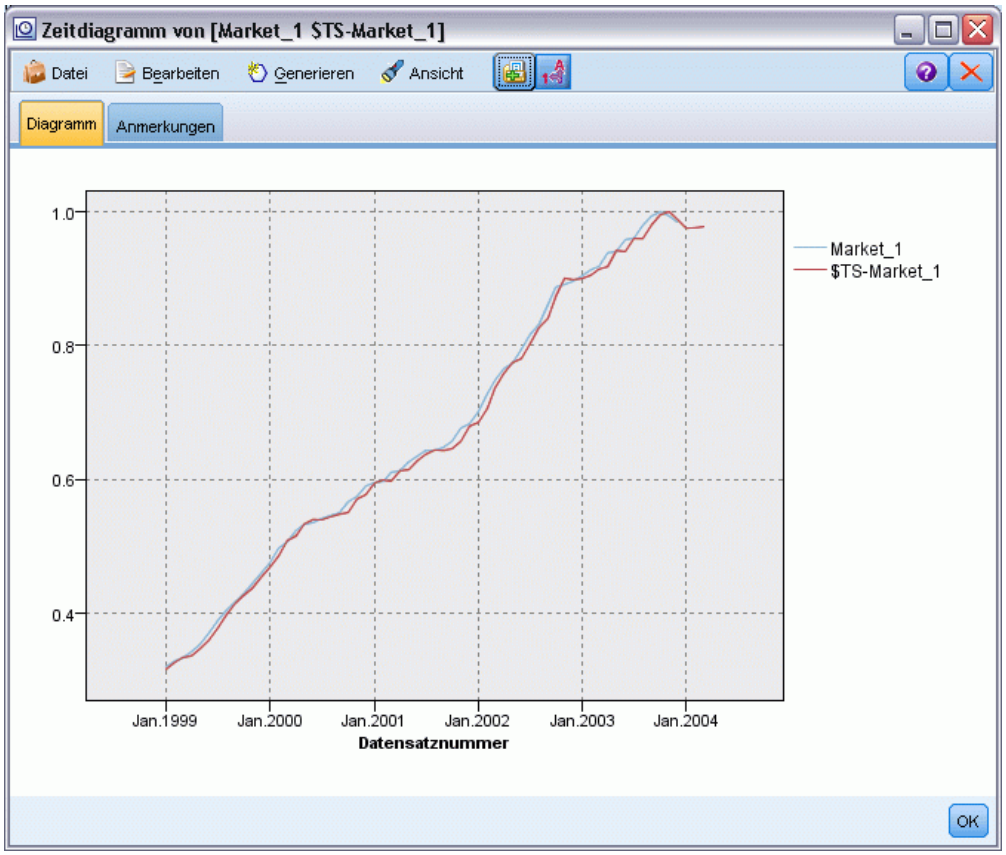
Abbildung 15-18
Auswählen der zu plottenden Felder



Achten Sie darauf, wie die Prognoselinie (*$TS-Market_1*) über das Ende der Ist-Daten hinausgeht. Es liegt nun eine Vorhersage der erwarteten Nachfrage für die nächsten drei Monate in diesem Markt vor.

Die Linien für die Ist-Daten und die vorhergesagten Daten über die gesamte Zeitreihe liegen im Diagramm sehr eng beieinander. Dies weist darauf hin, dass es sich um ein zuverlässiges Modell für die konkrete Zeitreihe handelt.

Abbildung 15-19
Zeitdiagramm der Ist-Daten und der vorhergesagten Daten für Market_1



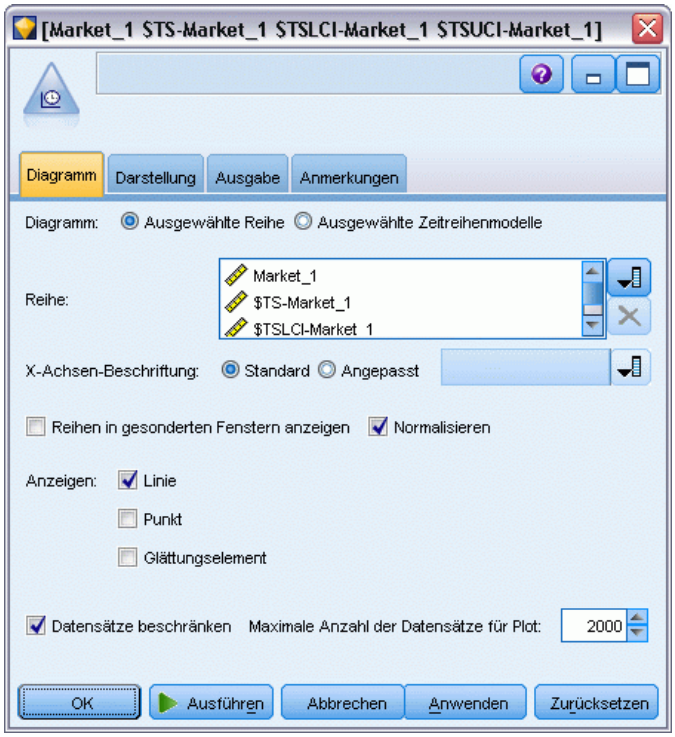
Speichern Sie das Modell in einer Datei, um es in einem zukünftigen Beispiel verwenden zu können:

- Klicken Sie auf OK, um das aktuelle Diagramm zu schließen.
- Öffnen Sie das Modell-Nugget vom Typ “Zeitreihe”.
- Wählen Sie die Optionsfolge Datei > Knoten speichern und geben Sie den Speicherort für die Datei an.
- Klicken Sie auf Speichern.

Sie verfügen über ein zuverlässiges Modell für den betreffenden Markt, aber welche Fehlermarge weist die Vorhersage auf? Einen Hinweis darauf erhalten Sie durch Untersuchung des Konfidenzintervalls.

- Doppelklicken Sie auf den letzten Zeitdiagrammknoten im Stream (den Knoten mit der Beschriftung Market_1 $TS-Market_1), um das zugehörige Dialogfeld erneut zu öffnen.
- Klicken Sie auf die Feldauswahlschaltfläche und fügen Sie die Felder *$TSLCI-Market_1* und *$TSUCI-Market_1* zur Liste Series (Reihe) hinzu.
- Klicken Sie auf Ausführen.

Abbildung 15-20
Hinzufügen weiterer zu plottender Felder



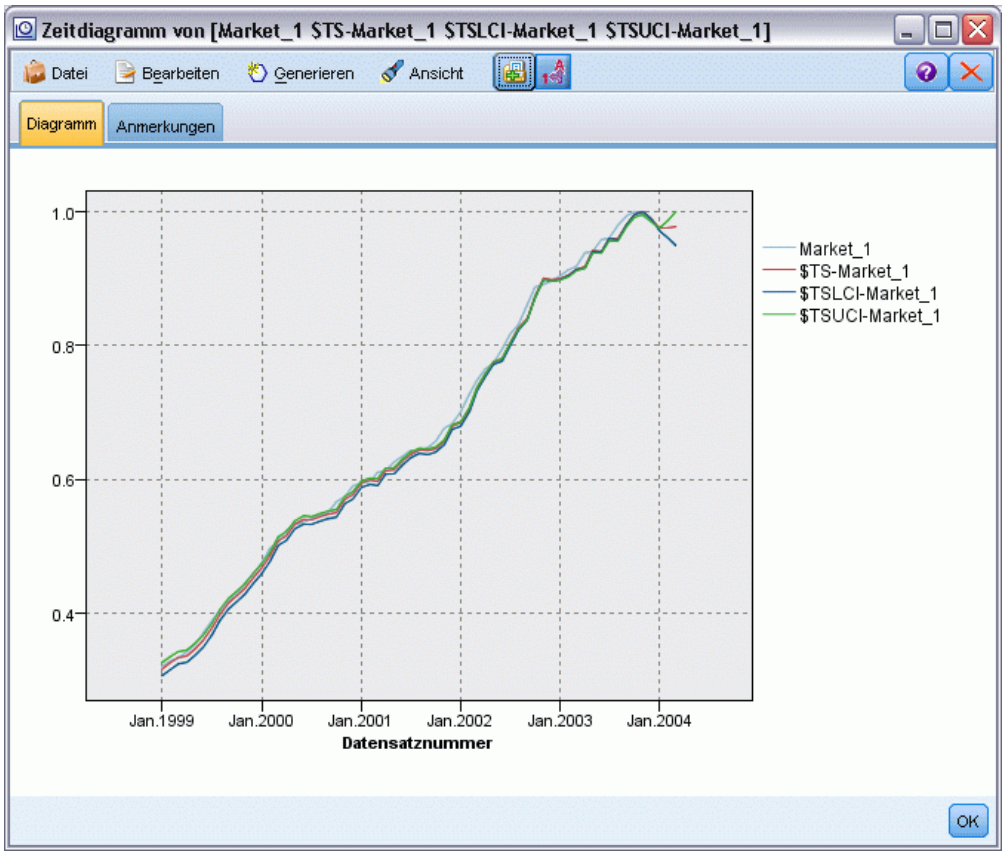
Sie erhalten dasselbe Diagramm wie zuvor, allerdings diesmal um die obere (*$TSUCI*) und untere lower (*$TSLCI*) Grenze des Konfidenzintervalls ergänzt.

Beachten Sie, wie die Grenzen des Konfidenzintervalls über die Vorhersageperiode divergieren, was auf zunehmende Unsicherheit hindeutet, je weiter sich die Vorhersage in die Zukunft erstreckt.

Allerdings haben sie nach Ablauf der einzelnen Zeitperioden jeweils einen weiteren (in diesem Fall) Monat mit tatsächlichen Nutzungsdaten, die Sie der Vorhersage zugrunde legen können. Sie können die neuen Daten in den Stream einlesen und Ihr Modell erneut anwenden, nun da Sie

wissen, dass es zuverlässig ist. Für weitere Informationen siehe Thema Erneutes Anwenden eines Zeitreihenmodells auf S. 202.

Abbildung 15-21
Zeitdiagramm, um Konfidenzintervall ergänzt



Zusammenfassung

Sie haben nun erfahren, wie Sie mit dem Expert Modeler Vorhersagen für mehrere Zeitreihen erstellen können, und Sie haben die so entstandenen Modelle in einer externen Datei gespeichert.

Im nächsten Beispiel sehen Sie, wie Sie nicht dem Standard entsprechende Zeitreihendaten in ein Format transformieren können, das sich für die Eingabe in einen Zeitreihenknoten eignet.

Erneutes Anwenden eines Zeitreihenmodells

In diesem Beispiel werden die Zeitreihenmodelle aus dem ersten Zeitreihenbeispiel angewendet. Die Modelle können jedoch auch unabhängig davon verwendet werden. Für weitere Informationen siehe Thema Prognoseerstellung mit dem Zeitreihenknoten auf S. 180.

Wie im ursprünglichen Szenario muss ein Analyst eines nationalen Breitbandanbieters monatliche Prognosen der Benutzerabonnements für eine Reihe von lokalen Märkten erstellen, um en Bandbreitenbedarf zu prognostizieren. Sie haben bereits Modelle mit dem Expert Modeler erstellt und eine Vorhersage über drei Monate angefertigt.

Ihr Data Warehouse wurde nun mit den Ist-Daten für die ursprüngliche Vorhersageperiode aktualisiert und Sie möchten diese Daten verwenden, um den Prognosehorizont um weitere drei Monate auszudehnen.

In diesem Beispiel wird ein Stream namens *broadband_apply_models.str* verwendet, der Bezug nimmt auf die Datendatei *broadband_2.sav*. Die Dateien stehen im Ordner *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *broadband_apply_models.str* befindet sich im Ordner *streams*.

Abrufen des Streams

In diesem Fall erstellen Sie den Zeitreihenknoten aus dem im ersten Beispiel gespeicherten Zeitreihenmodell neu. Sie müssen kein Modell gespeichert haben. Wir haben eines für Sie im Ordner *Demos* zur Verfügung gestellt.

- Öffnen Sie den Stream *broadband_apply_models.str* im Ordner *streams* unter *Demos*.

Abbildung 15-22
Öffnen des Streams

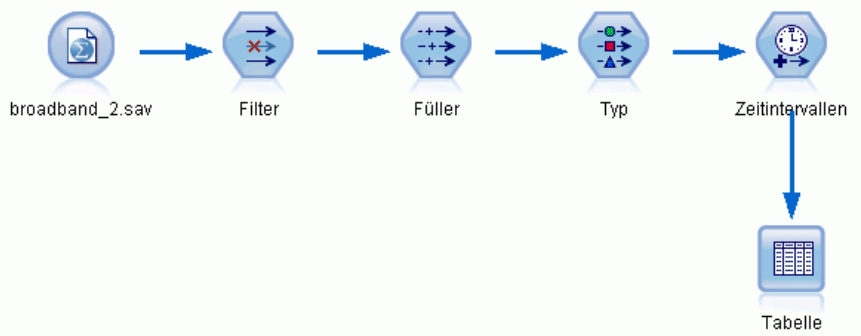


Abbildung 15-23
Aktualisierte Umsatzdaten

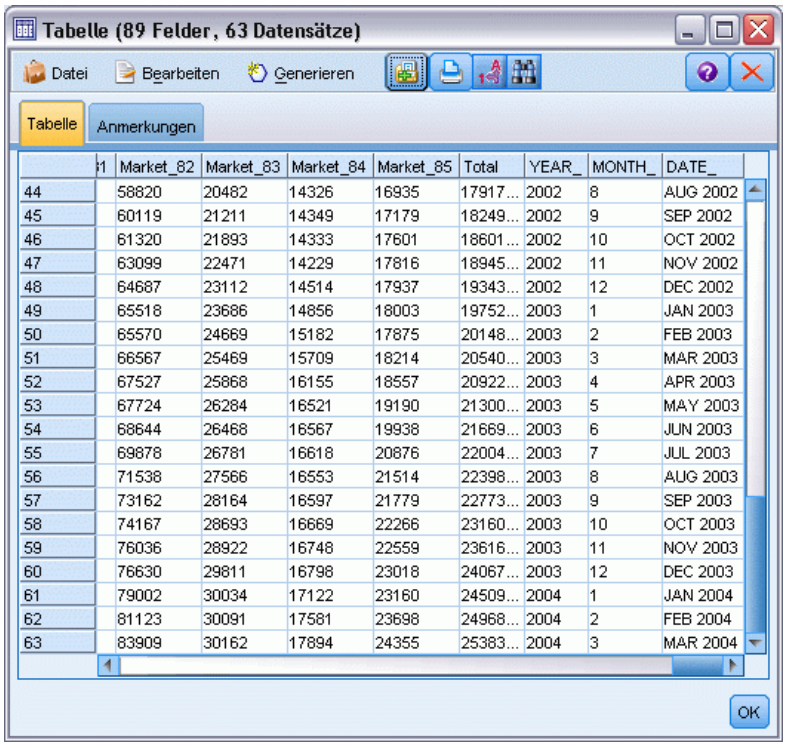


Tabelle (89 Felder, 63 Datensätze)

Datei Bearbeiten Generieren

Tabelle Anmerkungen

	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

OK

Die aktualisierten Monatsdaten finden Sie in der Datei *broadband_2.sav*.

- Verbinden Sie einen Tabellenknoten mit dem IBM® SPSS® Statistics-Dateiquellenknoten, öffnen Sie den Tabellenknoten und klicken Sie auf Ausführen.

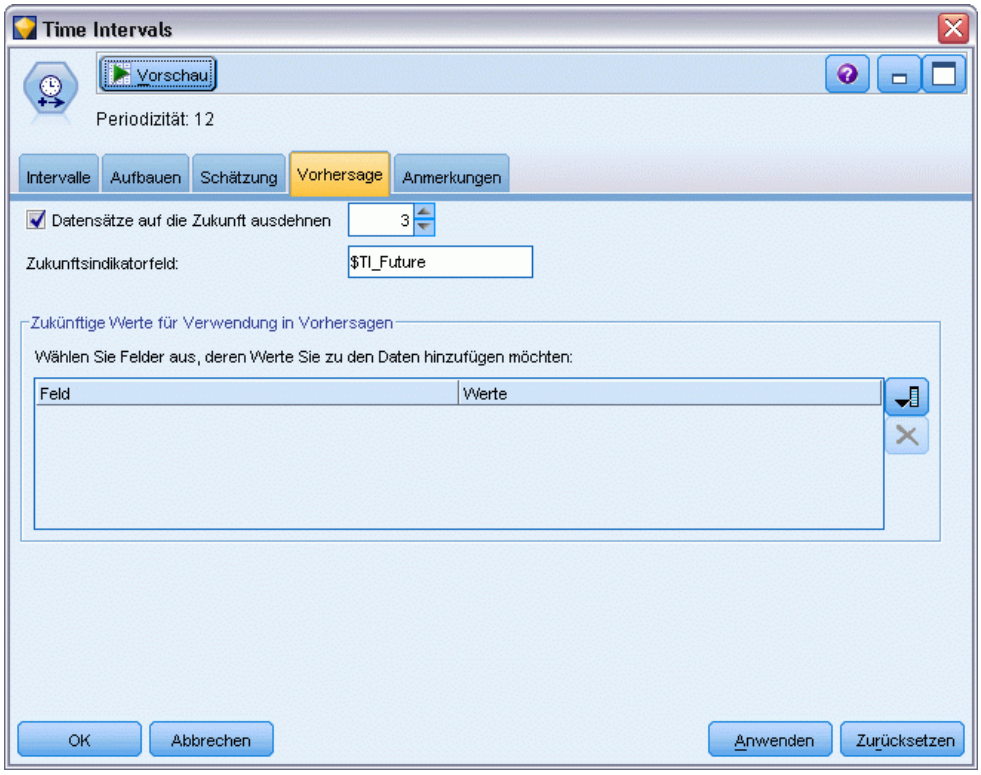
  Hinweis: Die Datendatei wurde mit den tatsächlichen Umsatzdaten für Januar bis März 2004 aktualisiert (Zeilen 61 bis 63).

- Öffnen Sie den Zeitintervallknoten im Stream.

- Klicken Sie auf die Registerkarte Vorhersage.

- Vergewissern Sie sich, dass Datensätze auf die Zukunft ausdehnen auf 3 gesetzt ist.

Abbildung 15-24
Überprüfen der Einstellung für die Vorhersageperiode

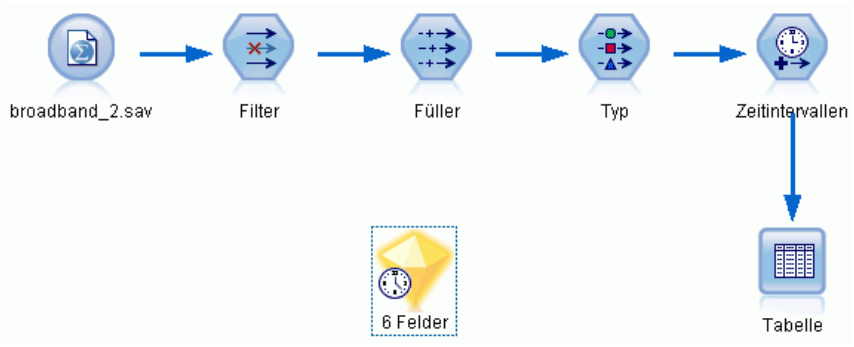


Abrufen des gespeicherten Modells

- Wählen Sie im IBM® SPSS® Modeler-Menü die Optionsfolge Einfügen > Knoten aus Datei und wählen Sie die Datei *TSmodel.nod* im Ordner *Demos* (oder verwenden Sie das Zeitreihenmodell, das Sie im ersten Zeitreihenbeispiel gespeichert haben).

Diese Datei enthält die Zeitreihenmodelle aus dem vorherigen Beispiel. Der Einfügevorgang platziert das entsprechende Zeitreihenmodell-Nugget im Zeichenbereich.

Abbildung 15-25
Hinzufügen des Modell-Nuggets

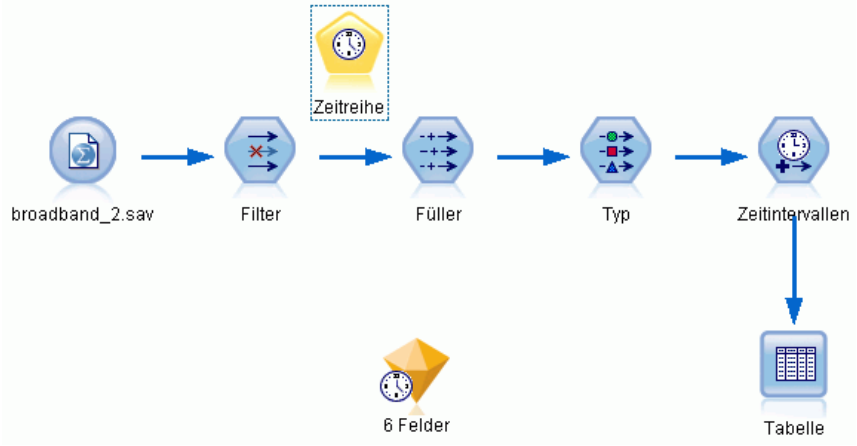


Generieren eines Modellknotens

- Öffnen Sie das Modell-Nugget vom Typ “Zeitreihe” und wählen Sie die Optionsfolge Generieren > Modellierungsknoten erzeugen.

Dadurch wird ein Zeitreihen-Modellierungsknoten im Zeichenbereich platziert.

Abbildung 15-26
Generieren eines Modellierungsknotens aus dem Modell-Nugget



Generieren eines neuen Modells

- Schließen Sie das Zeitreihenmodell-Nugget und löschen Sie es aus dem Zeichenbereich.

  Das alte Modell wurde anhand von 60 Datenzeilen erstellt. Sie müssen ein neues Modell auf der Grundlage der aktualisierten Umsatzdaten erstellen (63 Zeilen).

- Fügen Sie den neu erzeugten Zeitreihen-Erstellungsknoten zum Stream hinzu.

Abbildung 15-27
Hinzufügen des Modellierungsknotens zum Stream

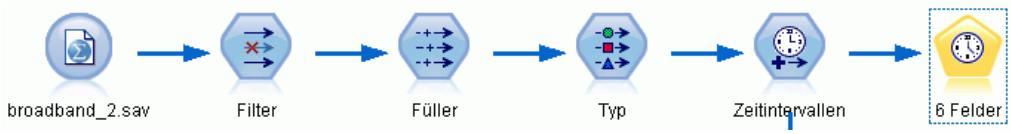
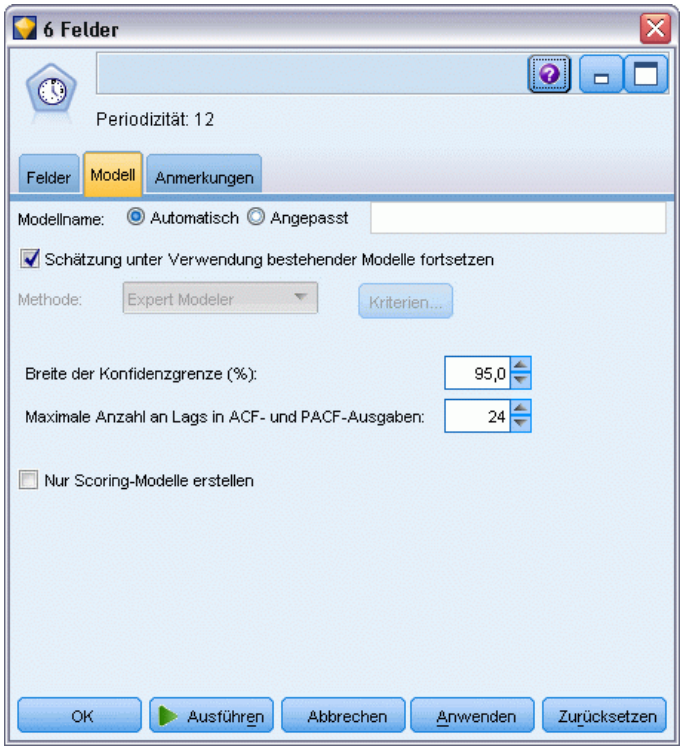


Abbildung 15-28
Wiederverwenden gespeicherter Einstellungen für das Zeitreihenmodell



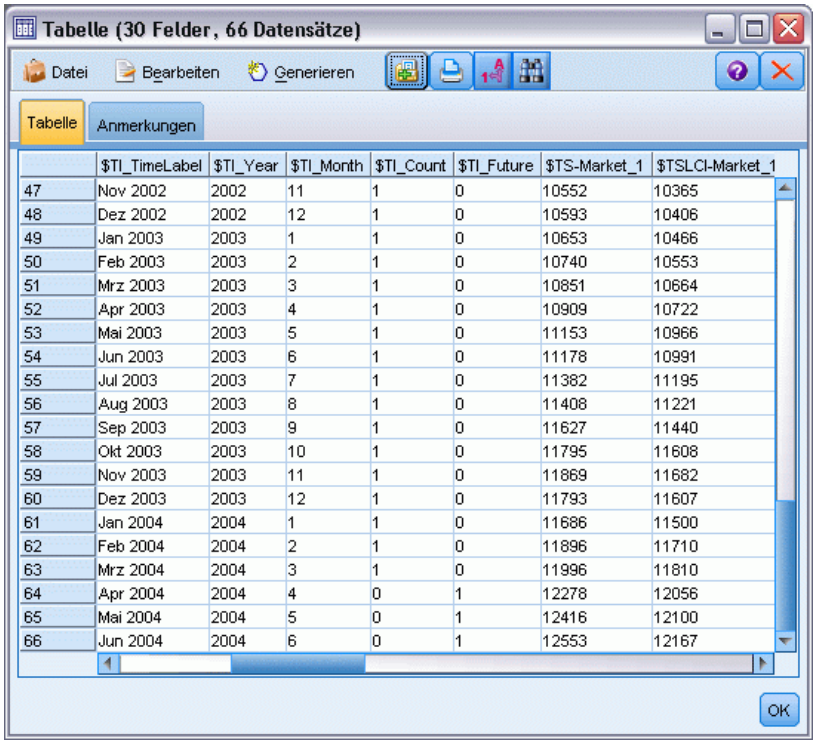
- Öffnen Sie den Zeitreihenknoten.

- Vergewissern Sie sich auf der Registerkarte Modell, dass die Option Schätzung unter Verwendung bestehender Modelle fortsetzen aktiviert ist.

- Klicken Sie auf Ausführen, um ein neues Modell-Nugget auf der Zeichenfläche und in der Modellpalette zu platzieren.

Untersuchen des neuen Modells

Abbildung 15-29
Tabelle mit neuer Vorhersage

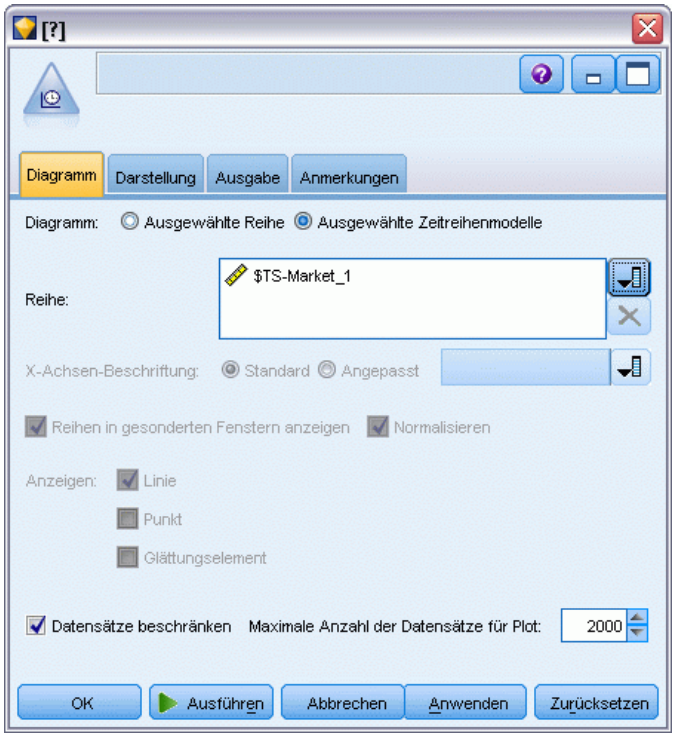


	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-Market_1	$TSLCI-Market_1
47	Nov 2002	2002	11	1	0	10552	10365
48	Dez 2002	2002	12	1	0	10593	10406
49	Jan 2003	2003	1	1	0	10653	10466
50	Feb 2003	2003	2	1	0	10740	10553
51	Mrz 2003	2003	3	1	0	10851	10664
52	Apr 2003	2003	4	1	0	10909	10722
53	Mai 2003	2003	5	1	0	11153	10966
54	Jun 2003	2003	6	1	0	11178	10991
55	Jul 2003	2003	7	1	0	11382	11195
56	Aug 2003	2003	8	1	0	11408	11221
57	Sep 2003	2003	9	1	0	11627	11440
58	Okt 2003	2003	10	1	0	11795	11608
59	Nov 2003	2003	11	1	0	11869	11682
60	Dez 2003	2003	12	1	0	11793	11607
61	Jan 2004	2004	1	1	0	11686	11500
62	Feb 2004	2004	2	1	0	11896	11710
63	Mrz 2004	2004	3	1	0	11996	11810
64	Apr 2004	2004	4	0	1	12278	12056
65	Mai 2004	2004	5	0	1	12416	12100
66	Jun 2004	2004	6	0	1	12553	12167

- Verbinden Sie einen Tabellenknoten mit dem neuen Zeitreihenmodell-Nugget auf der Zeichenfläche.
- Öffnen Sie den Tabellenknoten und klicken Sie auf Ausführen.

  Auch die Prognose des neuen Modells erstreckt sich drei Monate in die Zukunft, da Sie die gespeicherten Einstellungen wiederverwenden. Dieses Mal wird jedoch der Zeitraum von April bis Juni vorhergesagt, da der Schätzzeitraum (im Zeitintervallknoten angegeben) nun anstatt im Januar im März endet.

Abbildung 15-30
Angabe der zu plottenden Felder



- Fügen Sie einen Zeitdiagrammknoten zum Zeitreihenmodell-Nugget hinzu.

  Diesmal verwenden wir die speziell für Zeitreihenmodelle entworfene Zeitdiagrammanzeige.

- Wählen Sie auf der Registerkarte “Plot” die Option Ausgewählte Zeitreihenmodelle.

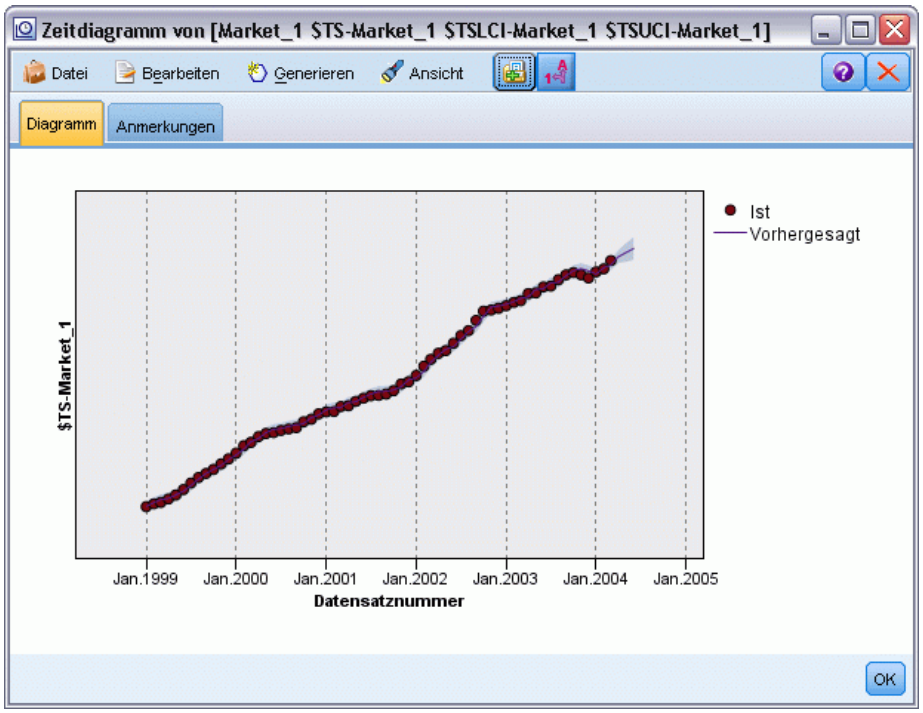
- Klicken Sie in der Liste Series (Reihe) auf die Feldauswahlschaltfläche, wählen Sie das Feld *$TS-Market_1* aus und klicken Sie auf OK, um es zur Liste hinzuzufügen.

- Klicken Sie auf Ausführen.

  Nun haben Sie ein Diagramm, das die aktuellen Umsatzdaten für *Market_1* bis März 2004 sowie die vorhergesagten Umsatzdaten und das Konfidenzintervall (durch den blau schattierten Bereich angezeigt) bis Juni 2004 angibt.

Wie im ersten Beispiel folgen die vorhergesagten Werte während der gesamten Zeitperiode eng den Ist-Daten, was nochmals anzeigt, dass Sie ein gutes Modell verwenden.

Abbildung 15-31
Bis Juni erweiterte Vorhersage



Zusammenfassung

Sie haben erfahren, wie Sie gespeicherte Modelle anwenden können, um Ihre vorherigen Prognosen zu erweitern, wenn aktuellere Daten verfügbar werden, und Sie haben dies getan, ohne die Modelle neu zu erstellen. Wenn es Grund zu der Annahme gibt, dass sich ein Modell geändert hat, sollten Sie es natürlich neu erstellen.

Kapitel
16

Vorhersage von Katalogverkäufen (Zeitreihen)

Ein Versandhaus möchte anhand der Umsatzdaten der letzten 10 Jahre die monatlichen Umsatzzahlen in seinem Herrenbekleidungssortiment vorhersagen.

In diesem Beispiel wird ein Stream namens *catalog_forecast.str* verwendet, der Bezug nimmt auf die Datendatei *catalog_forecast.str*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *catalog_forecast.str* befindet sich im Verzeichnis *streams*.

In einem früheren Beispiel haben Sie gesehen, wie Sie das am besten geeignete Modell für Ihre Zeitreihe vom Expert Modeler ermitteln lassen können. Nun wenden wir uns den beiden Methoden zu, die Ihnen zur Verfügung stehen, wenn Sie selbst ein Modell auswählen möchten: Exponentielles Glätten und ARIMA.

Bei der Suche nach einem geeigneten Modell sollte zunächst die Zeitreihe geplottet werden. Die optische Untersuchung einer Zeitreihe kann oft entscheidende Hinweise für die Auswahl geben. Insbesondere sollten Sie sich folgende Fragen stellen:

- Weist die Zeitreihe einen allgemeinen Trend auf? Wenn ja, scheint der Trend konstant zu sein oder scheint er mit der Zeit auszulaufen?
- Weist die Zeitreihe Saisonalität auf? Wenn ja, scheinen die saisonalen Schwankungen im Laufe der Zeit zuzunehmen oder scheinen sie über mehrere aufeinander folgende Perioden konstant zu sein?

Erstellen des Streams

▶ Erstellen Sie einen neuen Stream und fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *catalog_seasfac.sav* verweist.

Abbildung 16-1
Vorhersage von Katalogverkäufen

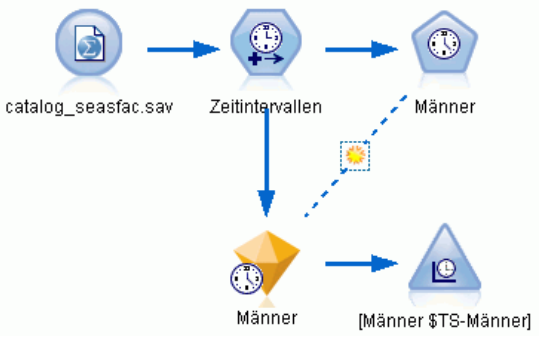
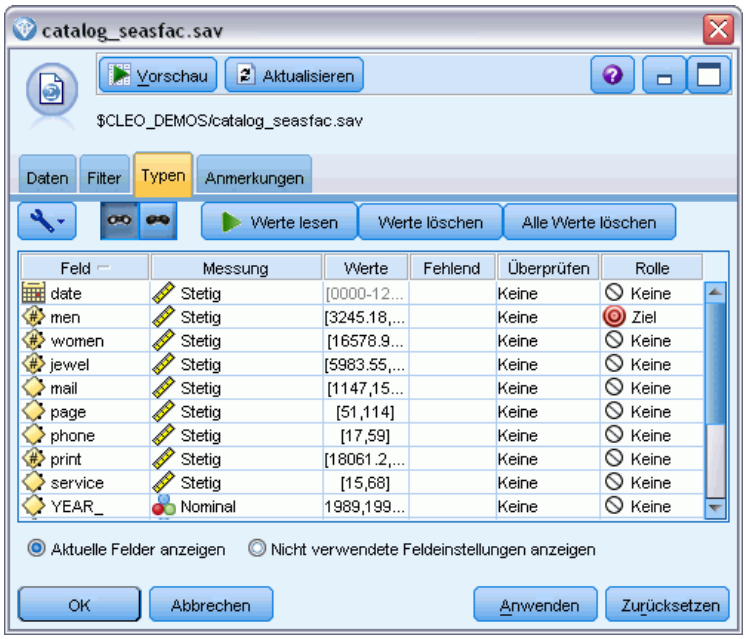
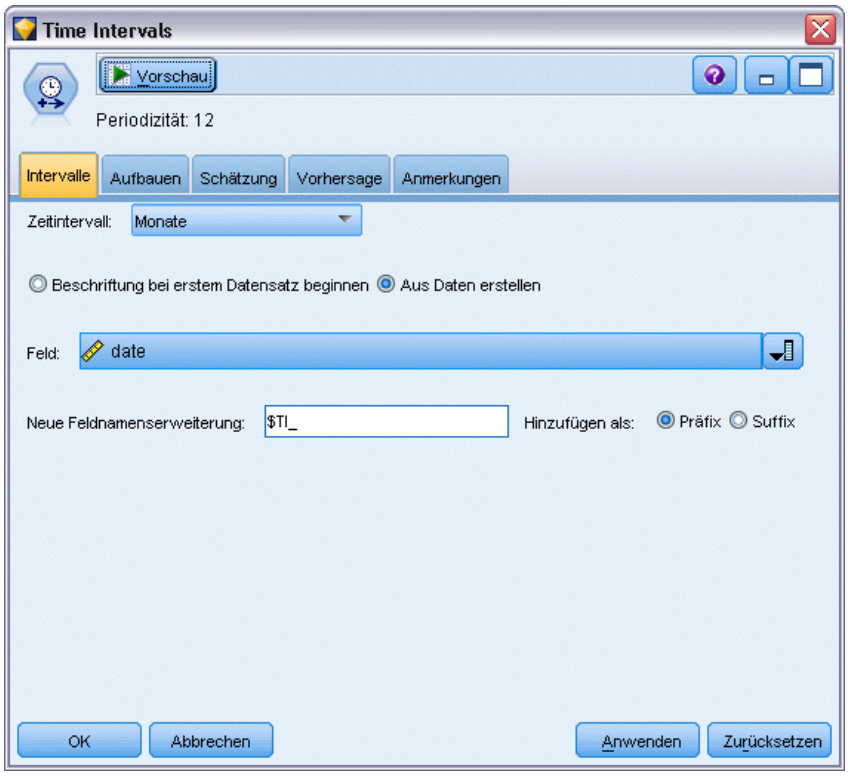


Abbildung 16-2
Angabe des Zielfelds



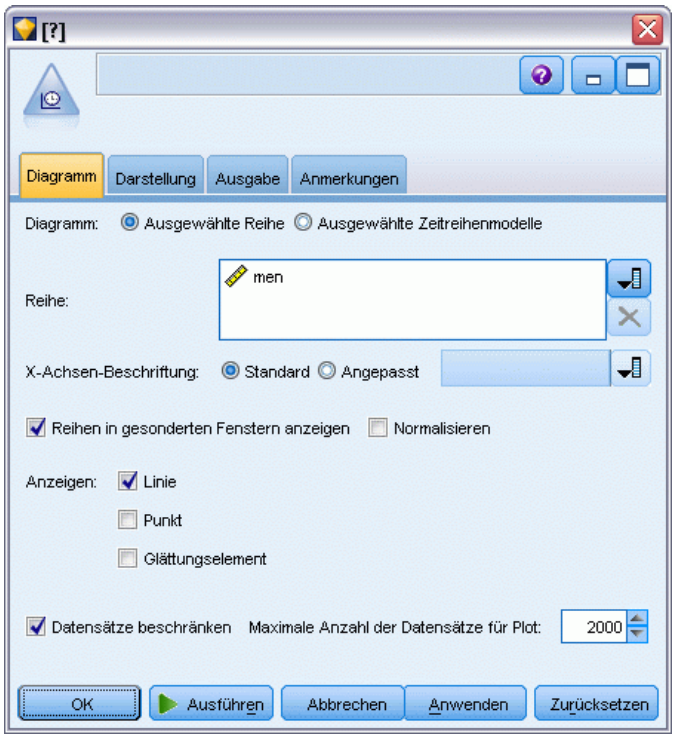
- Öffnen Sie den IBM® SPSS® Statistics-Dateiquellenknoten und wählen Sie die Registerkarte “Typen” aus.
- Klicken Sie auf Werte lesen und dann auf OK.
- Klicken Sie in die Spalte *Rolle* für das Feld *men* (Herren) und setzen Sie die Rolle auf Ziel.
- Setzen Sie die Rolle für alle anderen Felder auf Keine und klicken Sie auf OK.

Abbildung 16-3
Festlegen des Zeitintervalls



- Verbinden Sie einen Zeitintervallknoten mit dem SPSS Statistics-Dateiquellenknoten.
- Öffnen Sie den Zeitintervallknoten und setzen Sie Zeitintervall auf Monate.
- Wählen Sie die Option Aus Daten erstellen.
- Setzen Sie Feld auf date (Datum) und klicken Sie auf OK.

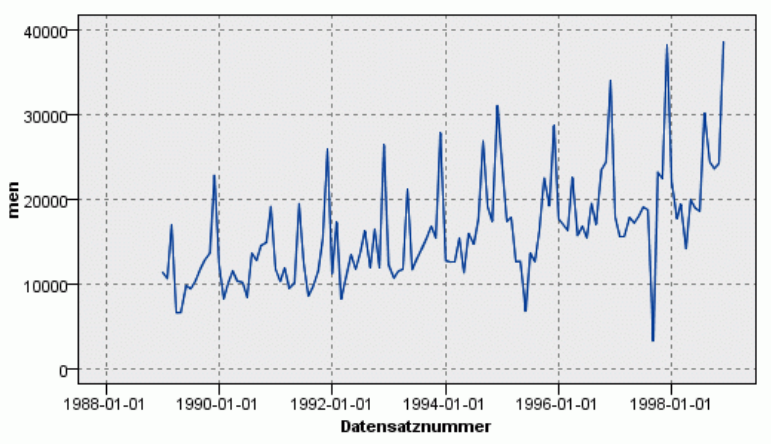
Abbildung 16-4
Plotten der Zeitreihe



- Verbinden Sie einen Zeitdiagrammknoten mit dem Zeitintervallknoten.
- Fügen Sie auf der Registerkarte “Plot” men (Herren) zur Liste “Reihe” hinzu.
- Deaktivieren Sie das Kontrollkästchen Normalisieren.
- Klicken Sie auf Ausführen.

Untersuchen der Daten

Abbildung 16-5
Tatsächlicher Umsatz bei der Herrenbekleidung



Die Zeitreihe weist einen allgemeinen Aufwärtstrend auf; d. h. die Werte der Zeitreihe nehmen tendenziell im Laufe der Zeit zu. Der Aufwärtstrend scheint konstant zu sein, was auf einen linearen Trend hindeutet.

Außerdem weist die Zeitreihe ein deutliches saisonales Muster mit jährlichen Spitzen im Dezember auf, wie durch die vertikalen Linien im Diagramm angedeutet. Die saisonalen Schwankungen scheinen mit dem Aufwärtstrend der Zeitreihe zu wachsen, was darauf hindeutet, dass vermutlich eher eine multiplikative und keine additive Saisonalität vorliegt.

▶ Klicken Sie auf OK, um den Plot zu schließen.

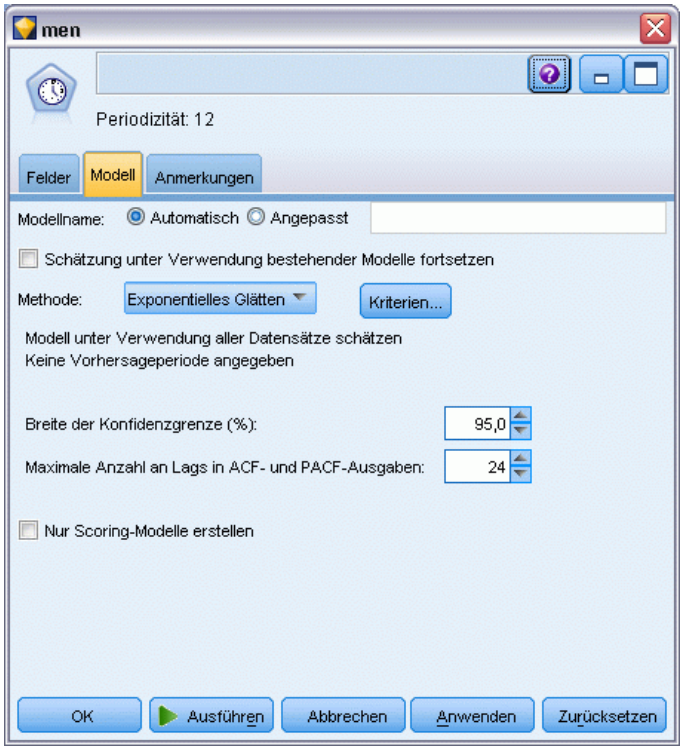
Nachdem Sie die Eigenschaften der Zeitreihe ermittelt haben, können Sie nun versuchen, sie zu modellieren. Das Verfahren der exponentiellen Glättung ist hilfreich für die Prognose von Zeitreihen, die Trend und/oder Saisonalität aufweisen. Wie wir gesehen haben, weisen die vorliegenden Daten beide Eigenschaften auf.

Exponentielles Glätten

Zur Konstruktion eines Modells der exponentiellen Glättung mit bester Anpassung gehört die Bestimmung des Modelltyps, also die Frage, ob das Modell Trend, Saisonalität oder beides enthalten muss, und die anschließende Ermittlung der am besten geeigneten Parameter für das ausgewählte Modell.

Das Diagramm für den Umsatz im Bereich Herrenbekleidung im Laufe der Zeit hat ein Modell mit linearer Trend-Komponente und multiplikativer Saisonalitätskomponente nahegelegt. Dies deutet auf ein Winter-Modell hin. Zunächst untersuchen wir jedoch ein einfaches Modell (ohne Trend und ohne Saisonalität) und anschließend ein Holt-Modell (der lineare Trend wird berücksichtigt, nicht jedoch die Saisonalität). Dadurch können Sie üben zu erkennen, wenn ein Modell keine gute Anpassung an die Daten darstellt. Dies ist eine entscheidende Fähigkeit für die erfolgreiche Modellerstellung.

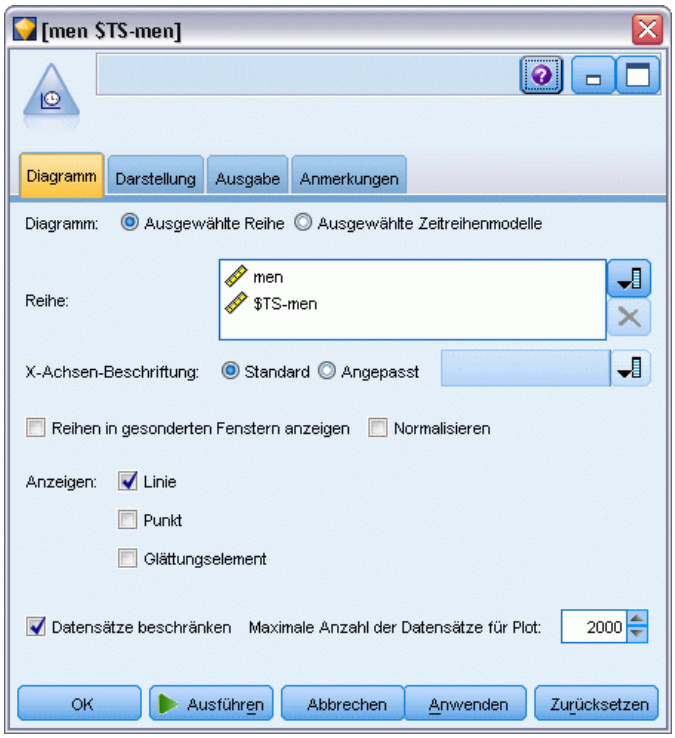
Abbildung 16-6
Angabe des exponentiellen Glättens



Wir beginnen mit einem einfachen Modell mit exponentiellem Glätten.

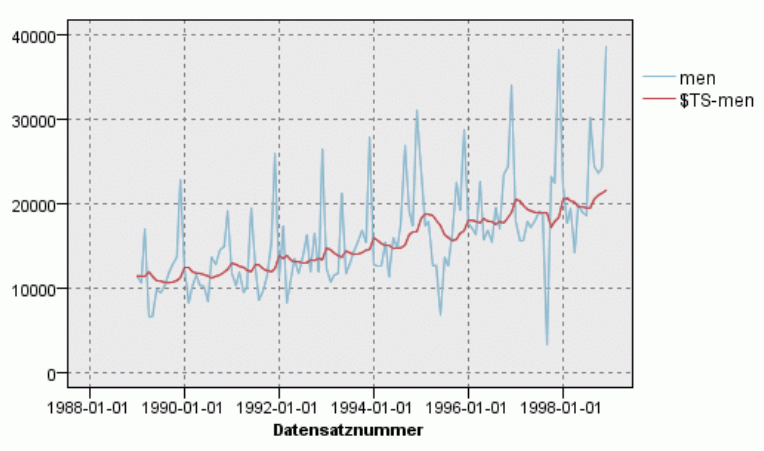
- Verbinden Sie einen Zeitreihenknoten mit dem Zeitintervallknoten.
- Legen Sie auf der Registerkarte Modell für Methode die Option Exponentielles Glätten fest.
- Klicken Sie auf Ausführen, um das Modell zu erstellen.

Abbildung 16-7
Plotten des Zeitreihenmodells



- Verbinden Sie einen Zeitdiagrammknoten mit dem Modell-Nugget.
- Fügen Sie auf der Registerkarte Plot*men* (Herren) und *$TS-men* zur Liste Reihe hinzu.
- Deaktivieren Sie die Kontrollkästchen Reihen in gesonderten Fenstern anzeigen und Normalisieren.
- Klicken Sie auf Ausführen.

Abbildung 16-8
Einfaches Modell mit exponentiellem Glätten

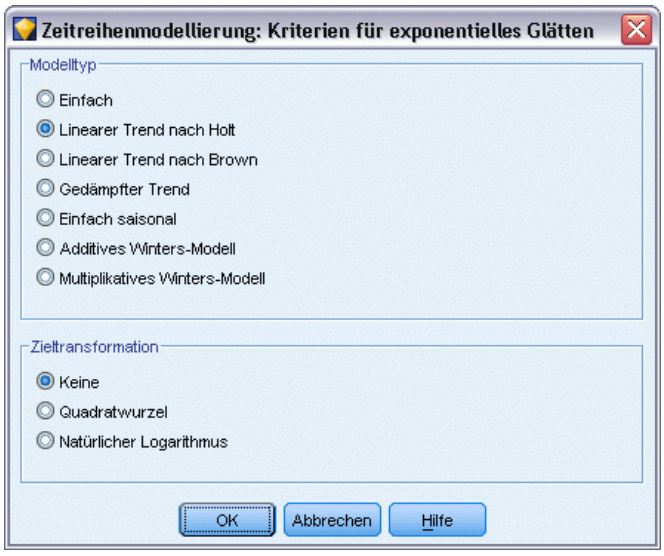


Der Plot men (Herren) stellt die aktuellen Daten dar, während $TS-men das Zeitreihenmodell angibt.

Das einfache Modell zeigt zwar einen graduellen (und ziemlich schwerfälligen) Aufwärtstrend, berücksichtigt jedoch keine Saisonalität. Sie können dieses Modell getrost verwerfen.

▶ Klicken Sie auf OK, um das Zeitdiagrammfenster zu schließen.

Abbildung 16-9
Auswahl des Holt-Modells

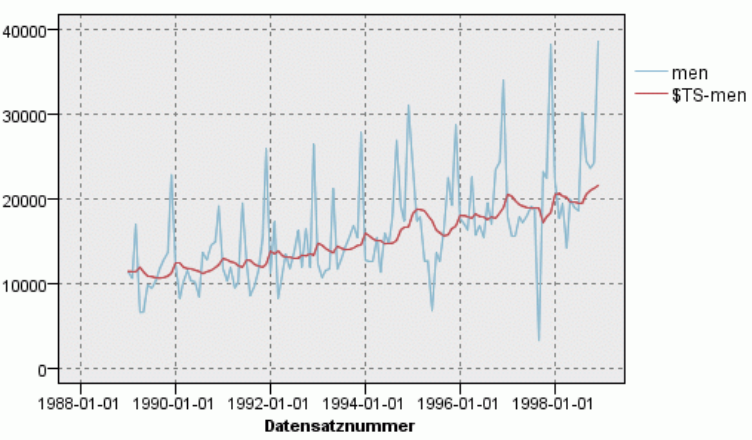


Testen wir das lineare Modell nach Holt. Dieses Modell sollte zumindest den Trend besser modellieren als das einfache Modell, obwohl auch hier die Saisonalität vermutlich nicht erfasst wird.

▶ Öffnen Sie den Zeitreihenknoten erneut.

- Klicken Sie auf der Registerkarte Modell bei aktivierter Auswahl von Exponentielles Glätten auf Kriterien.
- Wählen Sie im Dialogfeld “Kriterien für exponentielles Glätten” die Option Linearer Trend nach Holt aus.
- Klicken Sie auf OK, um das Dialogfeld zu schließen.
- Klicken Sie auf Ausführen, um das Modell-Nugget neu zu erstellen.
- Öffnen Sie erneut den Zeitdiagrammknoten und klicken Sie auf Ausführen.

Abbildung 16-10
Modell für den linearer Trend nach Holt

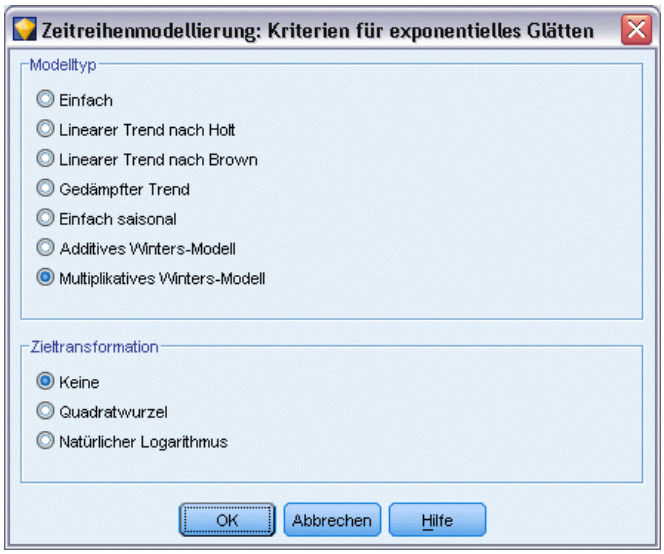


Das Holt-Modell zeigt einen gleichmäßigeren Aufwärtstrend als das einfache Modell, berücksichtigt jedoch noch immer nicht die Saisonalität. Daher können Sie auch dieses Modell verwerfen.

- Schließen Sie das Zeitdiagrammfenster.

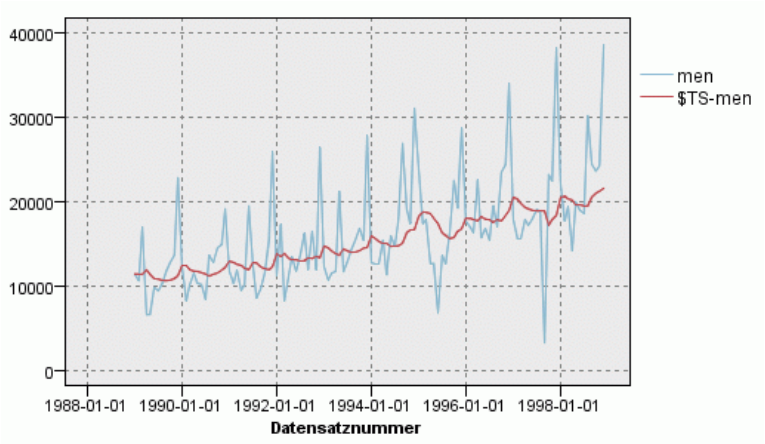
Sie erinnern sich, dass der ursprüngliche Plot für den Umsatz im Bereich Herrenbekleidung im Laufe der Zeit ein Modell nahelegte, das einen linearen Trend und multiplikative Saisonalität beinhaltet. Daher könnte das Winter-Modell ein geeigneterer Kandidat sein.

Abbildung 16-11
Auswahl des Winter-Modells



- Öffnen Sie den Zeitreihenknoten erneut.
- Klicken Sie auf der Registerkarte Modell bei aktivierter Auswahl von Exponentielles Glätten auf Kriterien.
- Wählen Sie im Dialogfeld “Kriterien für exponentielles Glätten” die Option Multiplikatives Winters-Modell aus.
- Klicken Sie auf OK, um das Dialogfeld zu schließen.
- Klicken Sie auf Ausführen, um das Modell-Nugget neu zu erstellen.
- Öffnen Sie den Zeitdiagrammknoten und klicken Sie auf Ausführen.

Abbildung 16-12
Multiplikatives Winters-Modell.



Dieses Modell sieht besser aus, es spiegelt sowohl den Trend als auch die Saisonalität der Daten wider.

Das Daten-Set deckt einen Zeitraum von 10 Jahren ab und enthält 10 saisonale Spitzen jeweils im Dezember der einzelnen Jahre. Die 10 Spitzen, die in den vorhergesagten Ergebnissen vorliegen, passen gut zu den 10 jährlichen Spitzen in den tatsächlichen Daten.

Die Ergebnisse zeigen jedoch auch die Grenzen des Verfahrens der exponentiellen Glättung auf. Wenn wir die nach oben und unten weisenden Spitzen betrachten, zeigt sich eine signifikante Struktur, die nicht berücksichtigt wurde.

Wenn Sie in erster Linie an der Modellierung eines langfristigen Trends mit saisonalen Schwankungen interessiert sind, kann das exponentielle Glätten eine gute Wahl sein. Wenn Sie eine komplexere Struktur modellieren möchten, wie beispielsweise die vorliegende, sollten Sie die Verwendung der Prozedur ARIMA in Erwägung ziehen.

ARIMA (X11 ARIMA)

Mit der Prozedur ARIMA können Sie ein Modell mit autoregressivem integriertem gleitendem Durchschnitt (AutoRegressive Integrated Moving Average) erstellen, das für eine feinabgestimmte Modellierung von Zeitreihen geeignet ist. ARIMA-Modelle bieten feinere Methoden für die Modellierung von Trend- und saisonalen Komponenten als die Modelle der exponentiellen Glättung und weisen den zusätzlichen Vorteil auf, dass Prädiktorvariablen in das Modell integriert werden können.

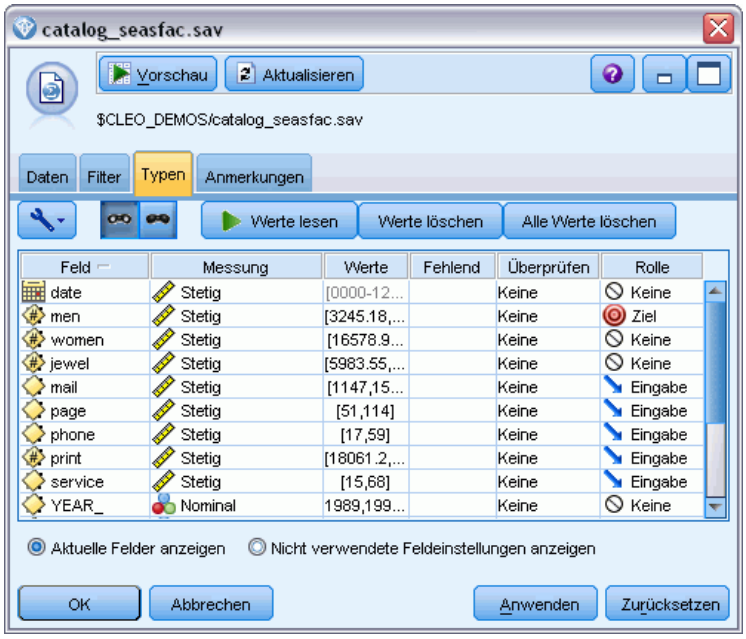
Bei der Fortsetzung des Beispiels des Versandhauses, das ein Prognosemodell entwickeln möchte, haben wir gesehen, wie das Unternehmen Daten zum monatlichen Umsatz im Bereich Herrenbekleidung sowie mehrere Zeitreihen gesammelt hat, die verwendet werden können, um einen Teil der Umsatzschwankungen zu erklären. Zu den möglichen Prädiktoren gehören die Anzahl der versendeten Kataloge und die Anzahl der Seiten im Katalog, die Anzahl der Telefonleitungen, über die eine Bestellung möglich ist, die Ausgaben für Werbung in Printmedien und die Anzahl der Kundendienstmitarbeiter.

Sind diese Einflussvariablen sinnvoll für die Prognostizierung? Ist ein Modell mit Prädiktoren wirklich besser als ein Modell ohne Prädiktoren? Mithilfe der Prozedur ARIMA können wir ein Prognosemodell mit Prädiktoren erstellen und untersuchen, ob gegenüber dem Modell mit exponentiellem Glätten ohne Prädiktoren ein signifikanter Unterschied in der Vorhersagekraft vorliegt.

Mit dem ARIMA-Verfahren können Sie eine Feinabstimmung des Modells durchführen, indem Sie die Ordnung für Autoregression, Differenzenbildung und gleitenden Durchschnitt sowie die saisonalen Gegenstücke dieser Komponenten angeben. Die manuelle Ermittlung der besten Werte für diese Komponenten kann ein zeitaufwendiger Vorgang sein, bei dem viel herumprobiert werden muss, daher lassen wir in diesem Beispiel das ARIMA-Modell automatisch durch den Expert Modeler auswählen.

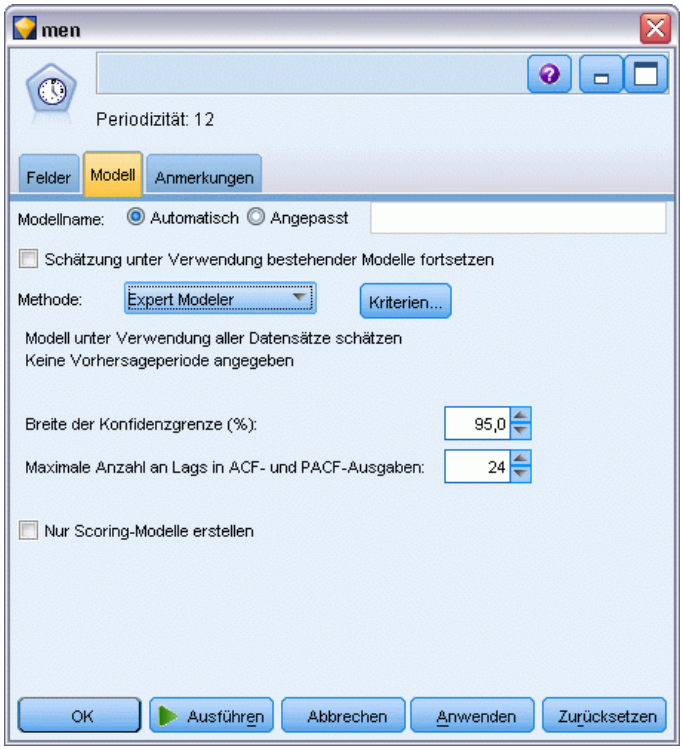
Wir versuchen, ein besseres Modell zu erstellen, indem wir einige der anderen Variablen im Daten-Set als Prädiktorvariablen behandeln. Als Prädiktoren am geeignetsten erscheinen folgende Variablen: Die Anzahl der versendeten Kataloge (*mail*), die Anzahl der Seiten im Katalog (*page*), die Anzahl der Telefonleitungen, über die eine Bestellung möglich ist (*phone*), die Ausgaben für Werbung in Printmedien *(print)* und die Anzahl der Kundendienstmitarbeiter *(service)*.

Abbildung 16-13
Festlegen der Prädiktorfelder



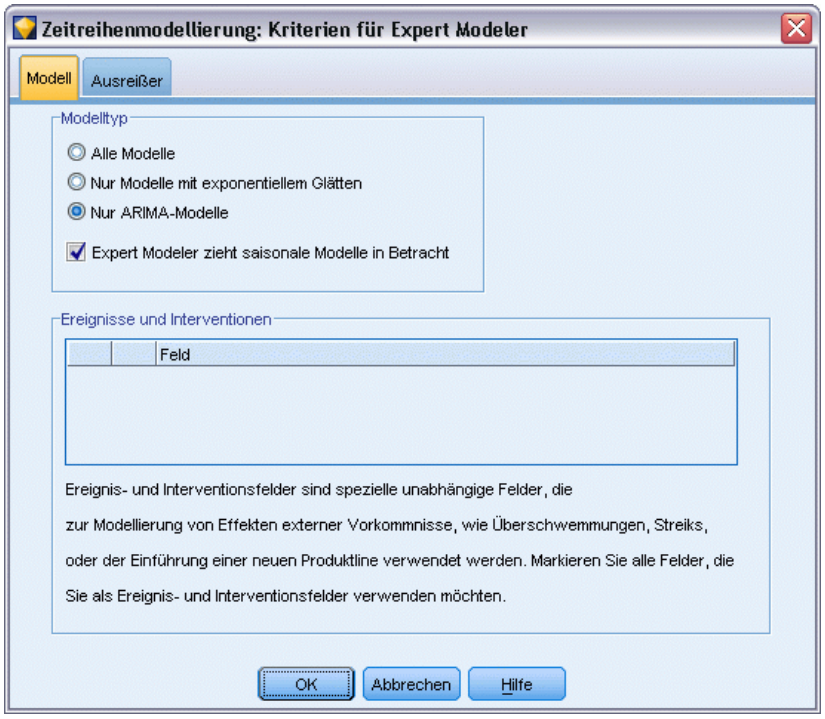
- Öffnen Sie den IBM® SPSS® Statistics-Dateiquellenknoten.
- Setzen Sie auf der Registerkarte “Typen” die *Rolle* für *mail*, *page*, *phone*, *print* und *service* auf Eingabe.
- Vergewissern Sie sich, dass die Rolle für men auf Ziel gesetzt ist und dass alle anderen Felder auf Keine gesetzt sind.
- Klicken Sie auf OK.

Abbildung 16-14
Auswahl des Expert Modeler



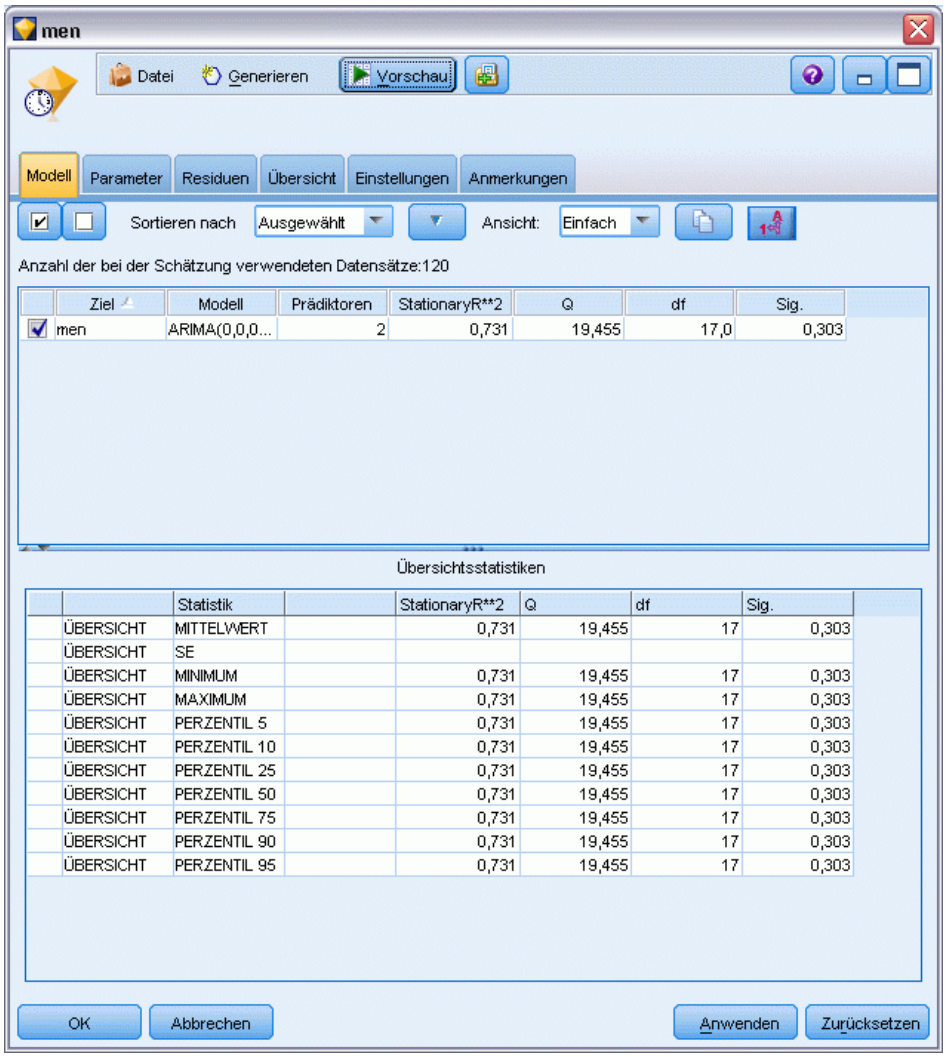
- Öffnen Sie den Zeitreihenknoten.
- Legen Sie auf der Registerkarte “Modell” für Methode die Option Expert Modeler fest und klicken Sie auf Kriterien.

Abbildung 16-15
Ausschließliche Auswahl von ARIMA-Modellen



- Wählen Sie im Dialogfeld “Kriterien für Expert Modeler” die Option Nur ARIMA-Modelle aus und vergewissern Sie sich, dass Expert Modeler zieht saisonale Modelle in Betracht aktiviert ist.
- Klicken Sie auf OK, um das Dialogfeld zu schließen.
- Klicken Sie auf Ausführen in der Registerkarte “Modell”, um das Modell-Nugget neu zu erstellen.

Abbildung 16-16
Der Expert Modeler wählt zwei Prädiktoren aus.



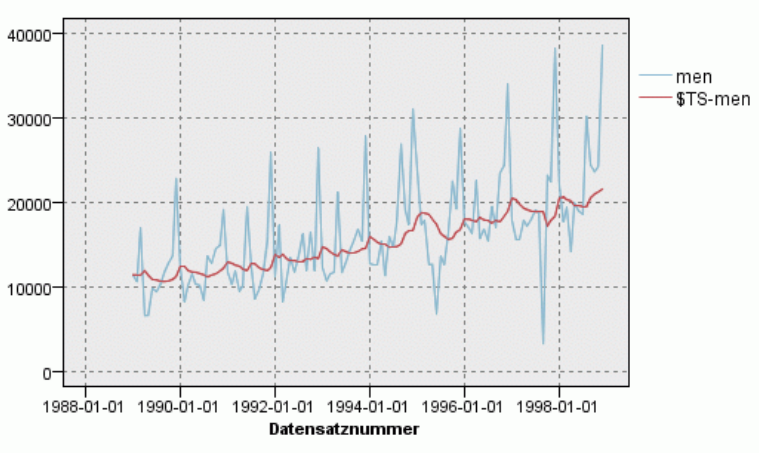
- Öffnen Sie das Modell-Nugget.

  Sie sehen, dass der Expert Modeler nur zwei der fünf angegebenen Prädiktoren als für das Modell signifikant ausgewählt hat.

- Klicken Sie auf OK, um das Modell-Nugget zu schließen.

- Öffnen Sie den Zeitdiagrammknoten und klicken Sie auf Ausführen.

Abbildung 16-17
ARIMA-Modell mit angegebenen Prädiktoren



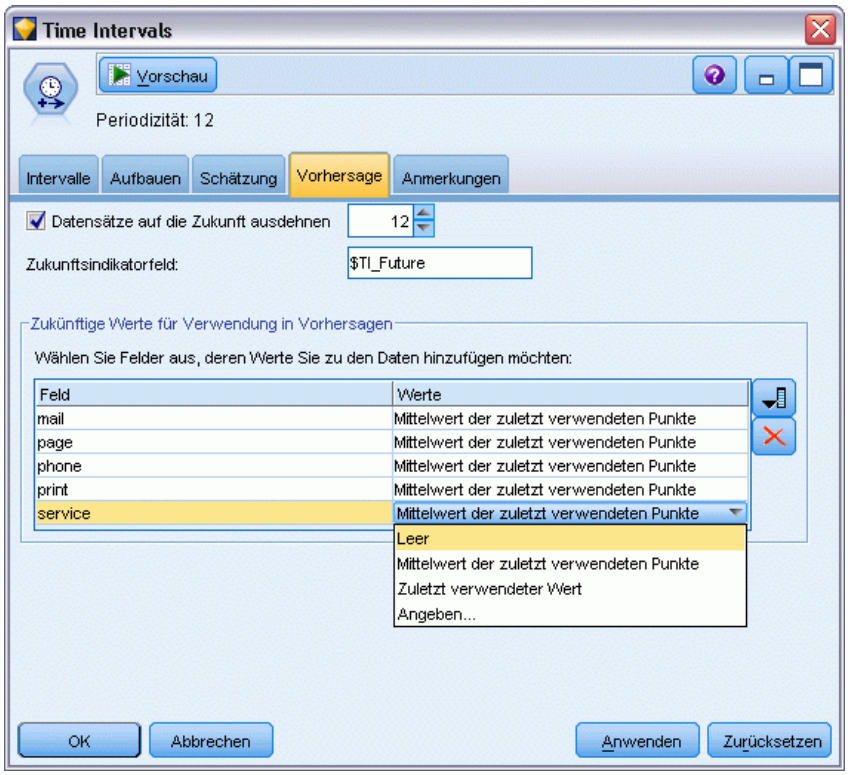
Dieses Modell stellt eine Verbesserung gegenüber dem vorherigen Modell dar, da auch die große Spitze nach unten erfasst wird. Damit stellt es die bisher beste Anpassung dar.

Wir könnten versuchen, das Modell noch weiter zu verfeinern, aber alle weiteren Verbesserungen würden wahrscheinlich nur noch äußerst gering sein. Wir haben festgstellt, dass das ARIMA-Modell mit Prädiktoren vorzuziehen ist. Daher werden wir nun das soeben erstellte Modell verwenden. In diesem Beispiel prognostizieren wir die Umsatzdaten für das kommende Jahr.

- Klicken Sie auf OK, um das Zeitdiagrammfenster zu schließen.
- Öffnen Sie den Zeitintervallknoten und wählen Sie die Registerkarte *Vorhersage*.
- Aktivieren Sie das Kontrollkästchen *Datensätze auf die Zukunft ausdehnen* und setzen Sie den Wert auf 12.

Die Verwendung von Prädiktoren bei der Vorhersage erfordert die Angabe von geschätzten Werten für die betreffenden Felder in der Vorhersageperiode, damit der Modeler das Zielfeld genauer vorhersagen kann.

Abbildung 16-18
Angabe zukünftiger Werte für Prädiktorfelder



- Klicken Sie in der Gruppe Zukünftige Werte für Verwendung in Vorhersagen auf die Feldauswahlschaltfläche rechts neben der Spalte "Werte".

- Wählen Sie im Dialogfeld "Felder auswählen" die Felder mail bis service aus und klicken Sie auf OK.

  In der Praxis würden Sie an dieser Stelle die zukünftigen Werte manuell eingeben, da diese fünf Prädiktoren sich jeweils auf Elemente beziehen, die in Ihrem Einflussbereich liegen. In diesem Beispiel verwenden wir jedoch eine der vordefinierten Funktionen, um nicht 12 Werte für jeden Prädiktor angeben zu müssen. (Wenn Sie genügend Erfahrung mit diesem Beispiel gesammelt haben, können Sie andere zukünftige Werte ausprobieren, um zu ermitteln, welche Wirkung sie auf das Modell haben.)

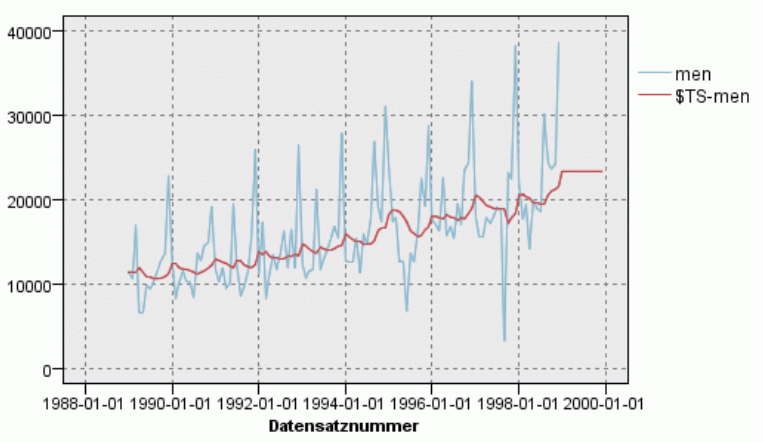
- Klicken Sie für jedes Feld abwechselnd auf das Feld Werte, um die Liste der möglichen Werte anzuzeigen, und wählen Sie die Option Mittelwert der zuletzt verwendeten Punkte. Diese Option berechnet den Mittelwert der letzten drei Datenpunkte für das betreffende Feld und verwendet ihn jeweils als geschätzten Wert.

- Klicken Sie auf OK.

- Öffnen Sie den Zeitreihenknoten und klicken Sie auf Ausführen, um das Modell-Nugget neu zu erstellen.
- Öffnen Sie den Zeitdiagrammknoten und klicken Sie auf Ausführen.

Die Prognose für 1999 sieht gut aus. Wie erwartet kehren die Umsatzzahlen nach der Spitze im Dezember wieder auf das normale Niveau zurück und es liegt ein stetiger Aufwärtstrend in der zweiten Jahreshälfte vor, wobei der Umsatz im Allgemeinen signifikant über dem des Vorjahrs liegt.

Abbildung 16-19
Umsatzprognose mit angegebenen Prädiktoren



Zusammenfassung

Sie haben erfolgreich eine komplexe Zeitreihe modelliert, die nicht nur einen Aufwärtstrend, sondern auch saisonale und andere Schwankungen beinhaltet. Außerdem haben Sie erfahren, wie Sie durch systematisches Ausprobieren eine immer engere Annäherung an ein genaues Modell erreichen können. Dieses Modell haben Sie anschließend zur Vorhersage des zukünftigen Umsatzes verwendet.

In der Praxis müssten Sie das Modell jedes Mal erneut anwenden, wenn die tatsächlichen Umsatzdaten aktualisiert werden (beispielsweise jeden Monat oder jedes Quartal), und aktualisierte Vorhersagen erstellen. Für weitere Informationen siehe Thema Erneutes Anwenden eines Zeitreihenmodells in Kapitel 15 auf S. 202.

Kapitel
17

Erstellen von Angeboten für Kunden (Selbstlernfunktion)

Der Selbstlern-Antwortmodellknoten (Self-Learning Response Model-(SLRM-)Knoten) generiert und aktiviert die Aktualisierung eines Modells, mit dem Sie prognostizieren können, welche Angebote für die Kunden am geeignetsten sind und mit welcher Wahrscheinlichkeit die Angebote angenommen werden. Modelle dieser Art sind am nützlichsten für Customer Relationship Management, beispielsweise in Marketinganwendungen oder im Callcenter.

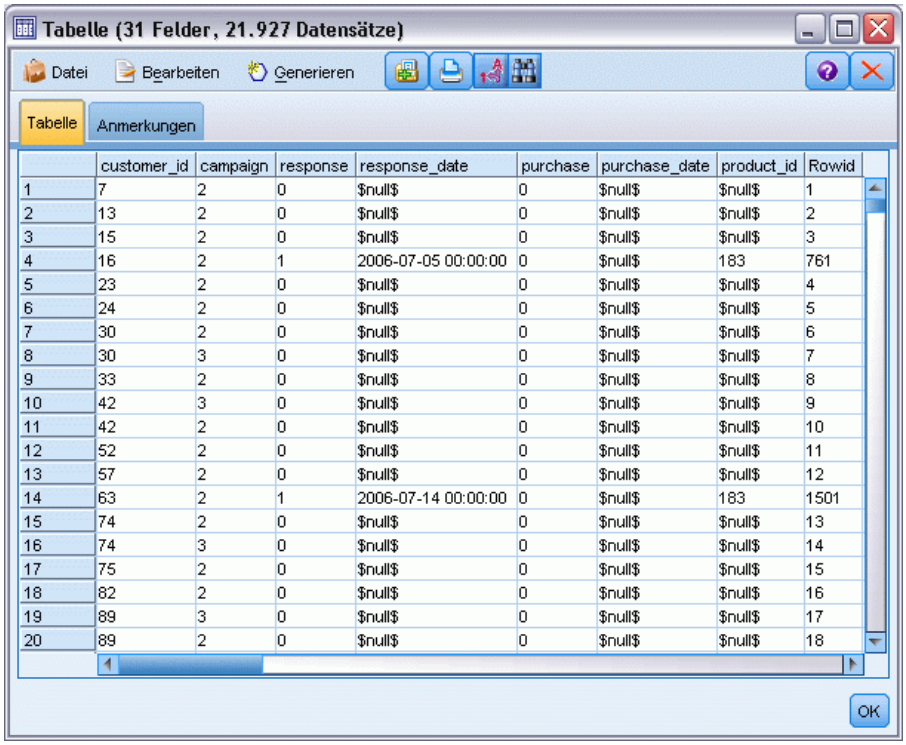
Dieses Beispiel beruht auf einem fiktiven Kreditinstitut. Die Marketingabteilung möchte in zukünftigen Kampagnen profitablere Ergebnisse erzielen, indem jedem Kunden ein speziell für ihn geeignetes Angebot an Finanzdienstleistungen unterbreitet wird. Insbesondere wird in dem Beispiel ein Selbstlern-Antwortmodell verwendet, mit dem auf der Grundlage früherer Angebote und Reaktionen die Eigenschaften der Kunden ermittelt werden, die mit der größten Wahrscheinlichkeit positiv reagieren werden, und auf der Grundlage der Ergebnisse das beste aktuelle Angebot beworben wird.

In diesem Beispiel wird der Stream *pm_selflearn.str* verwendet, der Bezug nimmt auf die Datendateien *pm_customer_train1.sav*, *pm_customer_train2.sav* und *pm_customer_train3.sav*. Die Dateien stehen im Ordner *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *pm_selflearn.str* befindet sich im Ordner *streams*.

Bestehende Daten

Das Unternehmen hat Daten über die Angebote aufgezeichnet, die den Kunden in früheren Kampagnen unterbreitet wurden, sowie über die Reaktionen auf diese Angebote. Diese Daten umfassen auch demografische Informationen und Finanzdaten, mit denen die Antwortquoten für verschiedene Kunden prognostiziert werden können.

Abbildung 17-1
Reaktionen auf frühere Angebote

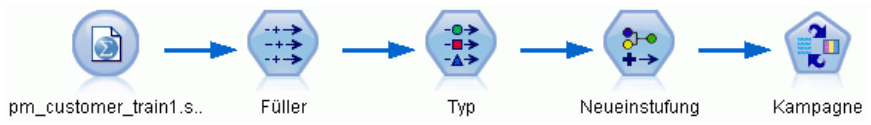


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

Erstellen des Streams

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf die Datei *pm_customer_train1.sav* im Ordner *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist.

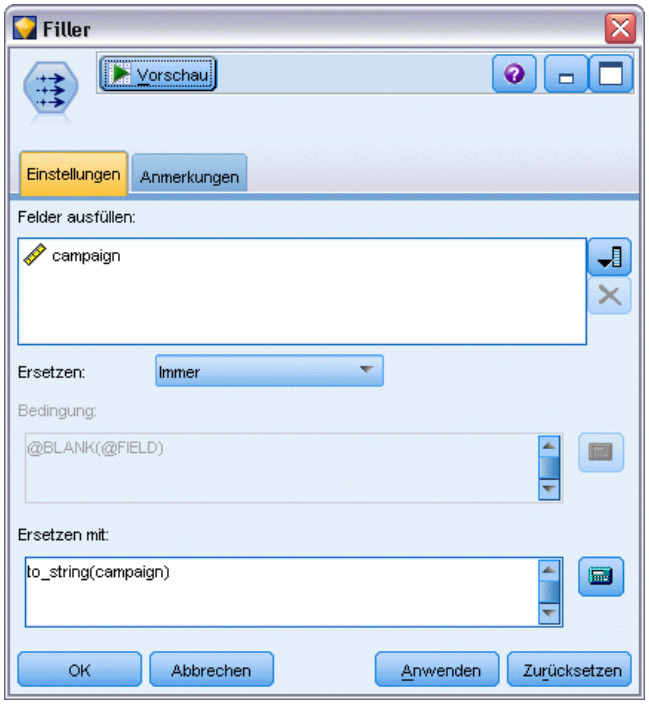
Abbildung 17-2
SLRM-Beispiel-Stream



- Fügen Sie einen Füllerknoten hinzu und wählen Sie campaign (Kampagne) als das Ausfüllfeld aus.
- Wählen Sie den Ersetzungstyp Immer aus.

▶ Geben Sie im Textfeld “Ersetzen mit” to_string(campaign) ein und klicken Sie auf OK.

Abbildung 17-3
Ableiten eines Kampagnenfelds



- Fügen Sie einen Typknoten hinzu und setzen Sie für die Felder *customer_id*, *response_date*, *purchase_date*, *product_id*, *Rowid* und *X_random* den Wert von *Rolle* auf Keine.

Abbildung 17-4
Ändern der Einstellungen für Typknoten

- Setzen Sie für die Felder *campaign* und *response* den Wert von *Rolle* auf Ziel. Auf diesen Feldern sollen die Vorhersagen beruhen.

  Setzen Sie die Messung für das Feld *response* (Antwort) auf den Wert Flag.

- Klicken Sie auf Werte lesen und dann auf OK.

  Da die Daten des Kampagnenfelds eine Liste mit Zahlen (1, 2, 3 und 4) enthalten, können Sie die Felder umkodieren, um ihnen aussagekräftigere Titel zu geben.

- Fügen Sie einen Umkodierungsknoten zum Typknoten hinzu.

- Wählen Sie im Feld Umkodieren die Option Vorhandenes Feld.

- Wählen Sie im Feld Umkodieren die Option campaign.

- Klicken Sie auf die Schaltfläche Ermitteln. Die Kampagnenwerte werden zur Spalte *Ursprünglicher Wert* hinzugefügt.

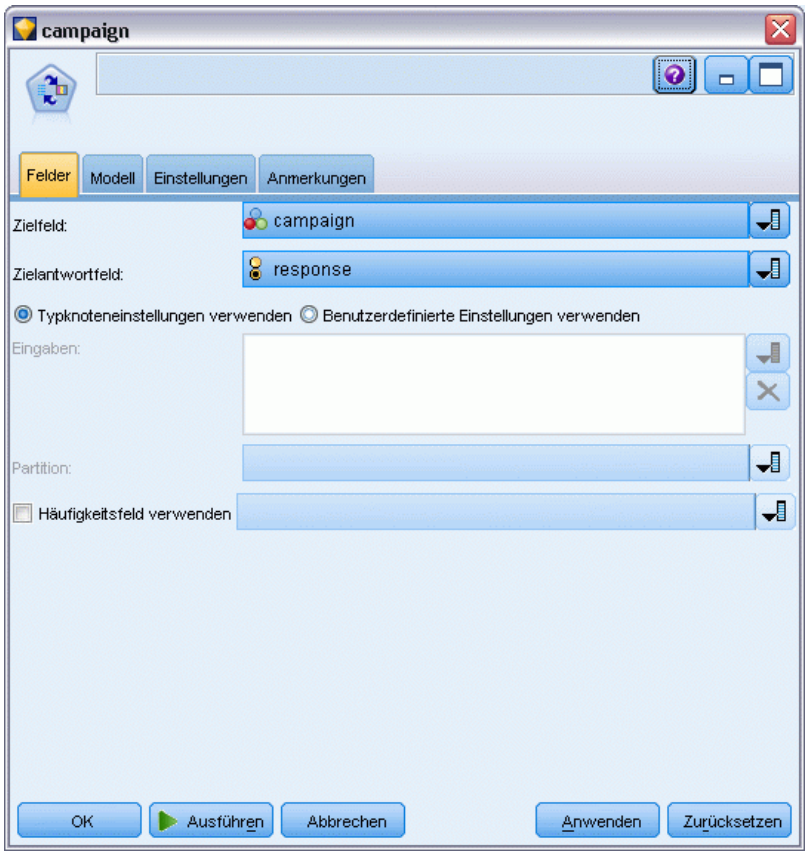
- Geben Sie in der Spalte *Neuer Wert* folgende Kampagnennamen in die ersten vier Zeilen ein:
  - Mortgage
  - Car loan
  - Savings
  - Pension

▶ Klicken Sie auf OK.

Abbildung 17-5
Umkodieren der Kampagnennamen

- Verbinden Sie einen SLRM-Modellierungsknoten mit dem Umkodierungsknoten. Wählen Sie auf der Registerkarte “Felder” als Zielfeld campaign und als Zielantwortfeld die Option response aus.

Abbildung 17-6
Auswahl von Ziel und Zielantwort

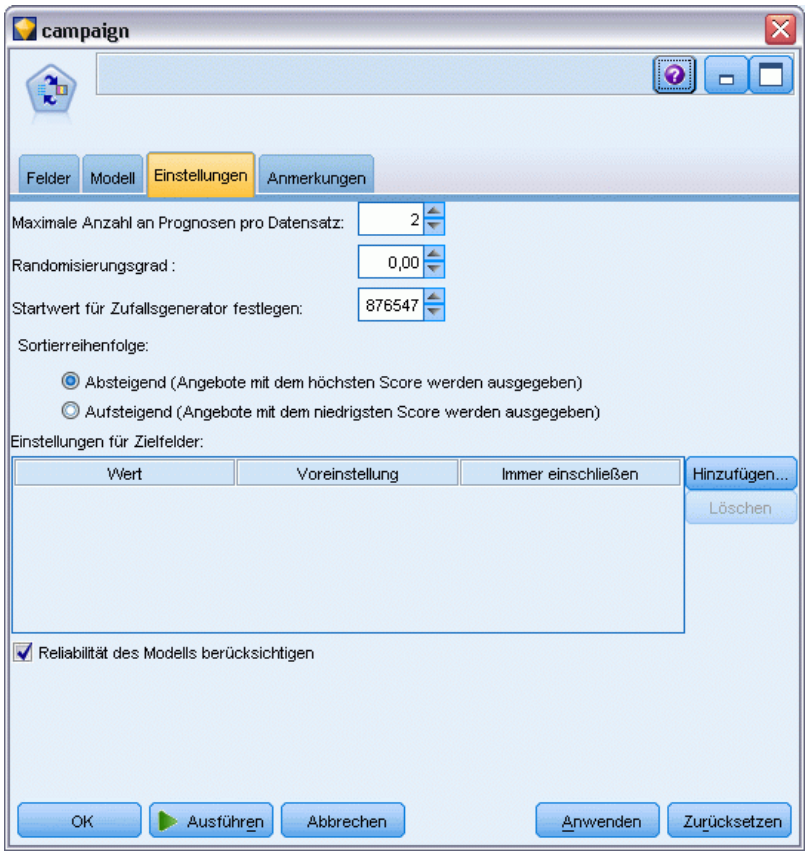


- Reduzieren Sie auf der Registerkarte “Einstellungen” im Feld “Maximale Anzahl an Prognosen pro Datensatz” den Wert auf 2.

  Auf diese Weise werden für jeden Kunden zwei Angebote ermittelt, bei denen die Wahrscheinlichkeit für die Annahme am höchsten ist.

▶ Vergewissern Sie sich, dass die Option Reliabilität des Modells berücksichtigen ausgewählt ist, und klicken Sie auf Ausführen.

Abbildung 17-7
SLRM-Knoten – Einstellungen



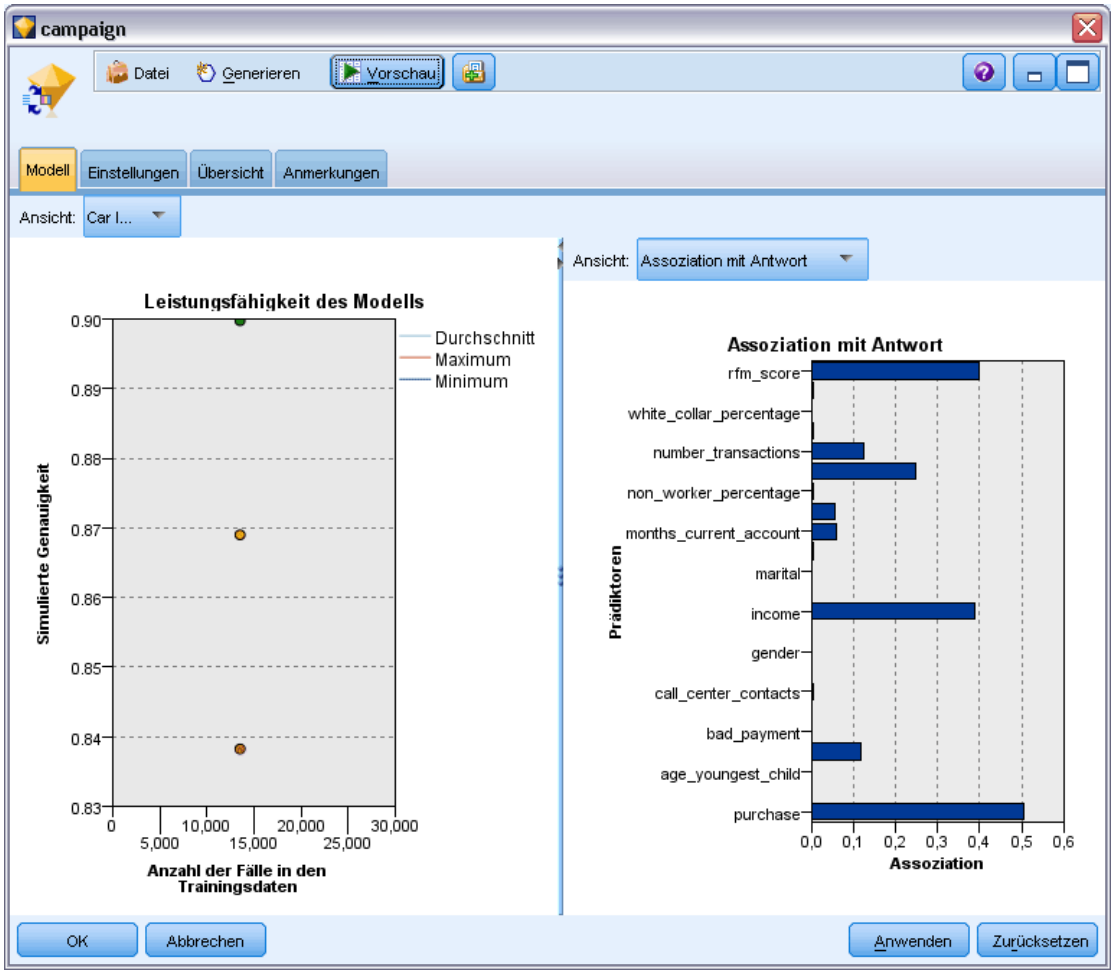
Durchsuchen des Modells

- Öffnen Sie das Modell-Nugget. Die Registerkarte “Modell” zeigt anfangs die geschätzte Genauigkeit der Vorhersagen für die einzelnen Angebote und die relative Wichtigkeit der einzelnen verwendeten Prädiktoren beim Schätzen des Modells.

  Zur Anzeige der Korrelation für jeden Prädiktor zur Zielvariablen wählen Sie Assoziation mit Antwort aus der Liste Ansicht im rechten Fensterbereich.

- Um zwischen den vier Angeboten zu wechseln, für die Vorhersagen vorhanden sind, wählen Sie das erforderliche Angebot aus der Liste Ansicht im linken Fensterbereich.

Abbildung 17-8
SLRM-Modell-Nugget

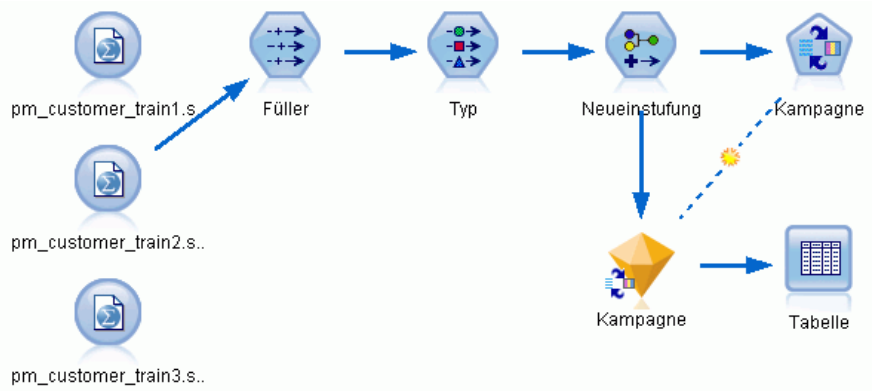


- Schließen Sie das Modell-Nugget-Fenster.

- Trennen Sie im Stream-Zeichenbereich den IBM® SPSS® Statistics-Dateiquellenknoten, der auf *pm_customer_train1.sav* verweist.

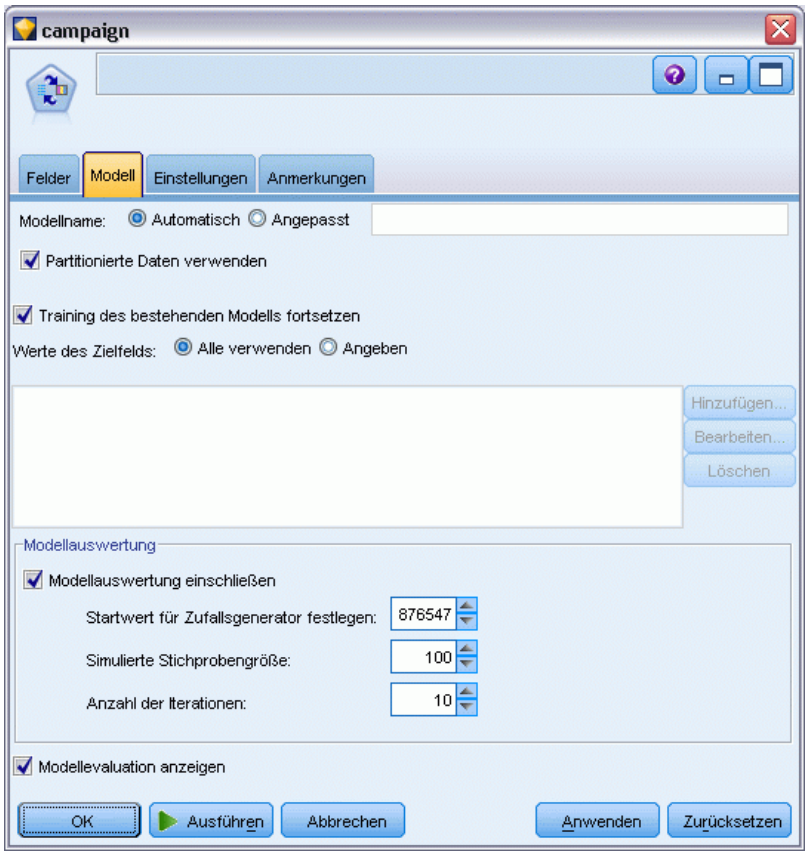
- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf die Datei *pm_customer_train2.sav* im Ordner *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist, und verbinden Sie ihn mit dem Füllerknoten.

Abbildung 17-9
Hinzufügen einer zweiten Datenquelle zu einem SLRM-Stream



- Wählen Sie auf der Registerkarte "Modell" des SLRM-Knotens die Option Training des bestehenden Modells fortsetzen.

Abbildung 17-10
Training des Modells fortsetzen

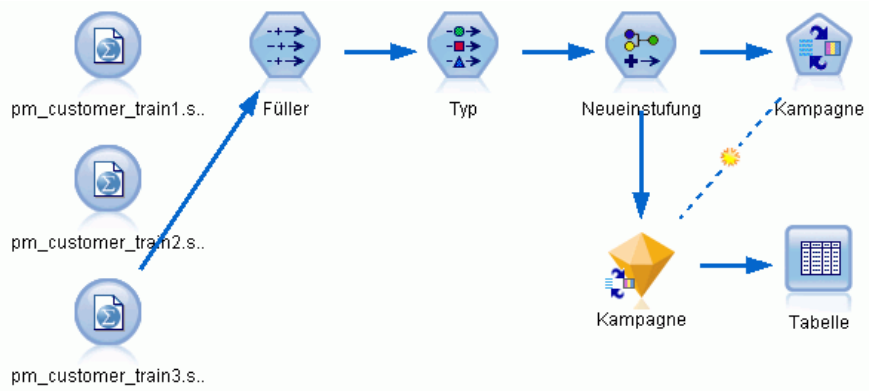


- Klicken Sie auf Ausführen, um das Modell-Nugget neu zu erstellen. Um die zugehörigen Details anzuzeigen, können Sie auf das Nugget in der Zeichenfläche doppelklicken.

  Auf der Registerkarte “Modell” werden nun die revidierten Schätzer für die Genauigkeit der Vorhersagen für die einzelnen Angebote angezeigt.

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf die Datei *pm_customer_train3.sav* im Ordner *Demos* Ihrer SPSS Modeler-Installation verweist, und verbinden Sie ihn mit dem Füllerknoten.

Abbildung 17-11
Hinzufügen einer dritten Datenquelle zu einem SLRM-Stream

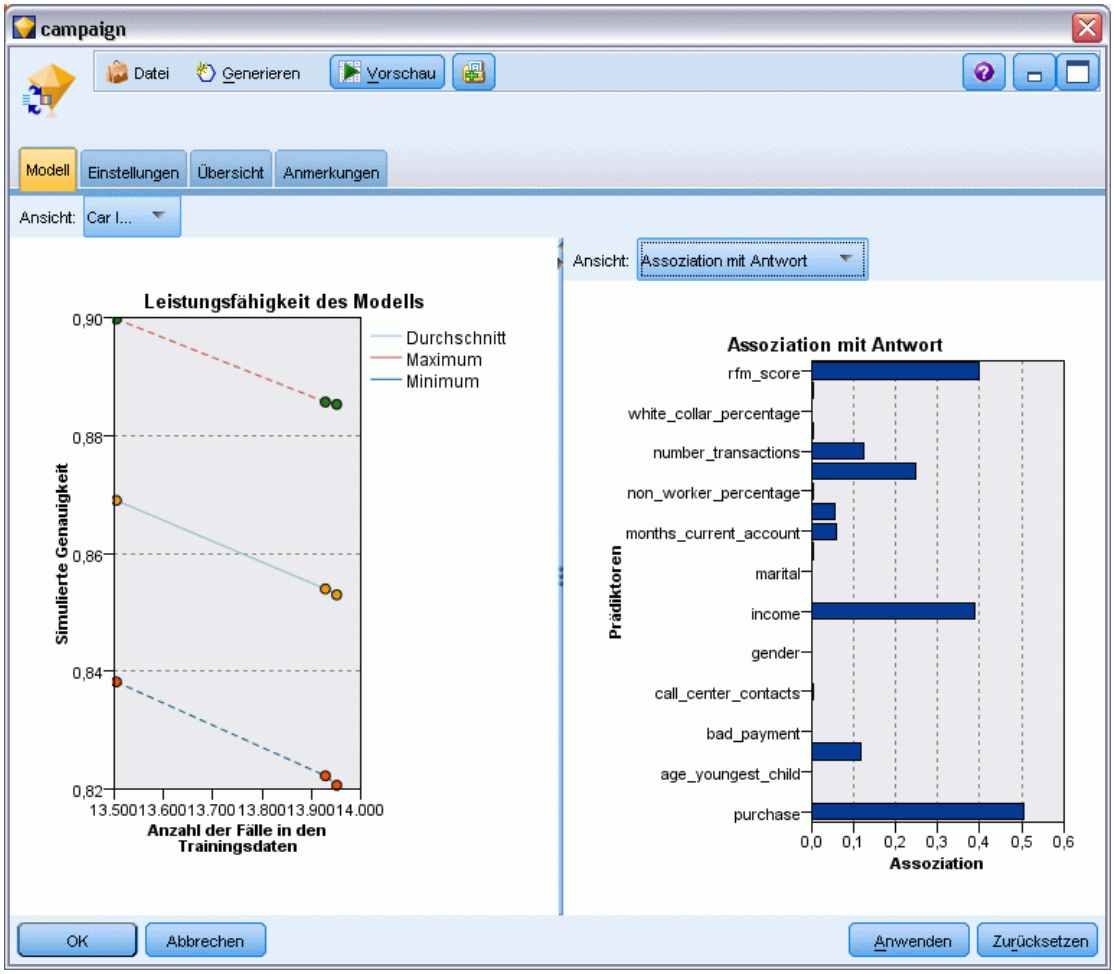


- Klicken Sie auf Ausführen, um das Modell-Nugget noch einmal neu zu erstellen. Um die zugehörigen Details anzuzeigen, können Sie auf das Nugget in der Zeichenfläche doppelklicken.

- Auf der Registerkarte “Modell” wird nun die endgültige geschätzte Genauigkeit der Vorhersagen für die einzelnen Angebote angezeigt.

Wie Sie sehen können, nahm die durchschnittliche Genauigkeit (von 86,9 % auf 85,4 %) geringfügig ab, als Sie die zusätzlichen Datenquellen hinzufügten. Bei dieser Fluktuation handelt es sich jedoch um einen minimalen Wert, der geringfügigen Anomalien innerhalb der verfügbaren Daten zugeschrieben werden kann.

Abbildung 17-12
Aktualisiertes SLRM-Modell-Nugget

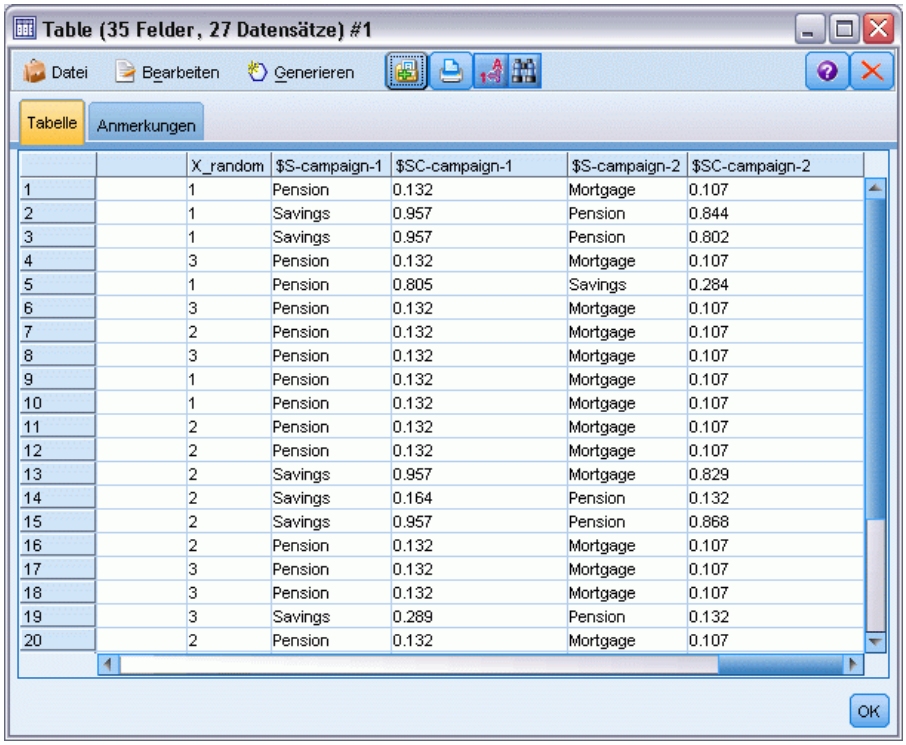


- Verbinden Sie einen Tabellenknoten mit dem letzten (dritten) generierten Modell und führen Sie den Tabellenknoten aus.
- Führen Sie einen Bildlauf zur rechten Seite der Tabelle durch. Die Vorhersagen zeigen, welche Angebote ein Kunde mit der größten Wahrscheinlichkeit annimmt, sowie die Konfidenz für die Annahme, je nach den Details der einzelnen Kunden.

  So liegt beispielsweise in der ersten Zeile der gezeigten Tabelle nur ein Konfidenzwert von 13,2 % (Wert 0,132 in der Spalte *$SC-campaign-1*) dafür vor, dass ein Kunde, der zuvor einmal einen Kredit für ein Auto aufgenommen hat, ein Angebot für einen Rentensparplan annehmen würde. Die zweite und die dritte Zeile zeigen jedoch zwei weitere Kunden, die ebenfalls einen Kredit für ein Auto aufgenommen haben; dort liegt ein Konfidenzwert von 95,7 % vor, dass sie und

andere Kunden mit einer ähnlichen Vorgeschichte ein Sparkonto eröffnen würden, wenn ihnen dies angeboten würde, und ein Konfidenzwert von mehr als 80 %, dass Sie einen Rentensparplan annehmen würden.

Abbildung 17-13
Modellausgabe – vorhergesagte Angebote und Konfidenzwerte



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

Erläuterungen der mathematischen Grundlagen für die in SPSS Modeler verwendeten Modellierungsmethoden finden Sie im *SPSS Modeler-Algorithmushandbuch*, das sich im Verzeichnis *Documentation* auf dem Produkt-DVD befindet.

Beachten Sie außerdem, dass diese Ergebnisse nur auf den Trainingsdaten beruhen. Um einzuschätzen, wie gut sich das Modell für andere Daten in der Praxis verallgemeinern lässt, könnten Sie mit einem Partitionsknoten eine Teilmenge der Datensätze für Test- und Validierungszwecke zurückhalten. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in *IBM SPSS Modeler 14.2- Quellen- , Prozess- und Ausgabeknoten*. Weitere Informationen zum SLRM-Knoten finden Sie hier: Kapitel 14 in der Knotenreferenz.

Kapitel
18

Vorhersage von Kreditausfällen (Bayes-Netzwerk)

Mithilfe des Knotens “Bayes-Netzwerk” können Sie ein Wahrscheinlichkeitsmodell erstellen, indem Sie beobachtete und aufgezeichnete Hinweise mit Weltwissen (“gesundem Menschenverstand”) kombinieren, um die Wahrscheinlichkeit des Vorkommens unter Verwendung scheinbar nicht miteinander verknüpfter Attribute zu ermitteln.

In diesem Beispiel wird ein Stream namens *bayes_bankloan.str* verwendet, der Bezug nimmt auf die Datendatei *bankloan.sav*. Diese Dateien finden Sie im Verzeichnis *Demos* jeder IBM® SPSS® Modeler-Installation. Sie können auch über die IBM® SPSS® Modeler-Programmgruppe im Windows-Startmenü aufgerufen werden. Die Datei *bayes_bankloan.str* befindet sich im Verzeichnis *streams*.

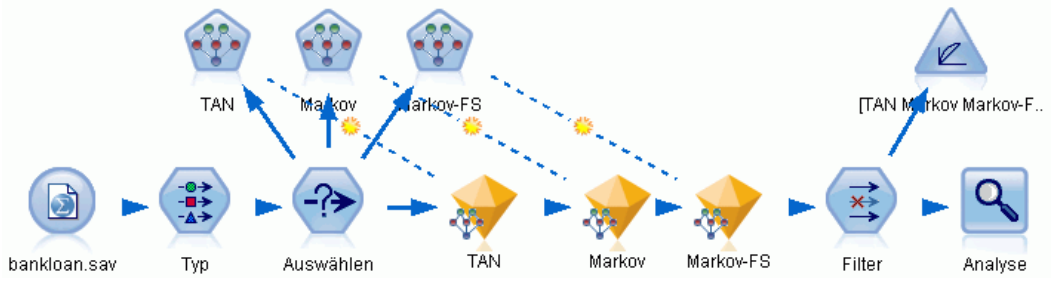
Nehmen Sie beispielsweise an, dass eine Bank Bedenken wegen Krediten hat, die möglicherweise nicht zurückgezahlt werden. Wenn Daten über frühere Kreditausfälle verwendet werden können, um vorherzusagen, welche Kunden mit hoher Wahrscheinlichkeit Probleme bei der Rückzahlung von Krediten haben werden, können diesen Kunden, die ein “hohes Risiko” aufweisen, Kredite verweigert oder alternative Produkte angeboten werden.

Dieses Beispiel konzentriert sich auf die Verwendung bestehender Daten zu Kreditausfällen zur Vorhersage potenziell zahlungsunfähiger Personen für die Zukunft. Dabei werden drei verschiedene Typen von Bayes-Netzwerk-Modellen untersucht, um zu ermitteln, welches in dieser Situation die besseren Prognosen bietet.

Erstellen des Streams

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *bankloan.sav* im Ordner *Demos* verweist.

Abbildung 18-1
Beispiel-Stream für Bayes-Netzwerk



- Gliedern Sie einen Typknoten an den Quellenknoten an und setzen Sie die Rolle des Standard-Felds auf Ziel. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

▶ Klicken Sie auf die Schaltfläche Werte lesen, um die Spalte *Werte* auszufüllen.

Abbildung 18-2
Auswahl des Zielfelds

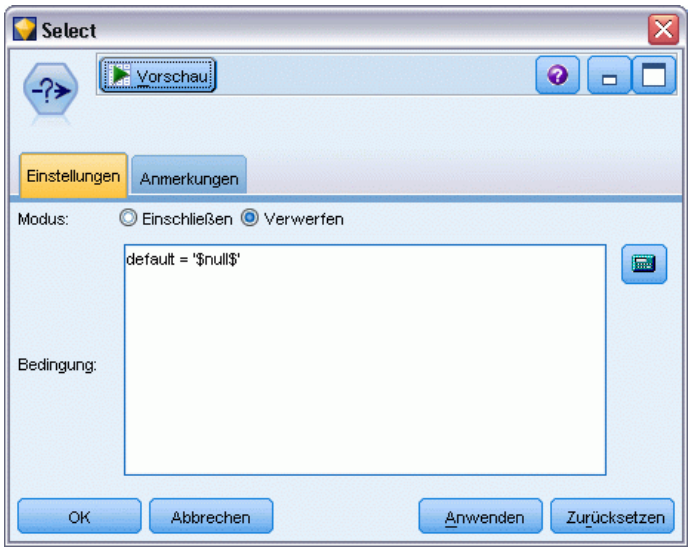
Fälle, bei denen das Ziel einen Nullwert aufweist, sind beim Erstellen des Modells nutzlos. Die können derartige Fälle ausschließen, damit sie nicht bei der Modellevaluation verwendet werden.

▶ Fügen Sie einen Auswahlknoten zum Typknoten hinzu.

▶ Wählen Sie als Modus die Option Verwerfen.

▶ Geben Sie im Feld “Bedingung” default = ‘$null$’ ein.

Abbildung 18-3
Verwerfen von Null-Zielen



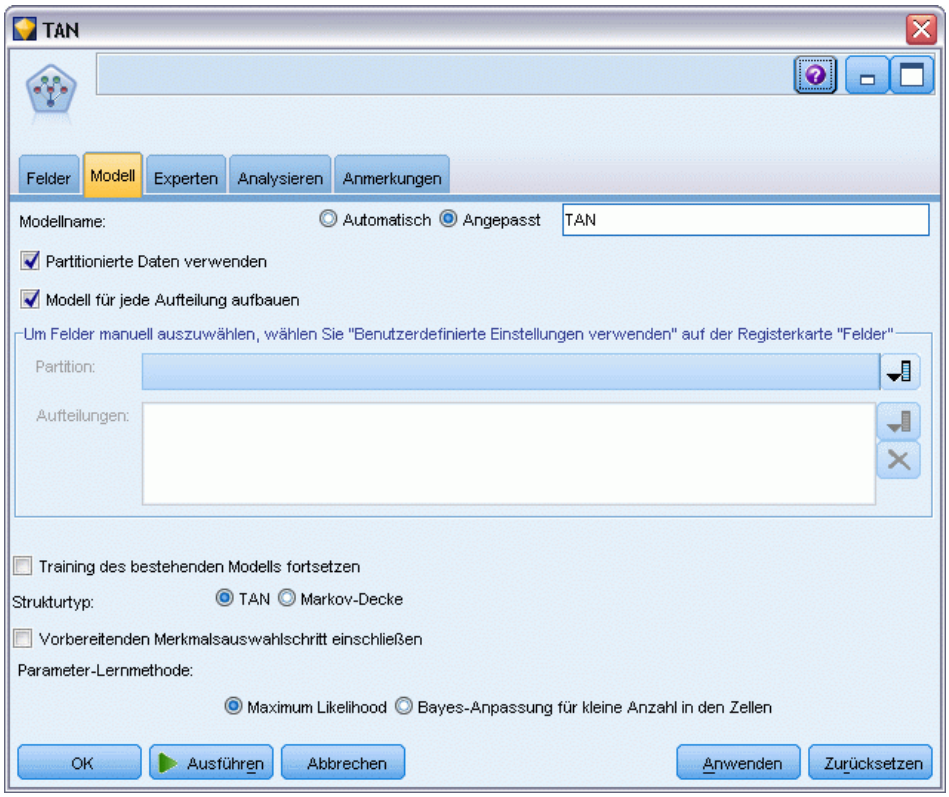
Da Sie mehrere verschiedene Typen von Bayes-Netzwerken erstellen können, lohnt es sich mehrere davon zu vergleichen, um zu ermitteln, welches Modell die besten Prädiktoren bietet. Als erstes soll ein Modell vom Typ “Tree Augmented Naïve Bayes” (TAN) erstellt werden.

▶ Gliedern Sie einen Bayes-Netzwerk-Knoten an den Auswahlknoten an.

▶ Wählen Sie auf der Registerkarte “Modell” als Modellnamen Benutzerdefiniert (Angepasst) und geben Sie im Textfeld den Ausdruck TAN ein.

- Wählen Sie als Strukturtyp TAN aus und klicken Sie auf OK.

Abbildung 18-4
Erstellen eines Modells vom Typ "Tree Augmented Naïve Bayes"

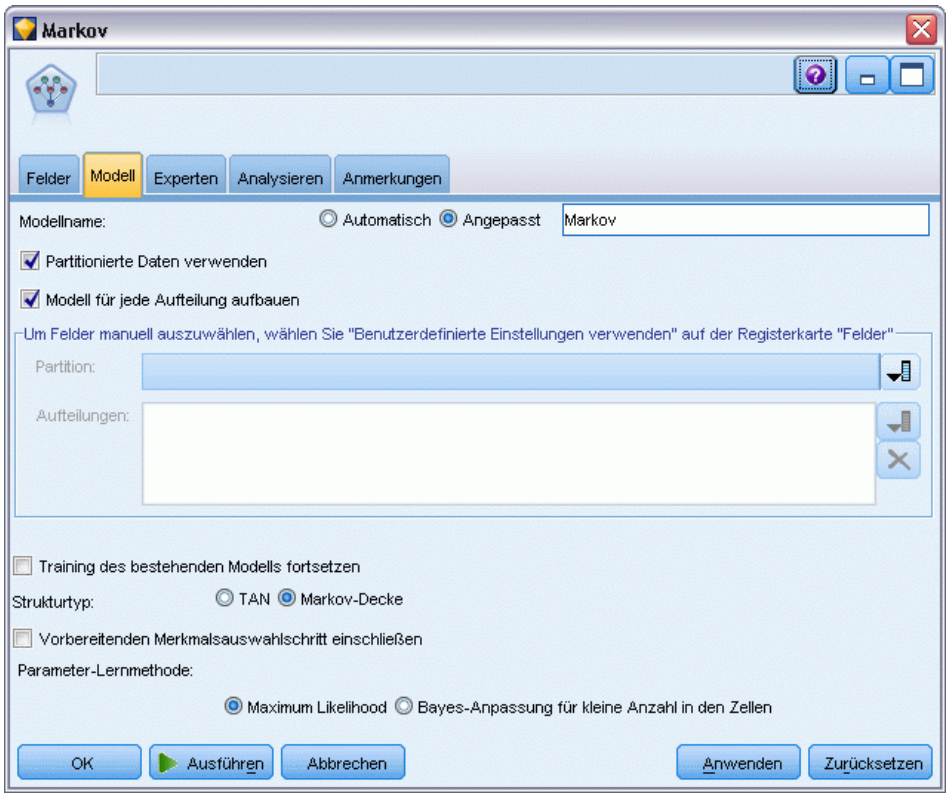


Der zweite zu erstellende Modelltyp weist eine Struktur vom Typ "Markov-Decke" auf.

- Gliedern Sie einen zweiten Bayes-Netzwerk-Knoten an den Auswahlknoten an.
- Wählen Sie auf der Registerkarte "Modell" als Modellnamen Benutzerdefiniert (Angepasst) aus und geben Sie im Textfeld den Ausdruck Markov ein.

- Wählen Sie als Strukturtyp Markov Blanket aus und klicken Sie auf OK.

Abbildung 18-5
Erstellen eines Modells vom Typ "Markov-Decke"

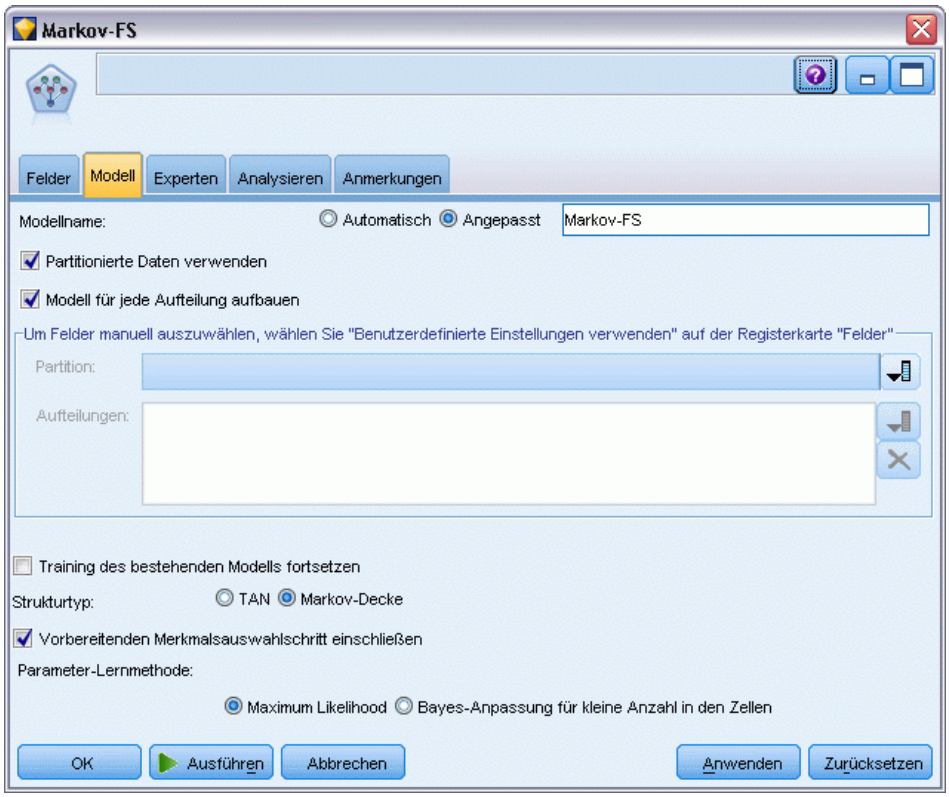


Das dritte zu erstellende Modell weist eine Struktur vom Typ "Markov-Decke" auf und verwendet außerdem eine vorbereitende Merkmalsauswahl, um die Eingaben auszuwählen, die in einer signifikanten Beziehung zur Zielvariablen stehen.

- Gliedern Sie einen dritten Bayes-Netzwerk-Knoten an den Auswahlknoten an.
- Wählen Sie auf der Registerkarte "Modell" als Modellnamen Benutzerdefiniert (Angepasst) aus und geben Sie im Textfeld den Ausdruck Markov-FS ein.
- Wählen Sie als Strukturtyp Markov-Decke aus.

- Wählen Sie die Option Vorbereitenden Merkmalsauswahlschritt einschließen aus und klicken Sie auf OK.

Abbildung 18-6
Erstellen eines Modells vom Typ "Markov-Decke" mit vorbereitender Merkmalsauswahl



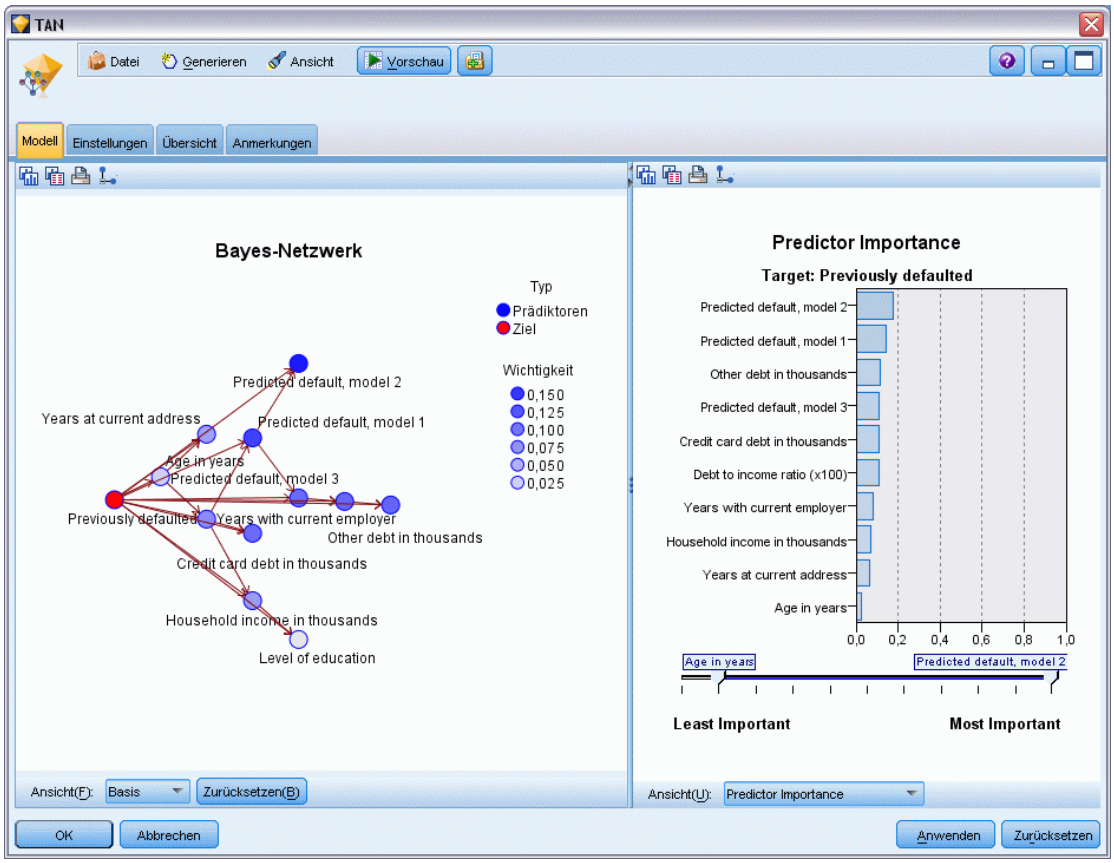
Durchsuchen des Modells

- Führen Sie den Stream aus, um die Modell-Nuggets zu generieren; diese werden dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die zugehörigen Details anzuzeigen, können Sie auf die Modell-Nuggets im Stream doppelklicken.

  Die Registerkarte "Modell" des Modell-Nuggets gliedert sich in zwei Bereiche. Der linke Bereich enthält ein Netzwerkdiagramm mit Knoten, das die Beziehung zwischen dem Ziel und seinen wichtigsten Prädiktoren sowie die Beziehung zwischen den Prädiktoren anzeigt.

Der rechte Bereich zeigt entweder die *Bedeutsamkeit der Prädiktoren*, also die relative Wichtigkeit der einzelnen Prädiktoren bei der Schätzung des Modells, oder die *Konditionalen Wahrscheinlichkeiten*, also den Wert der bedingten Wahrscheinlichkeit für die einzelnen Knoten und jede Kombination von Werten in ihren übergeordneten Knoten.

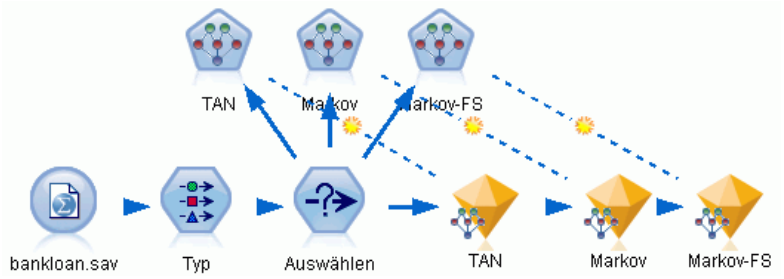
Abbildung 18-7
Anzeigen eines Modells vom Typ "Tree Augmented Naïve Bayes"



- Verbinden Sie das TAN-Modell-Nugget mit dem Markov-Nugget (wählen Sie Ersetzen im Warnungsdialog).
- Verbinden Sie das Markov-Nugget mit dem Markov-FS-Nugget (wählen Sie Ersetzen im Warnungsdialog).

- Richten Sie für bessere Übersichtlichkeit die drei Nuggets mit dem Auswahlknoten aus.

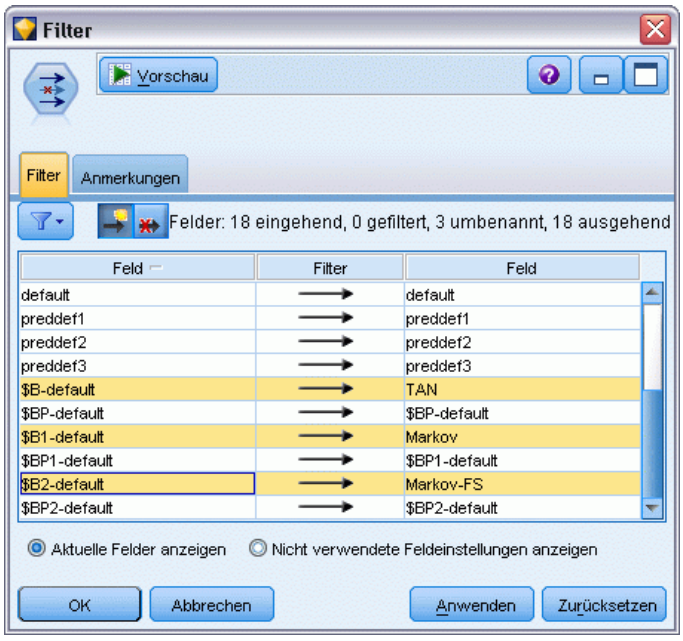
Abbildung 18-8
Ausrichten der Nuggets im Stream



- Um die Modellausgaben zugunsten größerer Klarheit im Evaluationsdiagramm (das Sie erstellen werden) umzubenennen, müssen Sie einen Filter-Knoten an das Markov-FS-Modell-Nugget angliedern.

- Benennen Sie in der rechten *Field*-Spalte "$B-default" in "TAN", "$B1-default" in "Markov" und "$B2-default" in Markov-FS um.

Abbildung 18-9
Umbenennen von Feldnamen für Modelle

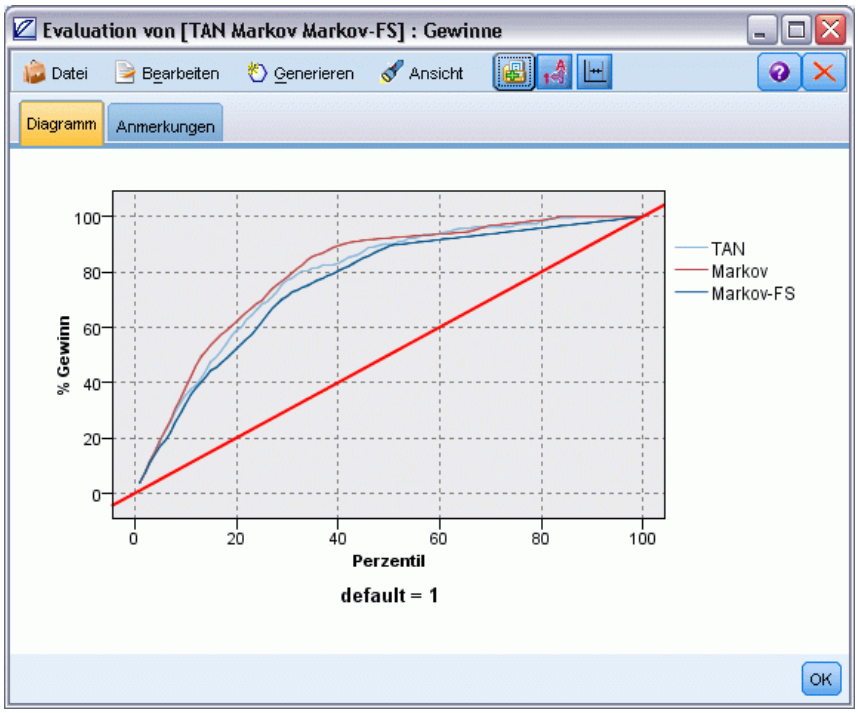


Um die Vorhersagegenauigkeit der Modelle zu vergleichen, können Sie ein Gewinndiagramm erstellen.

- Gliedern Sie einen Evaluationsdiagrammknoten an den Filterknoten an und führen Sie den Diagrammknoten mit den zugehörigen Standardeinstellungen aus.

Das Diagramm zeigt, dass die einzelnen Modelltypen zu ähnlichen Ergebnissen führen; das Markov-Modell ist jedoch geringfügig besser.

Abbildung 18-10
Evaluation der Modellgenauigkeit



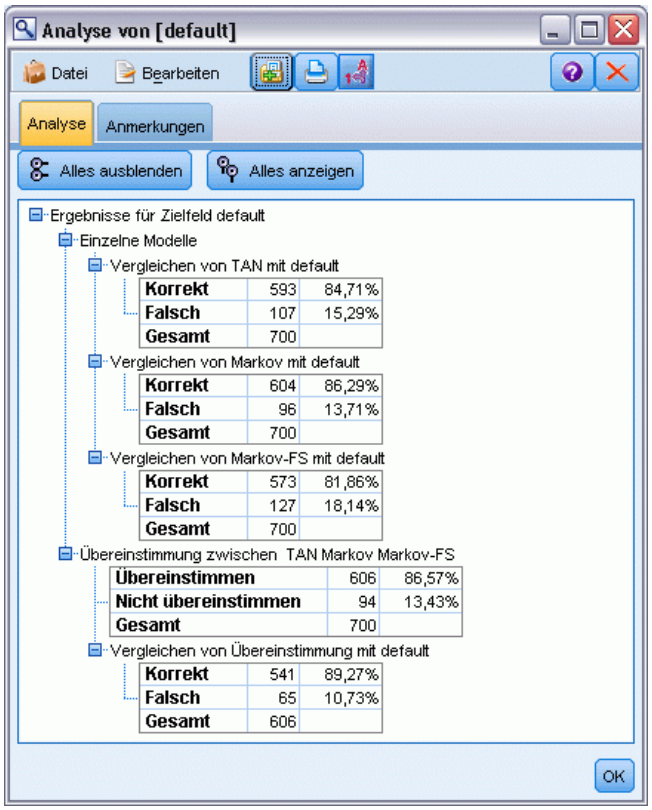
Um die Vorhersagequalität der einzelnen Modelle zu überprüfen, könnten Sie statt des Evaluationsdiagramms auch einen Analyseknoten verwenden. Dieser zeigt die Genauigkeit als Prozentsatz für richtige und falsche Vorhersagen.

- Gliedern Sie einen Analyseknoten an den Filterknoten an und führen Sie den Analyseknoten mit den zugehörigen Standardeinstellungen aus.

Wie beim Evaluationsdiagramm ergibt sich, dass das Markov-Modell eine geringfügig bessere Leistung bei der korrekten Vorhersage aufweist. Das Markov-FS-Modell liegt allerdings nur wenige Prozentpunkte hinter dem Markov-Modell. Dies kann bedeuten, dass es besser wäre, das Markov-FS-Modell zu verwenden, da es weniger Eingaben zur Berechnung der Ergebnisse

verwendet und somit einen geringeren Datenerfassungsaufwand und kürzere Eingabe- und Verarbeitungszeiten mit sich bringt.

Abbildung 18-11
Analysieren der Modellgenauigkeit



Erläuterungen der mathematischen Grundlagen für die in IBM® SPSS® Modeler verwendeten Modellierungsmethoden finden Sie im *SPSS Modeler-Algorithmushandbuch*, das sich im Verzeichnis *Documentation* des Installationsdatenträgers befindet.

Beachten Sie außerdem, dass diese Ergebnisse nur auf den Trainingsdaten beruhen. Um einzuschätzen, wie gut sich das Modell für andere Daten in der Praxis verallgemeinern lässt, könnten Sie mit einem Partitionsknoten eine Teilmenge der Datensätze für Test- und Validierungszwecke zurückhalten. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in *IBM SPSS Modeler 14.2- Quellen- , Prozess- und Ausgabeknoten*.

Kapitel
19

Erneutes Trainieren eines Modells auf monatlicher Basis (Bayes-Netzwerk)

Mithilfe des Knotens “Bayes-Netzwerk” können Sie ein Wahrscheinlichkeitsmodell erstellen, indem Sie beobachtete und aufgezeichnete Hinweise mit Weltwissen (“gesundem Menschenverstand”) kombinieren, um die Wahrscheinlichkeit des Vorkommens unter Verwendung scheinbar nicht miteinander verknüpfter Attribute zu ermitteln.

In diesem Beispiel wird ein Stream namens *bayes_churn_retrain.str* verwendet, der Bezug nimmt auf die Datendateien *telco_Jan.sav* und *telco_Feb.sav*. Diese Dateien finden Sie im Verzeichnis *Demos* jeder IBM® SPSS® Modeler-Installation. Sie können auch über die IBM® SPSS® Modeler-Programmgruppe im Windows-Startmenü aufgerufen werden. Die Datei *bayes_churn_retrain.str* befindet sich im Verzeichnis *streams*.

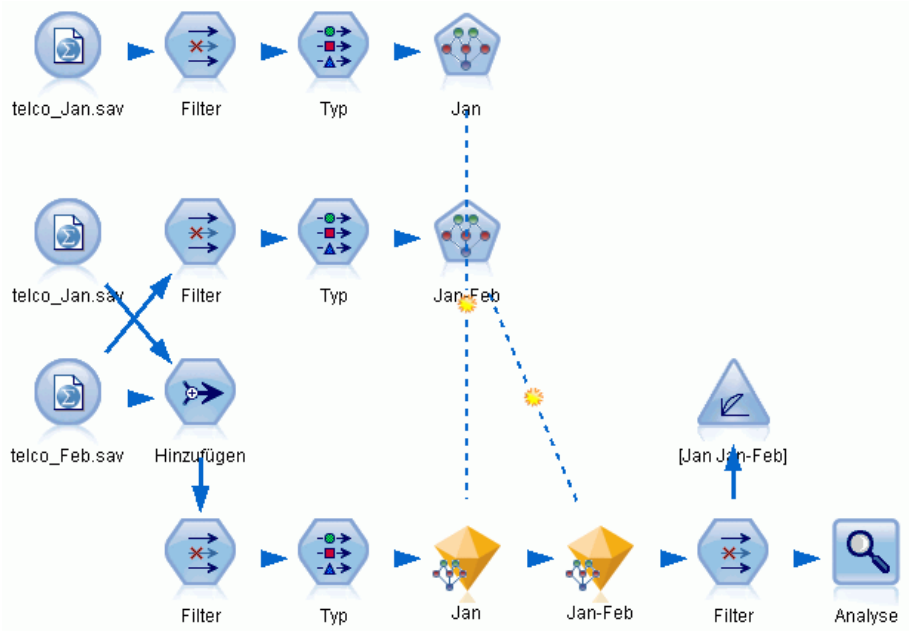
Hier ein Beispiel: Ein Telekommunikationsanbieter ist besorgt über die Anzahl an Kunden, die er an Mitbewerber verliert (Abwanderung). Wenn historische Kundendaten verwendet werden können, um vorherzusagen, welche Kunden in der Zukunft mit höherer Wahrscheinlichkeit abwandern, können gezielt Anreize oder andere Angebote für diese Kunden erstellt werden, um sie von Ihrem Wechsel zu einem anderen Anbieter abzubringen.

In diesem Beispiel wird anhand der Abwanderungsdaten für einen bestimmten Monat vorhergesagt, welche Kunden wahrscheinlich in Zukunft abwandern. Anschließend werden die Daten des Folgemonats ergänzt, um das Modell zu verfeinern und neu zu trainieren.

Erstellen des Streams

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *telco_Jan.sav* im Ordner *Demos* verweist.

Abbildung 19-1
Beispiel-Stream für Bayes-Netzwerk

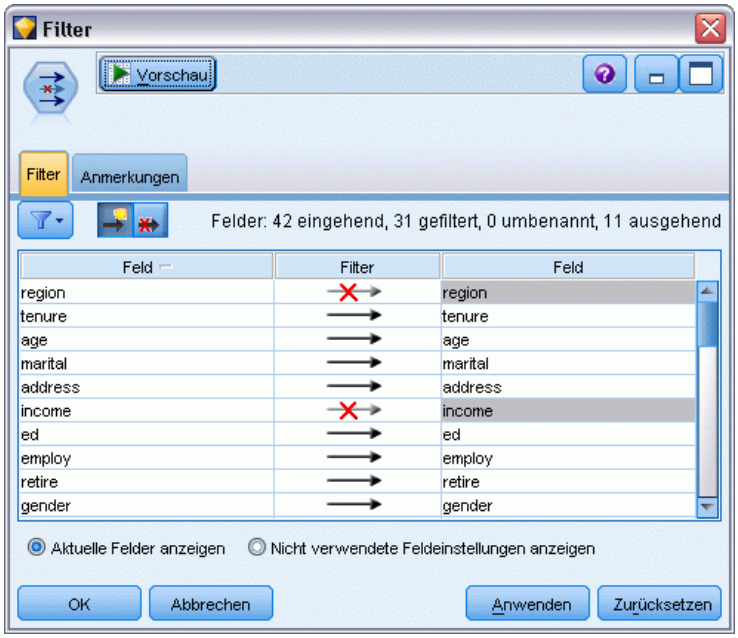


Vorangegangene Analysen haben gezeigt, dass mehrere Datenfelder bei der Vorhersage der Abwanderung kaum von Bedeutung sind. Diese Felder können aus dem Daten-Set herausgefiltert werden, um die Verarbeitungsgeschwindigkeit beim Erstellen und Scoren von Modellen zu erhöhen.

- Fügen Sie einen Filterknoten zum Quellenknoten hinzu.
- Schließen Sie alle Felder mit Ausnahme von *address*, *age*, *churn*, *custcat*, *ed*, *employ*, *gender*, *marital*, *reside*, *retire* und *tenure* aus.

▶ Klicken Sie auf OK.

Abbildung 19-2
Filtern unnötiger Felder



▶ Fügen Sie einen Typknoten zum Filterknoten hinzu.

▶ Öffnen Sie den Typknoten und klicken Sie auf die Schaltfläche Werte lesen, um die Spalte *Werte* auszufüllen.

- Damit der Evaluationsknoten einschätzen kann, welcher Wert wahr und welcher falsch ist, setzen Sie das Messniveau für das Feld *churn* auf Flag und setzen Sie die Rolle auf Ziel. Klicken Sie auf OK.

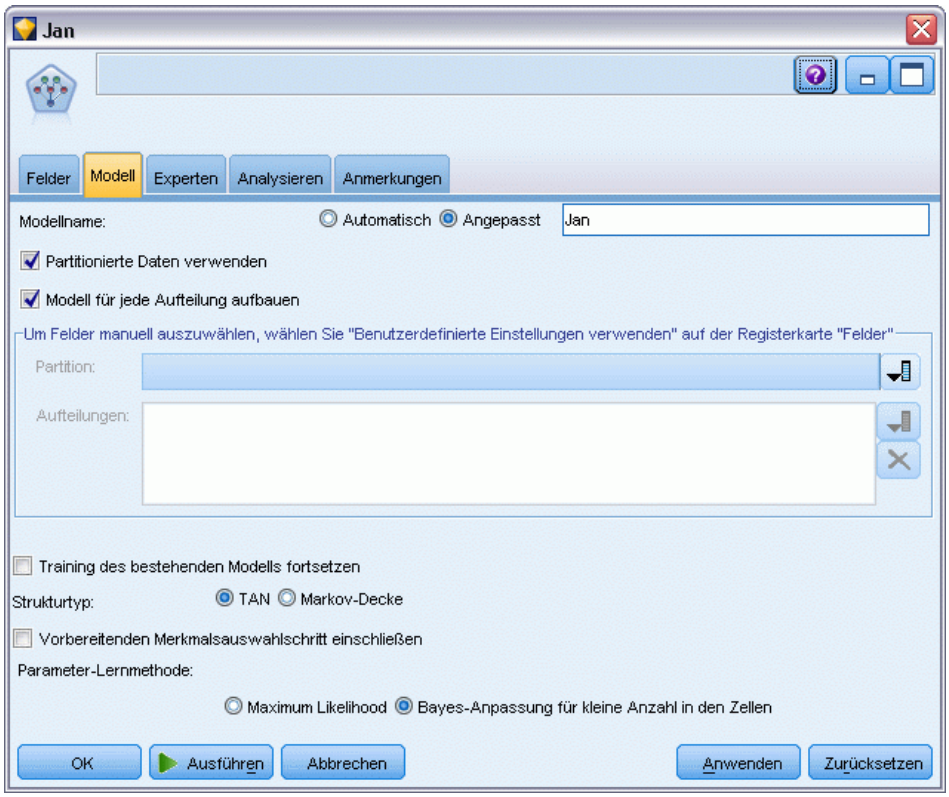
Abbildung 19-3
Auswahl des Zielfelds

Sie können mehrere verschiedene Typen von Bayes-Netzwerken erstellen. Für dieses Beispiel wird jedoch ein Modell vom Typ “Tree Augmented Naïve Bayes (TAN)” erstellt. Dadurch wird ein großes Netzwerk erstellt und gewährleistet, dass alle möglichen Verbindungen zwischen Datenvariablen aufgenommen wurden. Somit wurde ein robustes ursprüngliches Modell erstellt.

- Gliedern Sie einen Bayes-Netzwerk-Knoten an den Typknoten an.
- Wählen Sie auf der Registerkarte “Modell” als Modellnamen Benutzerdefiniert (Angepasst) und geben Sie im Textfeld den Ausdruck Jan ein.
- Wählen Sie als Parameter-Lernmethode Bayes-Anpassung für kleine Anzahl in den Zellen.

- Klicken Sie auf Ausführen. Das generierte Modell-Nugget wird zum Stream und zur Modellpalette in der rechten oberen Ecke hinzugefügt.

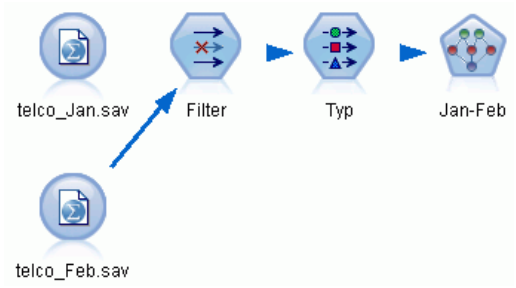
Abbildung 19-4
Erstellen eines Modells vom Typ "Tree Augmented Naïve Bayes"



- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *telco_Feb.sav* im Ordner *Demos* verweist.

- Hängen Sie diesen neuen Quellenknoten an den Filterknoten an (wählen Sie im Warnungsdialogfeld Ersetzen, um die Verbindung zum vorherigen Quellenknoten zu ersetzen).

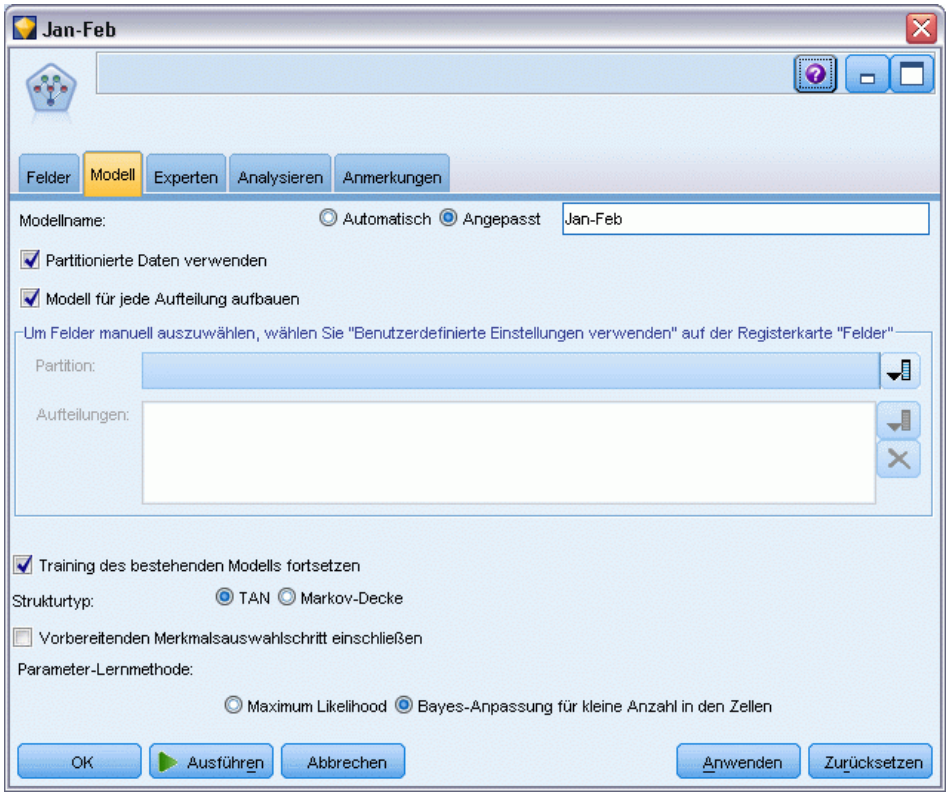
Abbildung 19-5
Hinzufügen der Daten des zweiten Monats



- Wählen Sie auf der Registerkarte "Modell" des Bayes-Netzwerk-Knotens als Modellnamen Benutzerdefiniert (Angepasst) aus und geben Sie im Textfeld den Ausdruck Jan-Feb ein.

- Wählen Sie die Option Training des bestehenden Modells fortsetzen.
- Klicken Sie auf Ausführen. Das generierte Modell-Nugget überschreibt den bestehenden im Stream, wird jedoch auch zur Modellpalette in der rechten oberen Ecke hinzugefügt.

Abbildung 19-6
Erneutes Trainieren des Modells



Bewertung des Modells

Um die Modelle vergleichen zu können, müssen Sie die beiden Daten-Sets kombinieren.

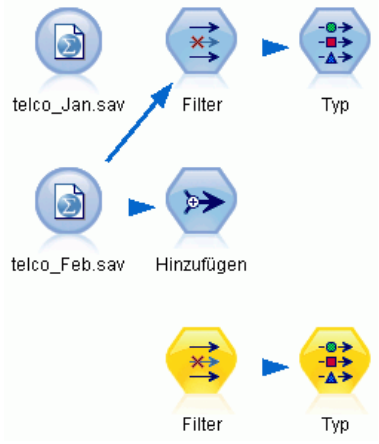
- Fügen Sie einen Anhangknoten hinzu und gliedern Sie die Quellenknoten *telco_Jan.sav* und *telco_Feb.sav* daran an.

Abbildung 19-7
Gliedern Sie die beiden Datenquellen an

- Kopieren Sie den Filter- und den Typknoten von weiter oben im Stream und fügen Sie sie in den Stream-Zeichenbereich ein.

- Gliedern Sie den Anhangknoten an den neu kopierten Filterknoten an.

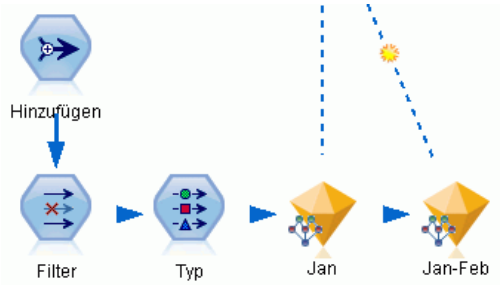
Abbildung 19-8
Einfügen der kopierten Knoten in den Stream



Die Nuggets für die beiden Bayes-Netzwerk-Modelle befinden sich in der Modell-Palette in der rechten oberen Ecke.

- Doppelklicken Sie auf das Jan-Modell-Nugget, um es in den Stream zu übernehmen und an den soeben kopierten Typknoten anzuhängen.
- Gliedern Sie das Modell-Nugget “Jan-Feb”, das sich bereits im Stream befindet, an das Modell-Nugget “Jan” an.
- Öffnen Sie das Jan-Modell-Nugget.

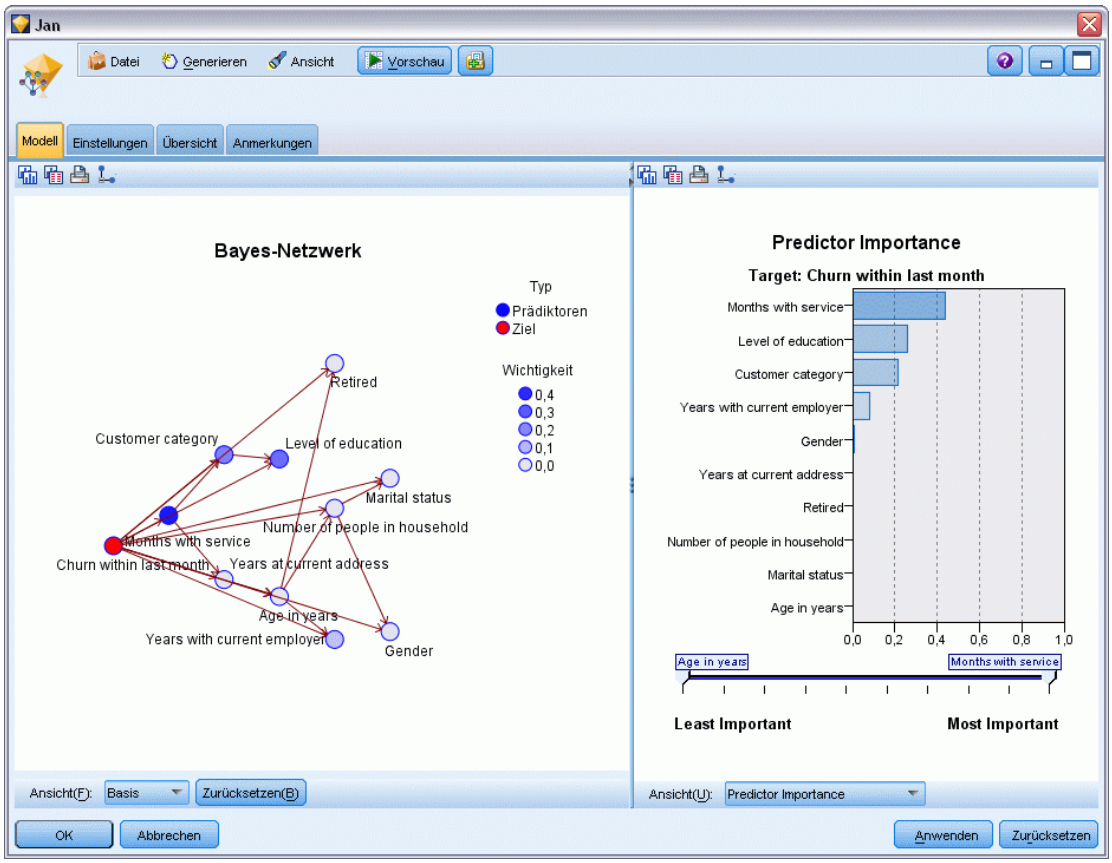
Abbildung 19-9
Hinzufügen der Nuggets zum Stream



Die Registerkarte “Modell” des Modell-Nuggets vom Typ “Bayes-Netzwerk” gliedert sich in zwei Spalten: Die linke Spalte enthält ein Netzwerkdiagramm mit Knoten, das die Beziehung zwischen dem Ziel und seinen wichtigsten Prädiktoren sowie die Beziehung zwischen den Prädiktoren anzeigt.

Die rechte Spalte zeigt entweder die *Bedeutsamkeit der Prädiktoren*, also die relative Wichtigkeit der einzelnen Prädiktoren bei der Schätzung des Modells, oder die *Konditionalen Wahrscheinlichkeiten*, also den Wert der bedingten Wahrscheinlichkeit für die einzelnen Knoten und jede Kombination von Werten in ihren übergeordneten Knoten.

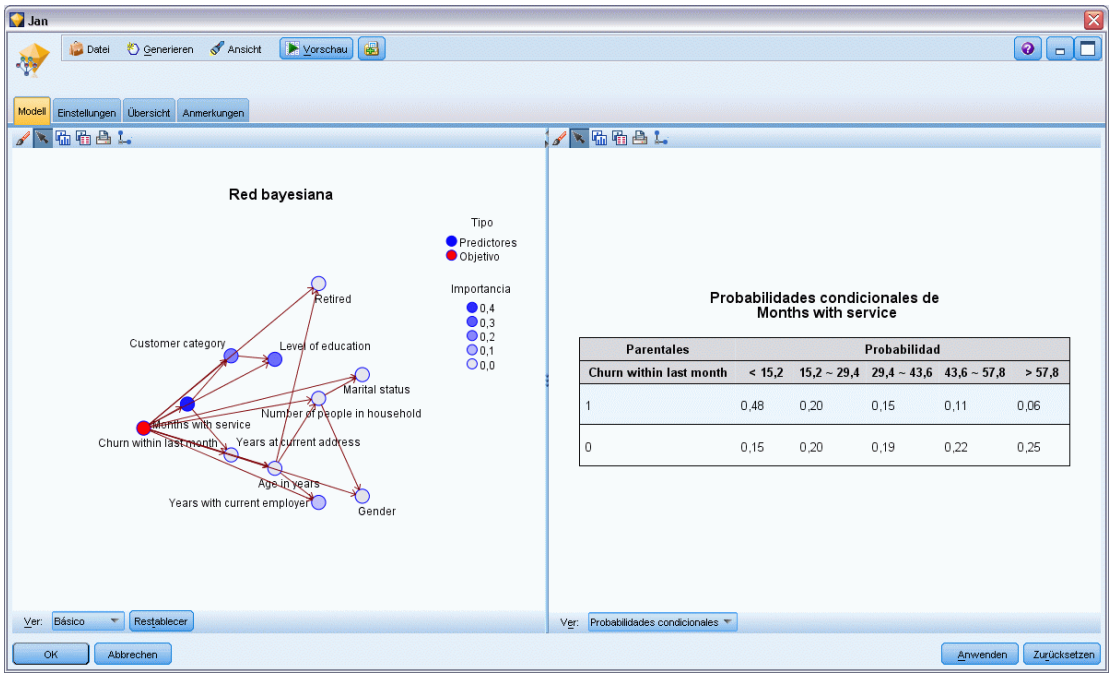
Abbildung 19-10
Bayes-Netzwerk-Modell mit Bedeutsamkeit der Prädiktoren



Zur Anzeige der bedingten Wahrscheinlichkeiten für einen Knoten klicken Sie auf den Knoten in der linken Spalte. Die rechte Spalte wird mit den erforderlichen Details aktualisiert.

Die bedingten Wahrscheinlichkeiten werden für jede Klasse angezeigt, in die die Datenwerte unterteilt wurden - relativ zum übergeordneten Knoten und den gleichrangigen Knoten.

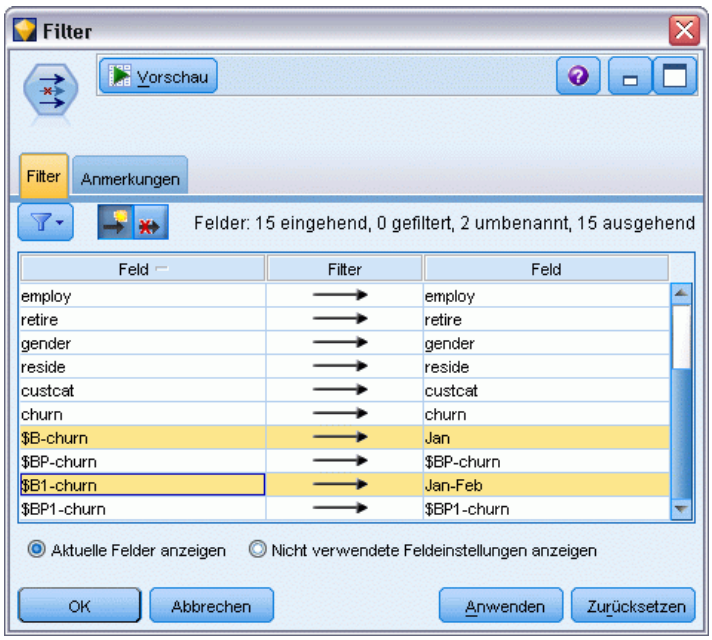
Abbildung 19-11
Bayes-Netzwerk-Modell mit bedingten Wahrscheinlichkeiten



- Um die Modellausgaben zugunsten größerer Klarheit umzubenennen, müssen Sie einen Filter-Knoten an das Modell-Nugget “Jan-Feb” angliedern.

▶ Benennen Sie in der rechten *Field*-Spalte "$B-churn" in "Jan" und "$B1-churn" in "Jan-Feb" um.

Abbildung 19-12
Umbenennen von Feldnamen für Modelle



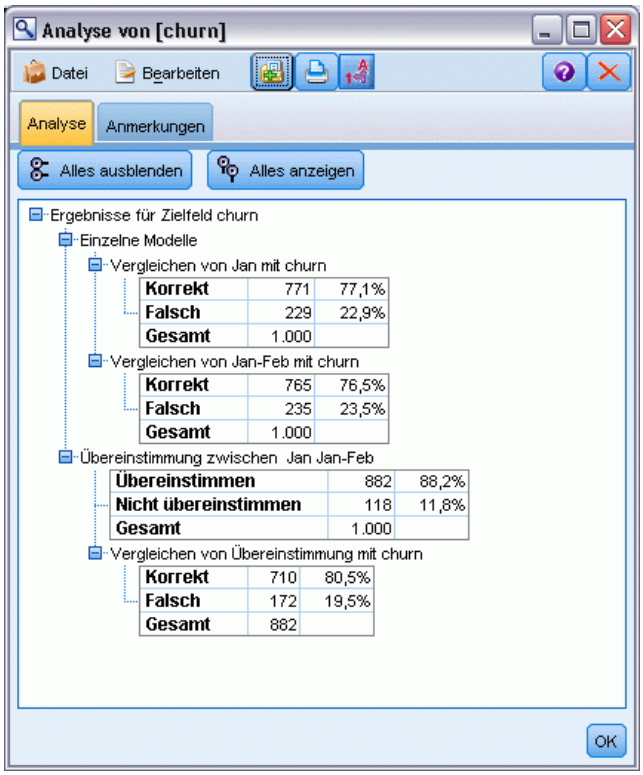
Mit einem Analyseknoten können Sie überprüfen, wie gut die einzelnen Modelle die Abwanderung vorhersagen. Dieser Knoten gibt die Genauigkeit als Prozentsätze für die korrekten und falschen Vorhersagen an.

▶ Gliedern Sie einen Analyseknoten an den Filterknoten an.

▶ Öffnen Sie den Analyseknoten und klicken Sie auf Ausführen.

Dies zeigt, dass beide Modelle ein ähnliches Maß an Genauigkeit für die Vorhersage der Abwanderung aufweisen.

Abbildung 19-13
Analysieren der Modellgenauigkeit



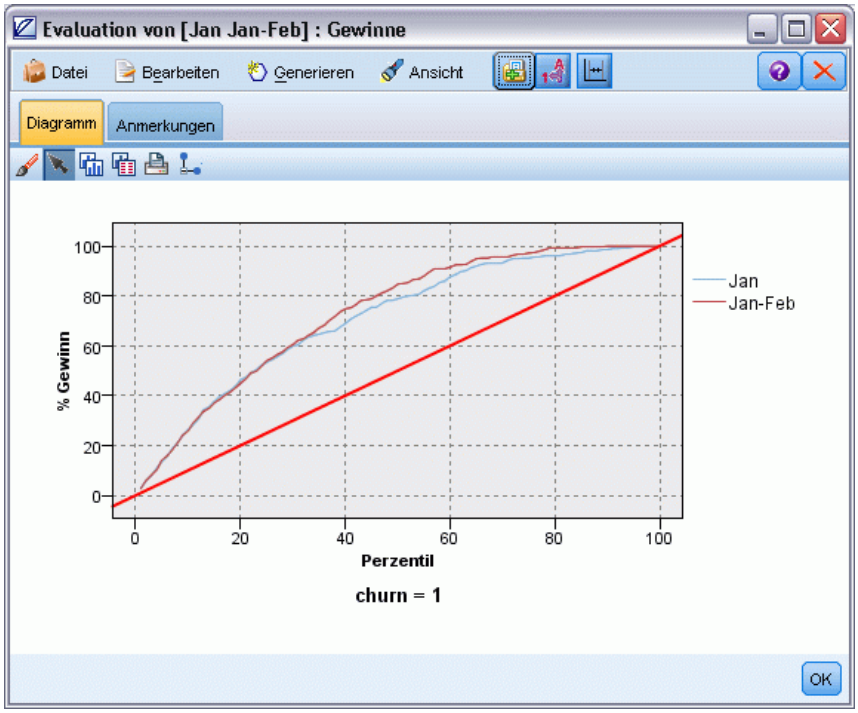
Als Alternative zum Analyseknoten können Sie mit einem Evaluationsdiagramm die Vorhersagegenauigkeit der Modelle durch Erstellung eines Gewinndiagramms vergleichen.

- Gliedern Sie einen Diagrammknoten vom Typ “Evaluation” an den Filterknoten an.

Führen Sie den Diagrammknoten mit all seinen Standardeinstellungen aus.

Wie der Analyseknoten zeigt auch das Diagramm, dass die einzelnen Modelltypen zu ähnlichen Ergebnissen führen. Das erneut trainierte Modell, bei dem die Daten aus beiden Monaten verwendet wurden, ist jedoch geringfügig besser, da seine Vorhersagen ein höheres Konfidenzniveau aufweisen.

Abbildung 19-14
Evaluation der Modellgenauigkeit



Erläuterungen der mathematischen Grundlagen für die in IBM® SPSS® Modeler verwendeten Modellierungsmethoden finden Sie im *SPSS Modeler-Algorithmushandbuch*, das sich im Verzeichnis *Documentation* des Installationsdatenträgers befindet.

Beachten Sie außerdem, dass diese Ergebnisse nur auf den Trainingsdaten beruhen. Um einzuschätzen, wie gut sich das Modell für andere Daten in der Praxis verallgemeinern lässt, könnten Sie mit einem Partitionsknoten eine Teilmenge der Datensätze für Test- und Validierungszwecke zurückhalten. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in *IBM SPSS Modeler 14.2- Quellen- , Prozess- und Ausgabeknoten*.

Kapitel
20

Werbeaktion für Einzelhandelumsatz (Netzwerk/C&RT)

Dieses Beispiel befasst sich mit Daten zu Produktlinien im Einzelhandel und den Auswirkungen von Werbeaktionen auf den Umsatz. (Die Daten sind frei erfunden.) Ziel dieses Beispiels ist es, die Auswirkungen zukünftiger Werbeaktionen vorherzusagen. Ähnlich wie beim Zustandsüberwachungsbeispiel besteht der Data-Mining-Vorgang aus Explorations-, Datenvorbereitungs-, Trainings- und Testphase.

In diesem Beispiel werden die Streams *goodsplot.str* und *goodslearn.str* verwendet, die auf die Datendateien *GOODS1n* und *GOODS2n* verweisen. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Der Stream *goodsplot.str* befindet sich im Ordner *streams*, während sich die Datei *goodslearn.str* im Verzeichnis *streams* befindet.

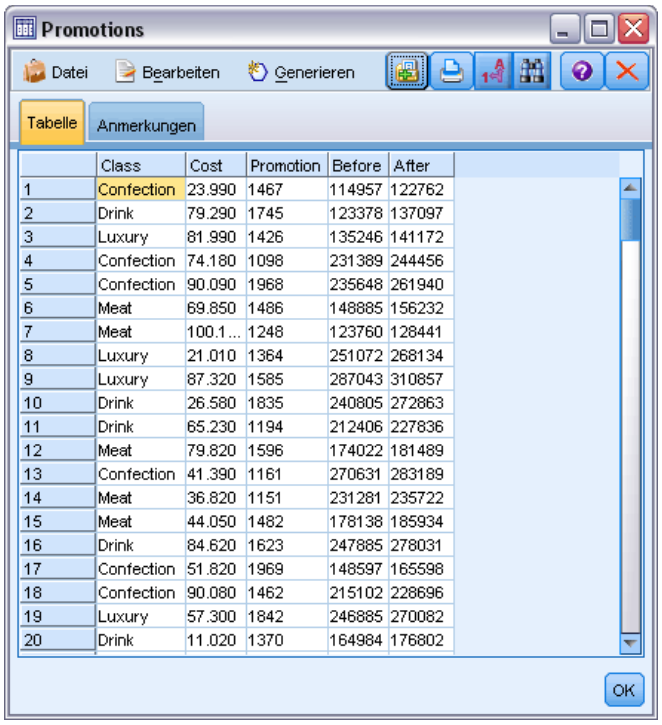
Untersuchen der Daten

Jeder Datensatz enthält Folgendes:

- *Klasse*. Produkttyp.
- *Kosten*. Preis einer Einheit.
- *Werbeaktion*. Index des Betrags, der für eine bestimmte Werbeaktion aufgebracht wird.
- *Vor*. Einkünfte vor der Werbeaktion.
- *Nach*. Einkünfte nach der Werbeaktion.

Der Stream *goodsplot.str* enthält einen einfachen Stream zum Anzeigen der Daten in einer Tabelle. Die beiden Felder für die Einkünfte (*Vor* und *Nach*) werden in absoluten Begriffen ausgedrückt; es ist jedoch wahrscheinlich, dass die Steigerung der Einkünfte nach der Werbeaktion (und wohl als Ergebnis davon) eine hilfreichere Abbildung darstellen würde.

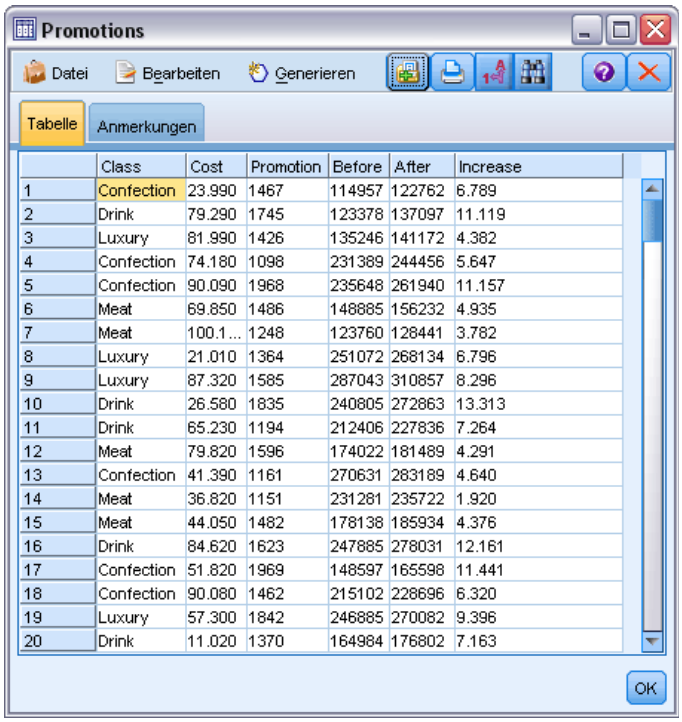
Abbildung 20-1
Auswirkungen von Werbeaktionen auf den Umsatz



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

Der Stream *goodsplot.str* enthält auch einen Knoten zur Ableitung dieses Werts, ausgedrückt als Prozentwert der Einkünfte vor der Werbeaktion, in einem Feld mit der Bezeichnung *Anstieg* und zeigt eine Tabelle mit diesem Feld an.

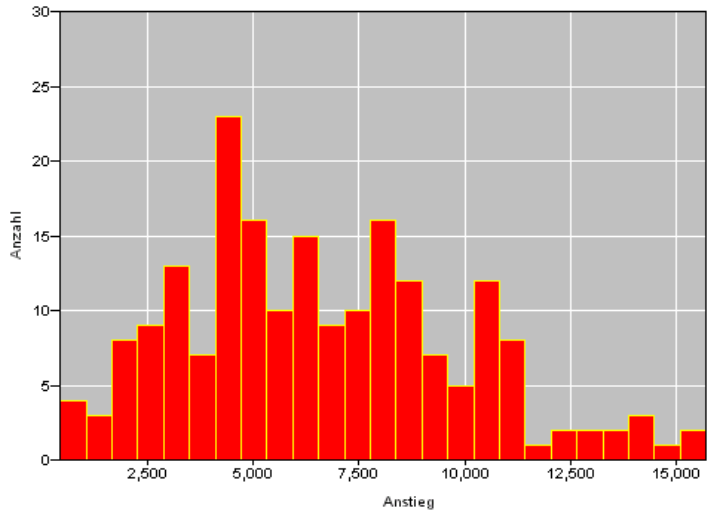
Abbildung 20-2
Anstieg der Einkünfte nach der Werbeaktion



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

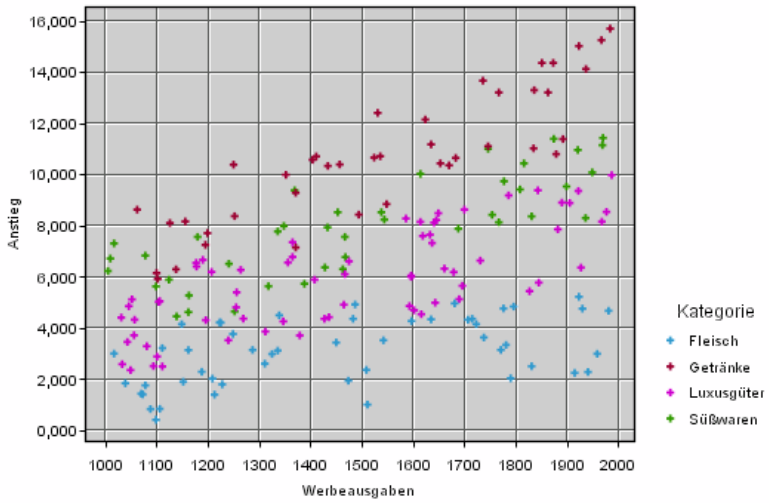
Außerdem zeigt der Stream ein Histogramm des Anstiegs sowie ein Streudiagramm des Anstiegs im Vergleich zu den Kosten für die Werbeaktion, überlagert von der betroffenen Produktkategorie.

Abbildung 20-3
Histogramm mit dem Anstieg der Einkünfte



Das Streudiagramm zeigt, dass für jede Produktklasse eine fast lineare Beziehung zwischen dem Anstieg an Einkünften und den Kosten für die Werbeaktion besteht. Deshalb ist es wahrscheinlich, dass ein Entscheidungsbaum oder ein neuronales Netz mit einer akzeptablen Genauigkeit den Anstieg der Einkünfte aus anderen verfügbaren Feldern vorhersagen könnte.

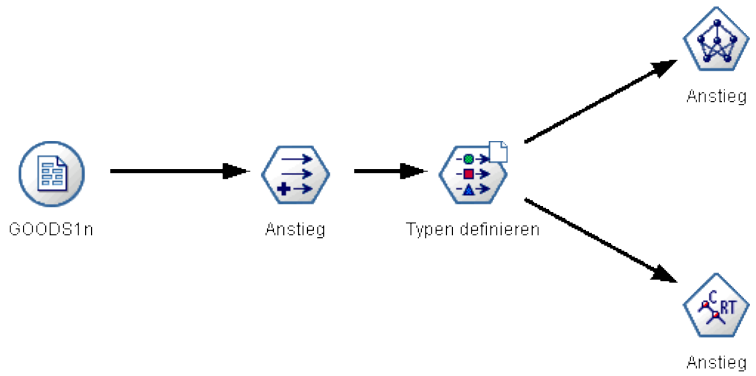
Abbildung 20-4
Anstieg der Einkünfte im Vergleich zu den Ausgaben für die Werbeaktion



Lernen und Testen

Der Stream *goodslearn.str* trainiert ein neuronales Netz und einen Entscheidungsbaum, diese Vorhersage bzgl. des Anstiegs an Einkünften zu treffen.

Abbildung 20-5
Modellierung des Streams "goodslearn.str"



Sobald Sie die Modellknoten ausgeführt und die tatsächlichen Modelle generiert haben, können Sie die Ergebnisse des Lernprozesses testen. Dazu verbinden Sie den Entscheidungsbaum und das Netz zwischen dem Typknoten und einem neuen Analyseknoten in Reihe, indem Sie die Eingabe-(Daten-)Datei zu *GOODS2n* ändern und den Analyseknoten ausführen. Anhand der

Ausgabe dieses Knotens, insbesondere aus der linearen Korrelation zwischen dem vorhergesagten Anstieg und der richtigen Antwort, werden Sie feststellen, dass die trainierten Systeme den Anstieg der Einkünfte mit einem hohen Erfolgsquotienten vorhersagen.

Eine weitere Exploration könnte sich auf Fälle konzentrieren, bei denen die trainierten Systeme relativ hohe Fehlerraten verursachen; diese könnten identifiziert werden, indem der prognostizierte Anstieg der Einkünfte mit dem tatsächlichen Anstieg verglichen wird. Ausreißer in diesem Diagramm könnten mithilfe der interaktiven Grafiken von IBM® SPSS® Modeler ausgewählt werden. Anhand ihrer Eigenschaften kann es möglich sein, die Datenbeschreibung bzw. den Lernprozess so zu optimieren, dass die Genauigkeit verbessert wird.

Kapitel
21

Bedingungsüberwachung (Netzwerk/C5.0)

Dieses Beispiel betrifft die Überwachungsstatusinformationen eines Rechners sowie das Problem der Erkennung und Vorhersage von Fehlerzuständen. Die Daten werden aus einer fiktiven Simulation erstellt und bestehen aus einer Reihe von verketteten Zeitreihen, die im Laufe der Zeit gemessen wurden. Jeder Datensatz ist ein "Schnappschuss"-Bericht auf dem Rechner in Bezug auf:

- *Zeit*. Eine ganze Zahl.
- *Energie*. Eine ganze Zahl.
- *Temperatur*. Eine ganze Zahl.
- *Druck*. 0 falls normal, 1 für eine momentane Druckwarnung.
- *Uptime*. Vergangene Zeit seit letzter Wartung.
- *Status*. Normal 0, wechselt zu Fehlercode bei Fehler (101, 202 oder 303).
- *Ergebnis*. Der Fehlercode, der in dieser Zeitreihe angezeigt wird, oder 0, wenn kein Fehler auftritt. (Diese Codes stehen nur rückwirkend zur Verfügung.)

In diesem Beispiel werden die Streams *condplot.str* und *condlearn.str* verwendet, die die Datendateien *COND1n* und *COND2n* referenzieren. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Dateien *condplot.str* und *condlearn.str* befinden sich im Verzeichnis *streams*.

Für jede Zeitreihe gibt es eine Reihe von Datensätzen aus einem Zeitraum des normalen Betriebs, gefolgt von einem Zeitraum, der zum Fehler führte, wie in der folgenden Tabelle dargestellt:

Time	Potenz	Temperatur	Druck	Uptime	Status	Ergebnis
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0

Time	Potenz	Temperatur	Druck	Uptime	Status	Ergebnis
52	965	251	0	209	0	0
53	938	251	0	209	0	101
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

Die folgende Vorgehensweise ist für die meisten Data-Mining-Projekte typisch:

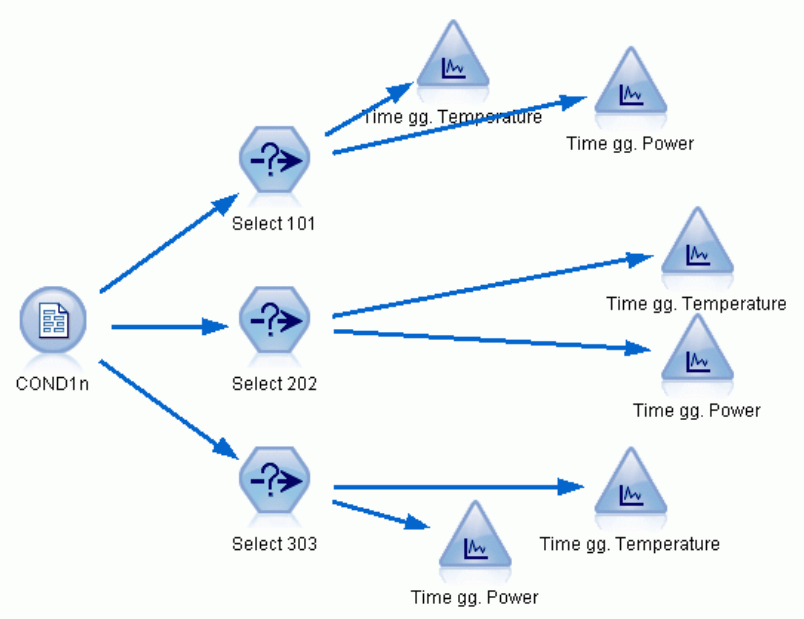
- Prüfen Sie die Daten, um zu ermitteln, welche Attribute für die Vorhersage oder Erkennung der gewünschten Zustände von Bedeutung sind.
- Behalten Sie diese Attribute (falls bereits vorhanden) bei oder leiten Sie diese ab und fügen Sie sie ggf. den Daten hinzu.
- Verwenden Sie die resultierenden Daten für das Training von Regeln und neuronalen Netzen.
- Testen Sie die trainierten Systeme unter Verwendung von unabhängigen Testdaten.

Untersuchen der Daten

Die Datei *condplot.str* stellt den ersten Teil des Prozesses dar. Sie enthält einen Stream, der eine Vielzahl von Diagrammen plottet. Wenn die Zeitreihe von Temperatur oder Energie sichtbare Muster enthält, können Sie zwischen bevorstehenden Fehlerbedingungen unterscheiden und möglicherweise ihr Auftreten vorhersagen. Für Temperatur und Energie stellt der nachfolgende Stream die mit den drei verschiedenen Fehlercodes verknüpften Zeitreihen in voneinander

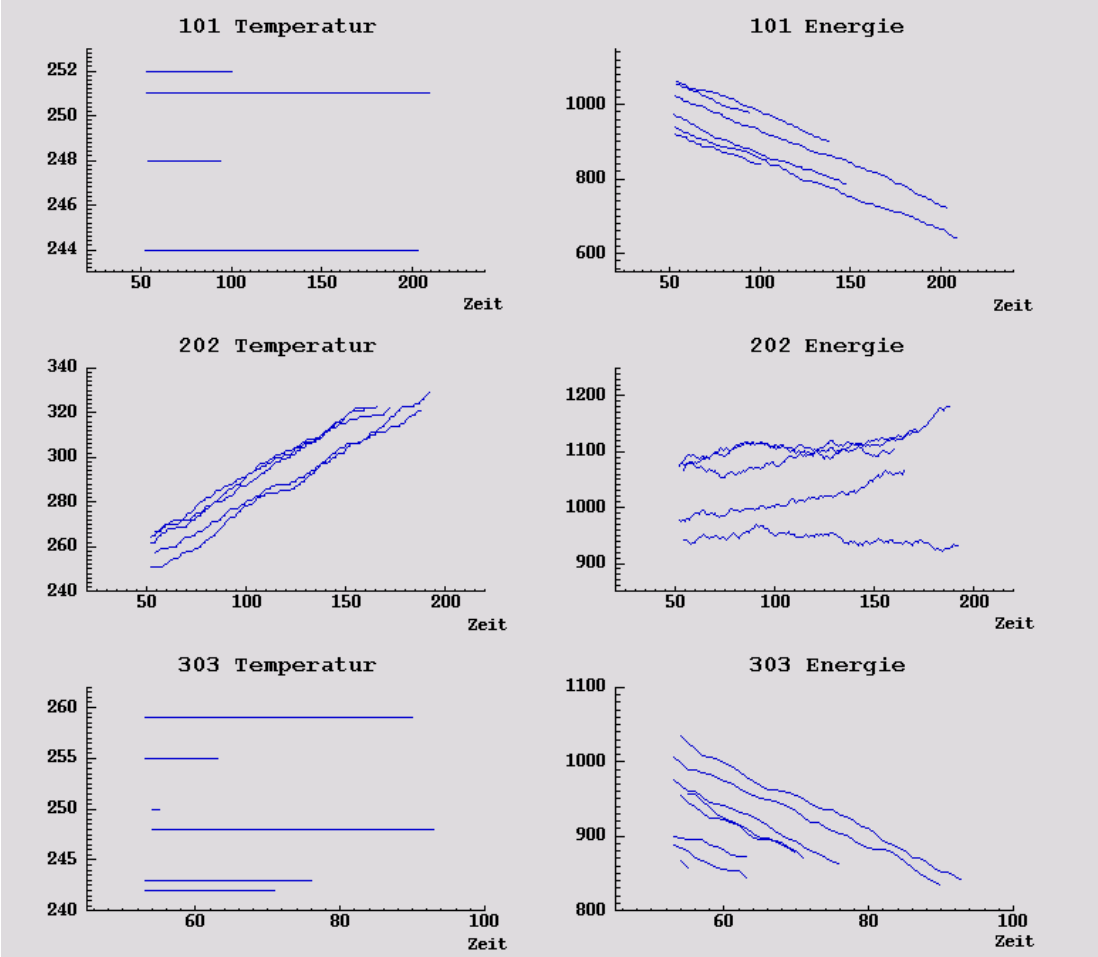
getrennten Diagrammen dar, was zu insgesamt sechs Diagrammen führt. Auswahlknoten trennen die mit den verschiedenen Fehlercodes verknüpften Daten voneinander.

Abbildung 21-1
Condplot-Stream



Die Ergebnisse dieses Streams werden in dieser Abbildung dargestellt.

Abbildung 21-2
Temperatur und Energie im Laufe der Zeit



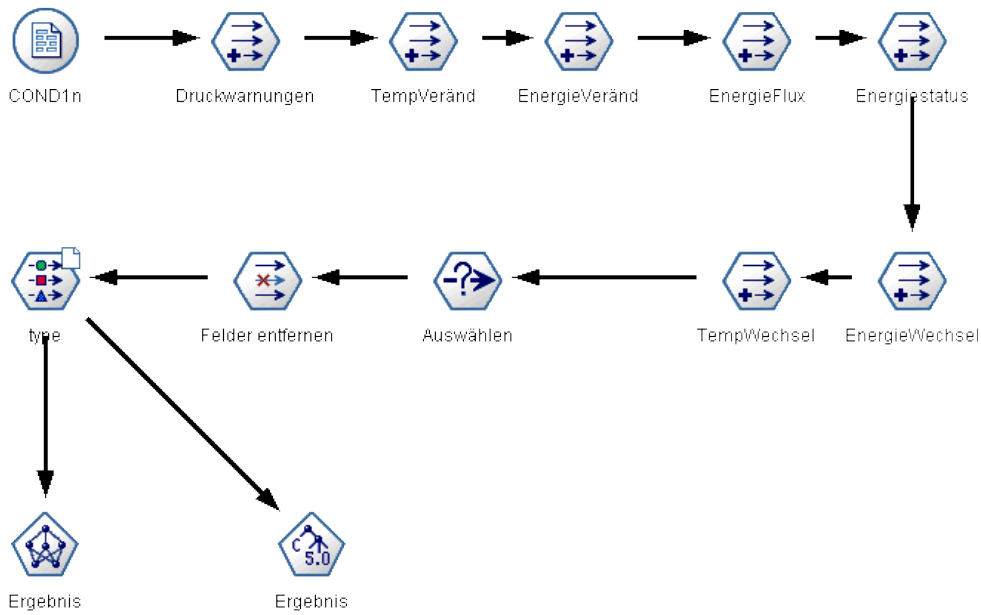
Die Diagramme zeigen deutlich Muster, die 202-Fehler von 101- und 303-Fehlern unterscheiden. Die 202-Fehler zeigen steigende Temperatur und fluktuierende Energie im Laufe der Zeit; die anderen Fehler jedoch nicht. Muster, die 101- von 303-Fehlern unterscheiden, sind weniger klar. Beide Fehler zeigen eine gleichmäßige Temperatur und ein Absinken der Energie, das Absinken der Energie ist jedoch offensichtlich für 303-Fehler steiler.

Basierend auf diesen Diagrammen ist es offensichtlich, dass das Vorhandensein und die Geschwindigkeit einer Änderung von Temperatur und Energie sowie das Vorhandensein und der Grad der Fluktuation für die Vorhersage und Unterscheidung von Fehlern relevant sind. Deshalb sollten diese Attribute den Daten hinzugefügt werden, bevor die Lernsysteme angewendet werden.

Data Preparation (Vorbereitung von Daten)

Basierend auf den Ergebnissen einer Datenexploration leitet der Stream *condlearn.str* die relevanten Daten ab und lernt, wie Fehler vorhergesagt werden.

Abbildung 21-3
Condlearn-Stream



Der Stream verwendet eine Reihe von Ableitungsknoten, um die Daten für die Modellierung vorzubereiten.

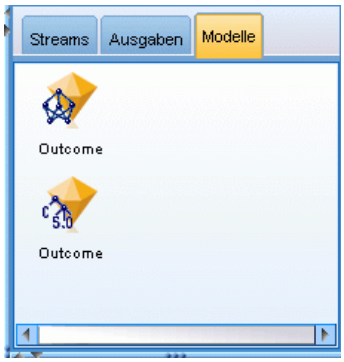
- **Knoten "Variable Datei"**. Liest die Datendatei *COND1n*.
- **Ableiten Druckwarnungen**. Zählt die Anzahl der momentan vorhandenen Druckwarnungen. Wird zurückgesetzt, wenn Zeit zu 0 zurückkehrt.
- **Ableiten TempInc**. Berechnet die momentane Geschwindigkeit einer Temperaturveränderung mithilfe von @DIFF1.
- **Ableiten PowerInc**. Berechnet die momentane Geschwindigkeit einer Energieveränderung mithilfe von @DIFF1.
- **Ableiten EnergieFlux**. Ein Flag, wahr, wenn die Energie in entgegengesetzte Richtungen im letzten Datensatz und in diesem abwich; das heißt, in Richtung einer Energiespitze oder eines Energietals.
- **Ableiten Energiestatus**. Ein Zustand, der bei *Stabil* beginnt und zu *Fluktuierend* wechselt, wenn zwei aufeinander folgende Energieflüsse entdeckt werden. Wechselt nur dann zurück zu *Stabil*, wenn es über fünf Zeitintervalle hinweg keinen Energiefluss gibt oder wenn *Zeit* zurückgesetzt wird.
- **EnergieVeränd**. Durchschnitt von *PowerInc* im Verlauf der letzten fünf Zeitintervalle.
- **TempVeränd**. Durchschnitt von *TempInc* im Verlauf der letzten fünf Zeitintervalle.

- **Ursprungswerte verwerfen (Auswahl)**. Verwirft den ersten Datensatz aller Zeitreihen, um große (inkorrekte) Sprünge in *Energie* und *Temperatur* an den Grenzen zu vermeiden.
- **Felder verwerfen**. Reduziert Datensätze zu *Uptime*, *Status*, *Ergebnis*, *Druckwarnungen*, *Energiestatus*, *EnergieVeränd* und *TempVeränd*.
- **Typ**. Definiert die Rolle von *Ergebnis* als Ziel (das vorherzusagende Feld). Definiert außerdem das Messniveau von *Ergebnis* als Nominal, *Druckwarnungen* als Stetig und *Energiestatus* als Flag.

Lernen

Die Ausführung des Streams in *condlearn.str* trainiert die C5.0-Regel und das neuronale Netzwerk (Netz). Das Training des Netzes kann einige Zeit in Anspruch nehmen, es kann aber früh unterbrochen werden, um ein Netz beizubehalten, das akzeptable Resultate liefert. Sobald das Lernen abgeschlossen ist, blinkt die Registerkarte "Modelle" oben rechts in den Manager-Fenstern, um Sie zu informieren, dass zwei neue Nuggets erstellt wurden: Einer stellt das neuronale Netz und einer die Regel dar.

Abbildung 21-4
Model Manager mit Modell-Nuggets



Die Modell-Nuggets werden ebenfalls dem vorhandenen Stream hinzugefügt, so dass Sie das System testen oder die Modellergebnisse exportieren können. In diesem Beispiel testen wir die Ergebnisse des Modells.

Testen

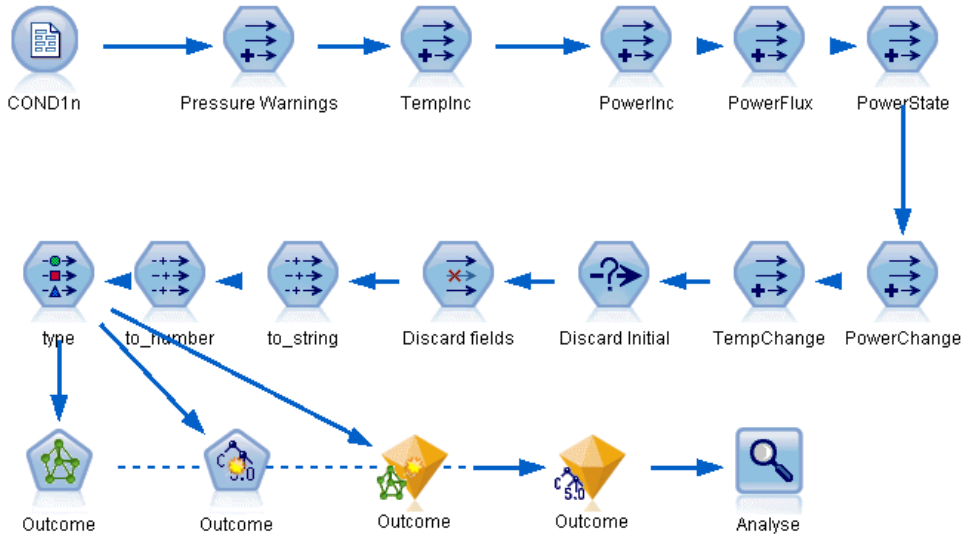
Die Modell-Nuggets werden in den Stream eingefügt und sind beide mit dem Typknoten verbunden.

▶ Positionieren Sie die Nuggets wie abgebildet, so dass der Typknoten an den Netzwerk-Nugget anschließt, der mit dem C5.0-Nugget verbunden ist.

▶ Gliedern Sie einen Analyseknoten an das C5.0-Nugget an.

- Bearbeiten Sie den ursprünglichen Quellknoten, um die Datei *COND2n* (anstelle von *COND1n*) zu lesen, da *COND2n* ungesehene Testdaten enthält.

Abbildung 21-5
Testen des trainierten Netzes



- Öffnen Sie den Analyseknoten und klicken Sie auf Ausführen.

  Dies führt zu Zahlen, die die Genauigkeit des trainierten Netzes und der trainierten Regel widerspiegeln.

Kapitel
22

Klassifizieren von Kunden im Telekommunikationsbereich (Diskriminanzanalyse)

Die Diskriminanzanalyse 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 Zielfelds ein kategoriales verwendet wird.

Nehmen wir beispielsweise an, dass ein Telekommunikationsanbieter seinen Kundenstamm nach Servicenutzungsmustern in vier Gruppen unterteilt hat. Wenn demografische Daten zum Vorhersagen der Gruppenzugehörigkeit verwendet werden können, sind angepasste Angebote für die einzelnen potenziellen Kunden möglich.

In diesem Beispiel wird ein Stream namens *telco_custcat_discriminant.str* verwendet, der Bezug nimmt auf die Datendatei *telco.sav*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *telco_custcat_discriminant.str* befindet sich im Verzeichnis *streams*.

Dieses Beispiel konzentriert sich auf die Verwendung von demografischen Daten zur Vorhersage von Nutzungsmustern. Das Zielfeld *custcat* weist vier mögliche Werte auf, die den vier Kundengruppen entsprechen:

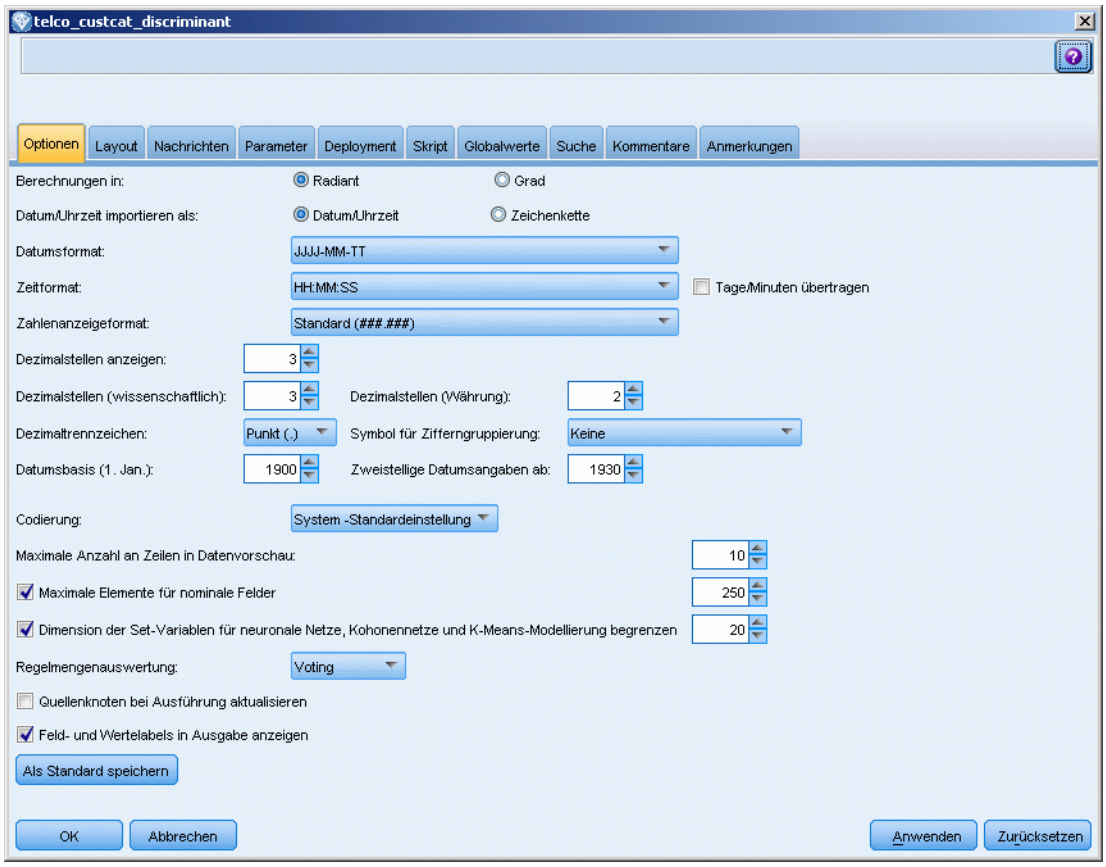
Wert	Label
1	Basic Service (Basis-Service)
2	E-Service
3	Plus Service (Plus-Service)
4	Total Service (Umfassender Service)

Erstellen des Streams

- Legen Sie zunächst die Stream-Eigenschaften fest, um die Feld- und Wertelabels in der Ausgabe anzuzeigen. Wählen Sie die folgenden Befehle aus den Menüs aus:
  Datei > Stream-Eigenschaften...

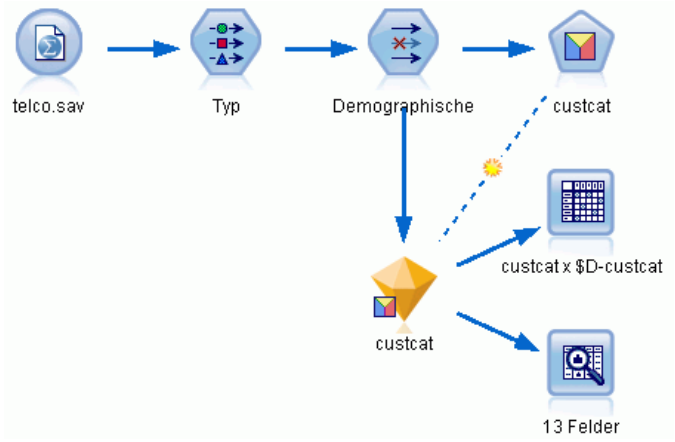
▶ Wählen Sie unbedingt die Option Feld- und Wertelabels in Ausgabe anzeigen und klicken Sie auf OK.

Abbildung 22-1
Stream-Eigenschaften



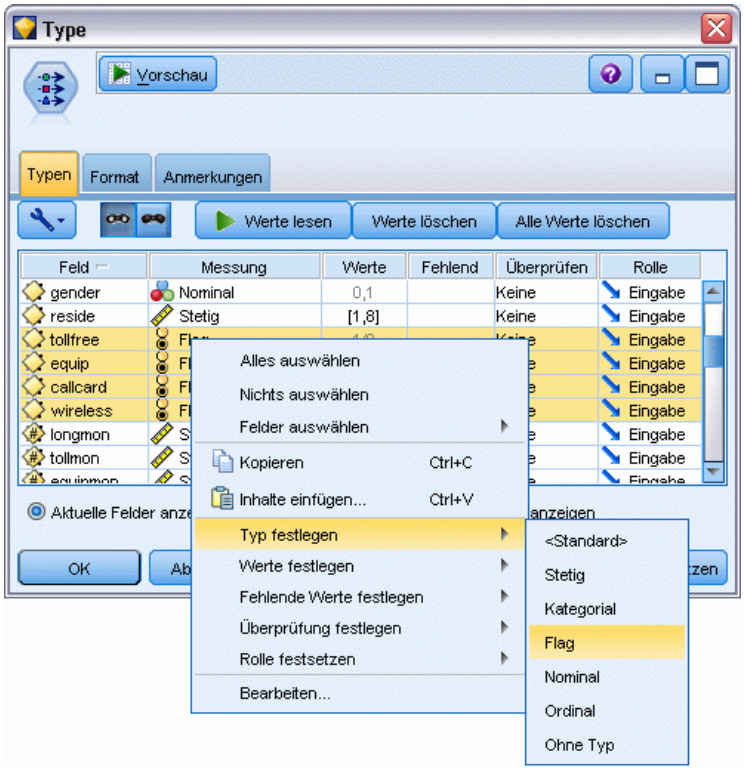
▶ Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

Abbildung 22-2
Beispiel-Stream zur Klassifizierung von Kunden mithilfe der Diskriminanzanalyse



- Fügen Sie einen Typknoten hinzu und klicken Sie auf Werte lesen. Achten Sie dabei darauf, dass alle Messniveaus korrekt festgelegt werden. Beispielsweise können die meisten Felder mit den Werten 1 und 0 als Flags betrachtet werden.

Abbildung 22-3
Festlegen des Messniveaus für mehrere Felder



Tipp: Um die Eigenschaften für mehrere Felder mit ähnlichen Werten (z. B. 0/1) zu ändern, klicken Sie auf die Überschrift der Spalte *Werte* (um die Felder nach ihrem Wert zu sortieren) und halten Sie anschließend die Umschalttaste gedrückt, während Sie mit der Maus oder den Pfeiltasten alle Felder auswählen, die geändert werden sollen. Anschließend können Sie mit der rechten Maustaste auf die Auswahl klicken, um das Messniveau oder andere Attribute der ausgewählten Felder zu ändern.

Beachten Sie, dass *Geschlecht* treffender als Feld mit einem Set von zwei Werten betrachtet wird denn als Flag. Belassen Sie also seinen Wert für “Messniveau” bei Nominal.

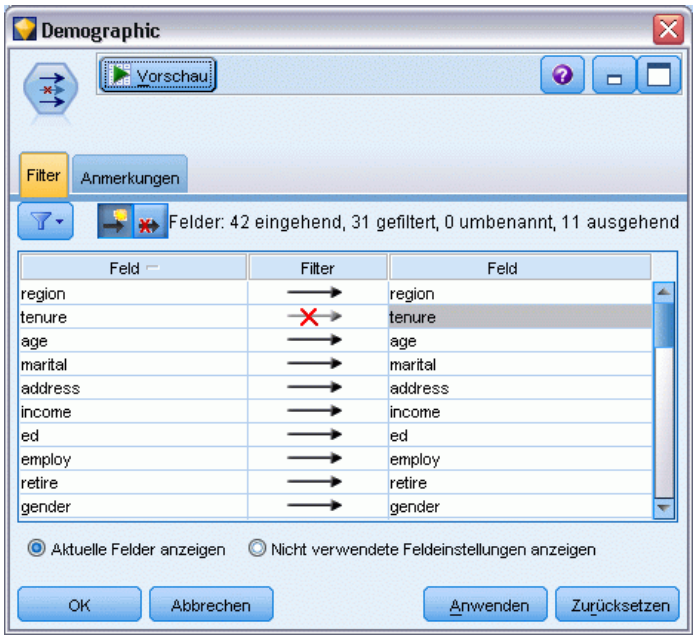
- Ändern Sie die Rolle für das Feld *custcat* in Ziel. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

Abbildung 22-4
Festlegen der Feldrolle

Da sich dieses Beispiel auf demografische Daten konzentriert, sollten Sie einen Filterknoten verwenden, mit dem nur die relevanten Felder (*region* (Region), *age* (Alter) *marital* (Familienstand), *address* (Adresse), *income* (Einkommen), *ed* (Bildung), *employ* (Beschäftigung),

retire (Ruhestand), *gender* (Geschlecht), *reside* (Wohnsitz) und *custcat* (Benutzerdef. Kategorie)) eingeschlossen werden. Die anderen Felder können für diese Analyse ausgeschlossen werden.

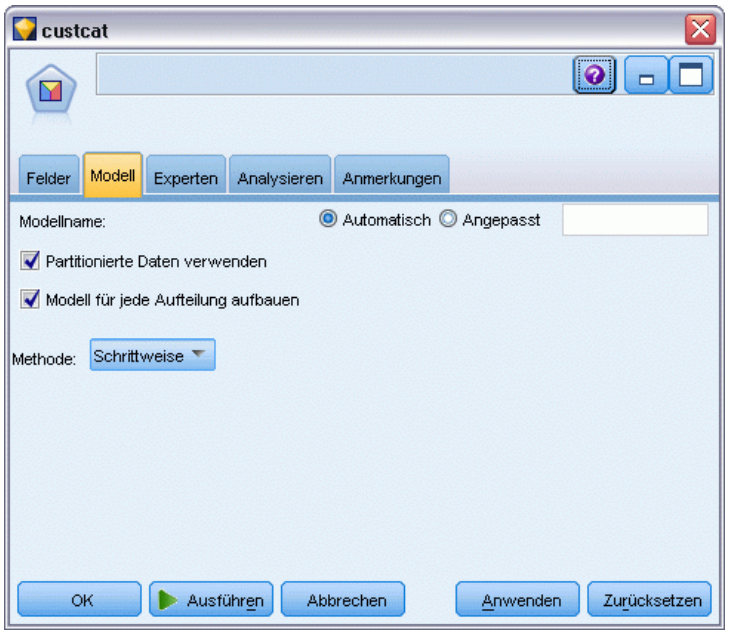
Abbildung 22-5
Filtern nach demografischen Feldern



(Alternativ können Sie die Rolle für diese Felder in Keine ändern, anstatt sie auszuschließen, oder die gewünschten Felder im Modellierungsknoten auswählen.)

- Klicken Sie im Diskriminanzknoten auf die Registerkarte “Modell” und wählen Sie die Methode Schrittweise aus.

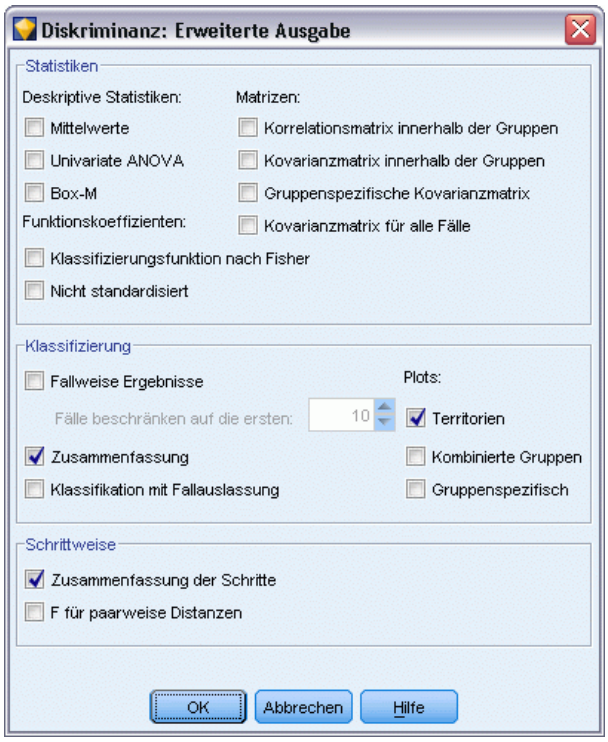
Abbildung 22-6
Auswählen Modelloptionen



- Setzen Sie auf der Registerkarte “Experten” den Modus auf Experten und klicken Sie auf Ausgabe.

- Wählen Sie auf der Registerkarte “Erweiterte Ausgabe” die Optionen Zusammenfassung, Territorien und Zusammenfassung der Schritte aus und klicken Sie auf OK.

Abbildung 22-7
Auswahl der Ausgabeoptionen

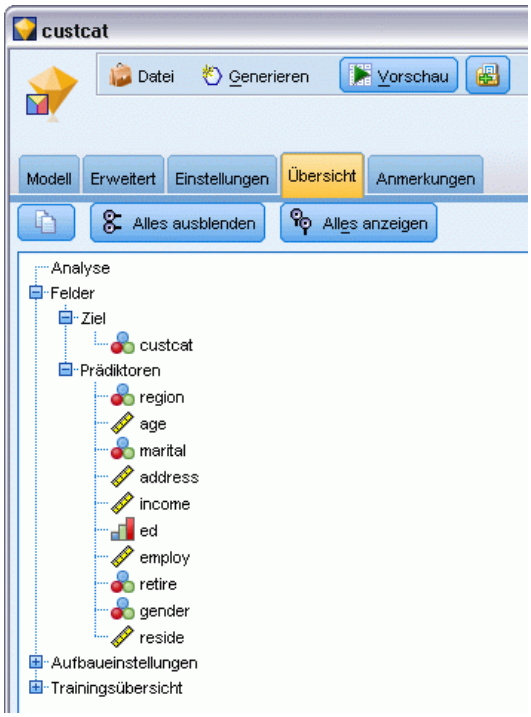


Untersuchen des Modells

- Klicken Sie auf Ausführen, um das Modell zu erstellen; dieses wird dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die zugehörigen Details anzuzeigen, doppelklicken Sie auf das Modell-Nugget im Stream.

Auf der Registerkarte “Übersicht” werden (unter anderem) die Ziele und die vollständige Liste der Eingaben (Prädiktorfelder) angezeigt, die zur Erwägung vorgelegt wurden.

Abbildung 22-8
Modellübersicht mit Ziel- und Eingabefeldern



So erhalten Sie Einzelheiten zu den Ergebnissen der Diskriminanzanalyse:

- Klicken Sie auf die Registerkarte “Erweitert”.
- Klicken Sie auf die Schaltfläche “In externem Browser starten” (direkt unter der Registerkarte “Modelle”), um die Ergebnisse in Ihrem Webbrowser anzuzeigen.

Schrittweise Diskriminanzanalyse

Abbildung 22-9
Nicht in die Analyse eingeschlossene Variablen, Schritt 0

Schritt		Toleranz	Minimale Toleranz	F-Wert für die Aufnahme	Wilks-Lambda
0	Age in years	1,000	1,000	7,521	,978
	Marital status	1,000	1,000	3,500	,990
	Years at current address	1,000	1,000	8,433	,975
	Household income in thousands	1,000	1,000	6,689	,980
	Level of education	1,000	1,000	61,454	,844
	Years with current employer	1,000	1,000	16,976	,951
	Retired	1,000	1,000	3,005	,991
	Gender	1,000	1,000	,373	,999
	Number of people in household	1,000	1,000	3,976	,988

Wenn eine große Anzahl von Einflussvariablen (Prädiktoren) vorhanden ist, kann die schrittweise Methode hilfreich sein, um automatisch die “besten” Variablen für das Modell auszuwählen. Die schrittweise Methode beginnt mit einem Modell, das keine der Einflussvariablen enthält. Bei jedem Schritt wird die Einflussvariable (Prädiktor) mit dem größten *F-Wert für Aufnahme*, der die Eintragskriterien überschreitet (standardmäßig 3,84), dem Modell hinzugefügt.

Abbildung 22-10
Nicht in die Analyse eingeschlossene Variablen, Schritt 3

Schritt		Toleranz	Minimale Toleranz	F-Wert für die Aufnahme	Wilks-Lambda
3	Age in years	,535	,535	,252	,795
	Marital status	,605	,593	1,507	,792
	Years at current address	,776	,771	3,514	,787
	Household income in thousands	,688	,657	,687	,794
	Retired	,917	,880	,353	,795
	Gender	,997	,931	,395	,795

Die Variablen, die beim letzten Analyseschritt übergangen wurden, weisen für *F-Wert für Aufnahme* jeweils einen Wert kleiner als 3,84 auf, sodass keine weiteren Variablen hinzugefügt werden.

Abbildung 22-11
Variablen in der Analyse

Schritt		Toleranz	F-Wert für den Ausschluß	Wilks-Lambda
1	Level of education	1,000	61,454	
2	Level of education	,953	59,108	,951
	Years with current employer	,953	14,933	,844
3	Level of education	,951	60,046	,940
	Years with current employer	,934	15,824	,834
	Number of people in household	,979	4,841	,807

In dieser Tabelle werden Statistiken für die Variablen angezeigt, die in den einzelnen Schritten in die Analyse aufgenommen werden. *Toleranz* ist der Anteil an der Varianz einer Variablen, der nicht durch andere unabhängige Variablen in der Gleichung erklärt wird. Eine Variable mit sehr geringer Toleranz trägt wenig zum Informationsgehalt eines Modells bei und kann zu Problemen bei der Berechnung führen.

Werte vom Typ *F-Wert für Ausschluss* sind nützlich, um zu beschreiben, was geschieht, wenn eine Variable aus dem aktuellen Modell entfernt wird (falls die anderen Variablen im Modell verbleiben). Der *F-Wert für Ausschluss* für die Aufnahmevariable entspricht dem *F-Wert für Aufnahme* im vorherigen Schritt (dargestellt in der Tabelle für nicht in die Analyse eingeschlossene Variablen).

Hinweis zu Problemen bei schrittweisen Methoden

Schrittweise Methoden sind praktisch, weisen jedoch Einschränkungen auf. Beachten Sie, dass bei schrittweisen Methoden die Modelle ausschließlich aufgrund des statistischen Vorteils ausgewählt werden. Die ausgewählten Einflussvariablen (Prädiktoren) haben daher möglicherweise keine **praktische Bedeutung**. Wenn Sie Erfahrungen mit den Daten haben und in etwa wissen, welche Einflussvariablen wichtig sind, sollten Sie diese Kenntnisse nutzen und keine schrittweisen Methoden verwenden. Wenn hingegen viele Einflussvariablen vorhanden sind und Ihnen kein geeigneter Ansatzpunkt bekannt ist, kann das Durchführen einer schrittweisen Analyse und das Anpassen des ausgewählten Modells zu einer besseren Vorhersage führen als gar kein Modell.

Überprüfen der Anpassungsgüte

Abbildung 22-12
Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,198(a)	80,2	80,2	,407
2	,048(a)	19,4	99,6	,214
3	,001(a)	,4	100,0	,031

Nahezu die gesamte Varianz, die durch das Modell erklärt wird, basiert auf den ersten beiden Diskriminanzfunktionen. Drei Funktionen werden automatisch angepasst. Aufgrund ihres sehr geringen Eigenwerts können Sie die dritte Funktion jedoch problemlos ignorieren.

Abbildung 22-13
Wilks-Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1 bis 3	,796	227,345	9	,000
2 bis 3	,953	47,486	4	,000
3	,999	,929	1	,335

Wilks-Lambda besagt ebenfalls, dass nur die ersten beiden Funktionen nützlich sind. Für die jeweiligen Funktionen wird damit die Hypothese getestet, dass die Mittelwerte der aufgelisteten Funktionen über mehrere Gruppen hinweg gleich sind. Der Test für Funktion 3 weist einen Signifikanzwert größer als 0,10 auf, sodass diese Funktion nur wenig zum Modell beiträgt.

Strukturmatrix

Abbildung 22-14
Strukturmatrix

	Funktion		
	1	2	3
Level of education	,966(*)	-,090	-,244
Years with current employer	-,182	,964(*)	-,193
Age in years(a)	-,162	,598(*)	-,285
Household income in thousands(a)	,109	,514(*)	-,190
Years at current address(a)	-,151	,394(*)	-,214
Retired(a)	-,108	,230(*)	-,137
Gender(a)	,008	,054(*)	,009
Number of people in household	,232	,097	,968(*)
Marital status(a)	,132	,134	,600(*)

Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

*. Größte absolute Korrelation zwischen jeder Variablen und einer Diskriminanzfunktion

a. Diese Variable wird in der Analyse nicht verwendet.

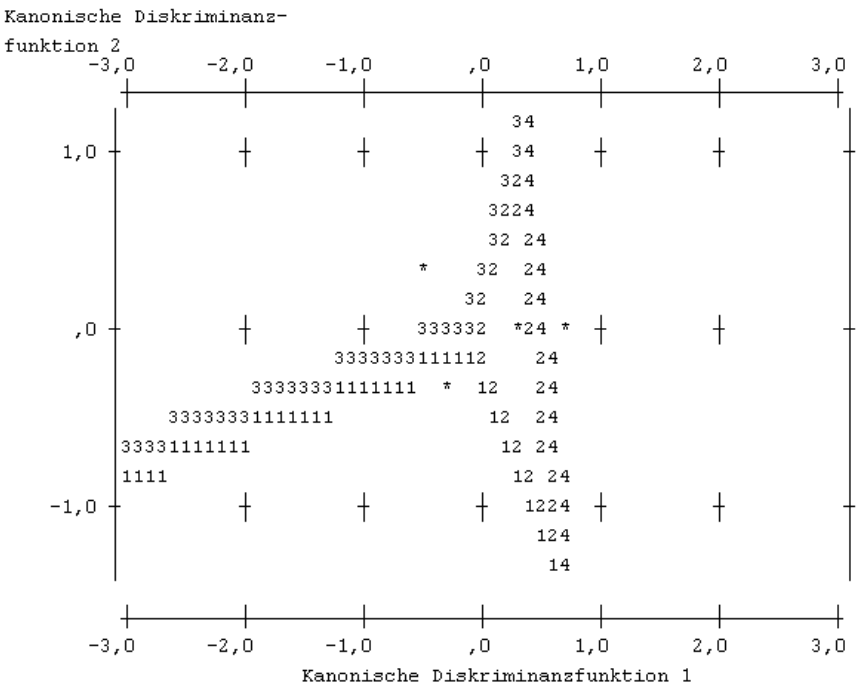
Wenn mehr als eine Diskriminanzfunktion vorhanden ist, markiert ein Sternchen (*) die größte absolute Korrelation der jeweiligen Variablen mit einer der kanonischen Funktionen. Innerhalb der jeweiligen Funktion werden diese markierten Variablen dann nach der Größe der Korrelation sortiert.

- *Level of education* (Bildungsniveau) korreliert am stärksten mit der ersten Funktion und es ist die einzige Variable, die am stärksten mit dieser Funktion korreliert.

- *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber), *Age in years* (Alter in Jahren), *Household income in thousands* (Haushaltseinkommen in Tausend), *Years at current address* (Wohnhaft an gleicher Adresse (in Jahren)) , *Retired* (Ruhestand) und *Gender* (Geschlecht) korrelieren am stärksten mit der zweiten Funktion, auch wenn *Gender* (Geschlecht) und *Retired* (Ruhestand) weniger stark korrelieren als die anderen Elemente. Die anderen Variablen markieren diese Funktion als "Stabilitätsfunktion".
- *Number of people in household* (Anzahl der Personen im Haushalt) und *Marital status* (Familienstand) korrelieren am stärksten mit der dritten Diskriminanzfunktion, doch da dies eine unnütze Funktion ist, sind diese Prädiktoren ebenso ohne Nutzen.

Territorien

Abbildung 22-15
Territorien



Territorien helfen Ihnen dabei, die Beziehungen zwischen den Gruppen und den Diskriminanzfunktionen zu untersuchen. Kombiniert mit den Ergebnissen der Strukturmatrix erhalten Sie eine grafische Interpretation der Beziehung zwischen Einflussvariablen (Prädiktoren) und Gruppen. Die erste Funktion, die auf der horizontalen Achse angezeigt wird, trennt Gruppe 4 (*Total Service*-Kunden) von den anderen. Da *Level of education* (Bildungsniveau) stark positiv mit der ersten Funktion korreliert, legt dies nahe, dass die *Total Service*-Kunden im Allgemeinen ein besonders hohes Bildungsniveau aufweisen. Die zweite Funktion trennt die Gruppen 1 und 3 (*Basic Service*- und *Plus Service*-Kunden). *Plus Service*-Kunden arbeiten üblicherweise bereits länger und sind älter als *Basic Service*-Kunden. *E-Service*-Kunden heben sich nicht besonders von den anderen ab, auch wenn die Territorien zeigen, dass sie üblicherweise über gute Bildung und eine gewisse Arbeitserfahrung verfügen.

Im Allgemeinen legt die Nähe der mit einem Sternchen (*) markierten Gruppenzentroiden zu den Territorienlinien nahe, dass die Trennung zwischen allen Gruppen nicht besonders stark ausgeprägt ist.

Nur die ersten beiden Diskriminanzfunktionen sind geplottet, doch da die dritte Funktion sich als eher unwichtig herausgestellt hat, bieten die Territorien eine umfassende Ansicht des Diskriminanzmodells.

Klassifikationsergebnisse

Abbildung 22-16
Klassifikationsergebnisse

			Vorhergesagte Gruppenzugehörigkeit				Gesamt
		Customer category	Basic service	E-service	Plus service	Total service	
Original	Anzahl	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47,0	4,1	22,9	25,9	100,0
		E-service	22,6	6,9	26,7	43,8	100,0
		Plus service	36,3	5,0	39,9	18,9	100,0
		Total service	16,9	6,8	15,7	60,6	100,0
a. 39,5% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.							

Durch Wilks-Lambda wissen Sie, dass Ihr Modell besser ist als bloßes Schätzen, doch Sie müssen die Klassifizierungsergebnisse untersuchen, um zu ermitteln, um wie viel besser. Anhand der beobachteten Daten würde das “Null”-Modell (d. h. das Modell ohne Einflussvariablen (Prädiktoren)) alle Kunden in die Gruppe, die dem Modalwert enstrpicht, einordnen, also *Plus Service*. Das Nullmodell wäre daher in 281/1000 = 28,1 % der Fälle richtig. Ihr Modell erfasst 39,5 % der Kunden richtig, dies entspricht einem Zuwachs von 11,4 %. Insbesondere ist das Modell beim Ermitteln der Kunden in *Total service* überlegen. Bei der Klassifikation der Kunden in *E-service* liegt dagegen ein außerordentlich schlechter Wert vor. Möglicherweise müssen Sie zum Trennen dieser Kunden eine andere Einflussvariable finden.

Zusammenfassung

Sie haben ein Diskriminanzmodell erstellt, das Kunden in eine von vier “Serviceverwendungsgruppen” einteilt, basierend auf den demografischen Daten der Kunden. Mithilfe der Strukturmatrix und der Territorien haben Sie herausgefunden, welche Variablen für die Segmentierung Ihres Kundenstammes besonders hilfreich sind. Zuletzt zeigen die Klassifizierungsergebnisse, dass das Modell nicht für die Klassifizierung von *E-Service*-Kunden geeignet ist. Es sind weitere Untersuchungen erforderlich, um eine andere Prädiktorvariable

festzulegen, die diese Kunden besser klassifiziert. Doch je nachdem, welche Elemente Sie vorhersagen möchten, kann das Modell für Ihre Bedürfnisse bestens geeignet sein. Wenn Sie beispielsweise die Identifizierung von *E-Service*-Kunden nicht benötigen, kann das Modell genau genug für Sie sein. Dies kann der Fall sein, wenn E-Service ein Lockangebot ist, das nicht viel Profit generiert. Wenn Sie beispielsweise den Großteil Ihrer Kapitalrendite durch *Plus-Service*- oder *Gesamtservice*-Kunden erwirtschaften, erhalten Sie durch dieses Modell alle notwendigen Informationen.

Beachten Sie außerdem, dass diese Ergebnisse nur auf den Trainingsdaten beruhen. Um einzuschätzen, wie gut sich das Modell für andere Daten verallgemeinern lässt, könnten Sie mit einem Partitionsknoten eine Teilmenge der Datensätze für Test- und Validierungszwecke zurückhalten. Für weitere Informationen siehe Thema Partitionsknoten in Kapitel 4 in *IBM SPSS Modeler 14.2- Quellen- , Prozess- und Ausgabeknoten*.

Erläuterungen der mathematischen Grundlagen der in IBM® SPSS® Modeler verwendeten Modellierungsverfahren sind im SPSS Modeler-Algorithmushandbuch aufgeführt. Es steht im Verzeichnis *Documentation* des Installationsdatenträgers zur Verfügung.

Kapitel
23

Analysieren von intervallzensierten Überlebensdaten (Verallgemeinerte lineare Modelle)

Wenn die Analyse von Überlebensdaten mit Intervallzensierung vorgenommen wird – d. h. wenn die exakte Zeit des betreffenden Ereignisses nicht bekannt ist, aber in einem bestimmten Zeitraum angesiedelt werden kann –, führt die intervallmäßige Anwendung des Cox-Modells auf Ereigns-Hazards zu einem komplementären Log-Log-Regressions-Modell.

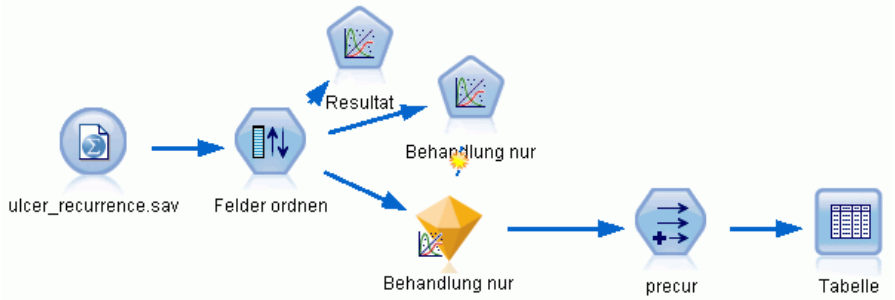
Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren sind in *ulcer_recurrence.sav* erfasst. Dieses Daten-Set wurde an anderer Stelle vorgestellt und analysiert. Mit verallgemeinerten linearen Modellen können Sie die Ergebnisse der komplementären Log-Log-Regressions-Modelle reproduzieren.

In diesem Beispiel wird der Stream *ulcer_genlin.str* verwendet, der sich auf die Datendatei *ulcer_recurrence.sav* bezieht. Die Datendatei befindet sich im Ordner *Demos* und die Stream-Datei im Unterordner *streams*. Für weitere Informationen siehe Thema Ordner "Demos" in Kapitel 1 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Erstellen des Modells

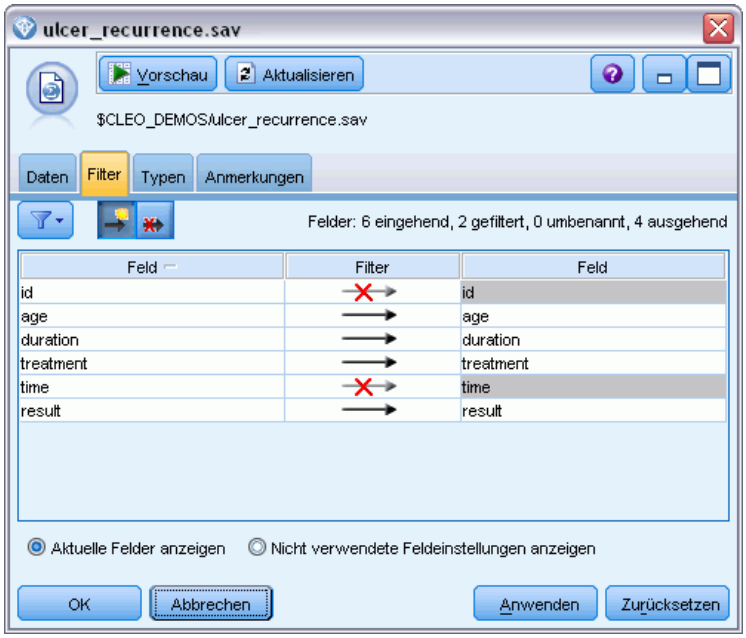
- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *ulcer_recurrence.sav* im Ordner *Demos* verweist.

Abbildung 23-1
Beispiel-Stream zur Vorhersage des erneuten Auftretens von Geschwüren



- Filtern Sie auf der Registerkarte “Filter” des Quellenknotens *id* und *time* heraus.

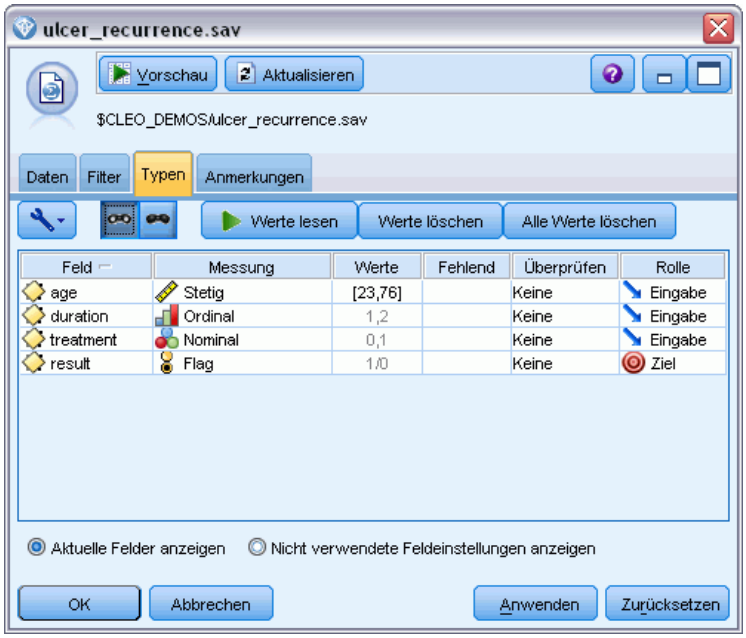
Abbildung 23-2
Filtern von unerwünschten Feldern



- Setzen Sie auf der Registerkarte “Typen” des Quellenknotens die Rolle für das Feld *result* auf Ziel und setzen Sie das Messniveau auf Flag. Ein Ergebnis von 1 gibt an, dass das Geschwür erneut aufgetreten ist. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

▶ Klicken Sie auf Werte lesen, um die Daten zu instanziieren.

Abbildung 23-3
Festlegen der Feldrolle



- Fügen Sie einen Knoten vom Typ “Felder ordnen” hinzu und legen Sie *duration*, *treatment* und *age* als Eingabereihenfolge fest. Dies legt die Reihenfolge fest, in der die Felder in das Modell eingegeben werden. So können Sie Colletts Ergebnisse leichter reproduzieren.

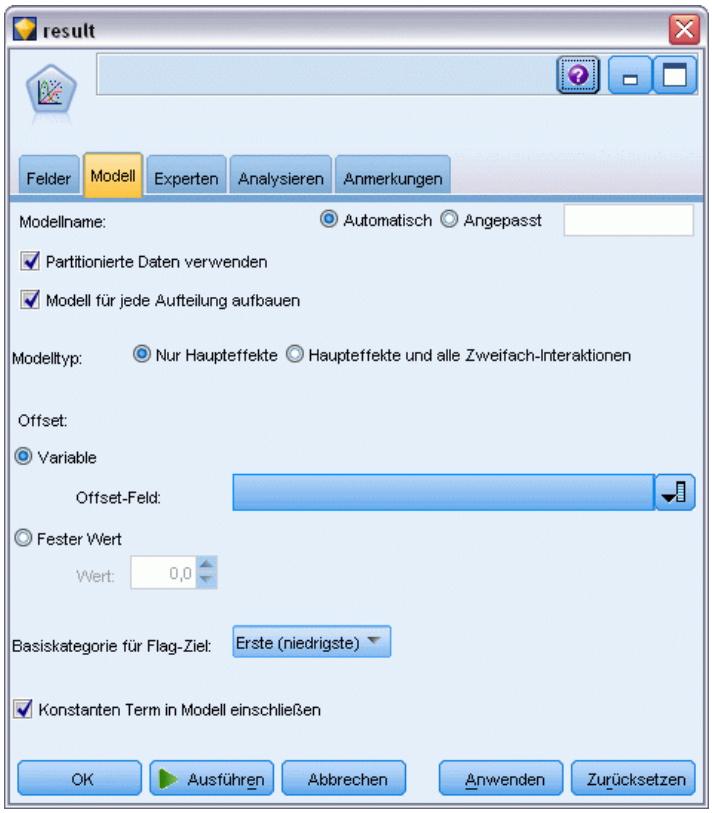
Abbildung 23-4
Neuordnung von Feldern für die gewünschte Eingabe in das Modell

- Fügen Sie dem Quellenknoten einen GenLin-Knoten hinzu. Klicken Sie im GenLin-Knoten auf die Registerkarte Modell.

- Wählen Sie Erste (niedrigster Wert) als Referenzkategorie für das Ziel. Dies gibt an, dass die zweite Kategorie das relevante Ereignis ist, und ihr Effekt auf das Modell liegt in der Interpretation der Parameterschätzer. Ein stetiger Prädiktor mit einem positiven Koeffizienten zeigt die gesteigerte Wahrscheinlichkeit eines erneuten Auftretens mit steigenden Werten des Prädiktors an. Kategorien

eines nominalen Prädiktors mit größeren Koeffizienten zeigen die gesteigerte Wahrscheinlichkeit eines erneuten Auftretens im Hinblick auf andere Kategorien des Sets an.

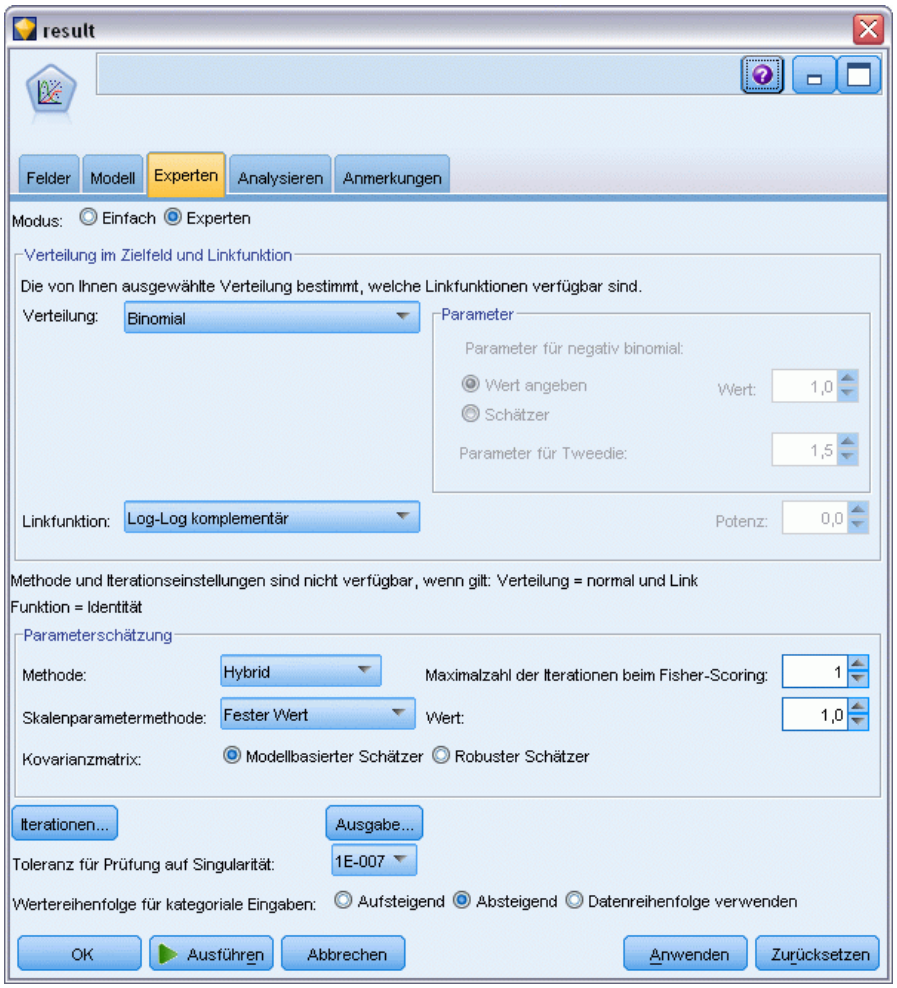
Abbildung 23-5
Auswählen Modelloptionen



- Klicken Sie auf die Registerkarte Experten und wählen Sie Experten, um die Expertenmodellierungsoptionen zu aktivieren.
- Wählen Sie Binomial als Verteilung und Log-Log komplementär als Verknüpfungsfunktion (Linkfunktion).
- Wählen Sie Fester Wert als Methode zur Schätzung des Skalenparameters und behalten Sie den Standardwert 1,0 bei.

- Wählen Sie Absteigend als Reiheinfolge der Kategorien für Faktoren. Dadurch wird angegeben, dass die erste Kategorie jedes Faktors als dessen Referenzkategorie dient; der Effekt dieser Auswahl auf das Modell besteht in der Interpretation von Parameterschätzern.

Abbildung 23-6
Auswählen von Expertenoptionen



- Führen Sie den Stream aus, um das Modell-Nugget zu generieren; dieses wird dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die Modell-Details anzuzeigen, können Sie mit der rechten Maustaste auf das Nugget klicken und Bearbeiten oder Durchsuchen auswählen.

Tests der Modelleffekte

Abbildung 23-7
Tests für Modelleffekte für das Haupteffektmodell

Quelle	Typ III		
	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,536	1	,464
duration	,003	1	,958
treatment	,382	1	,537
age	,358	1	,550
Abhängige Variable: ResultModell: (Konstanter Term), duration, treatment, age			

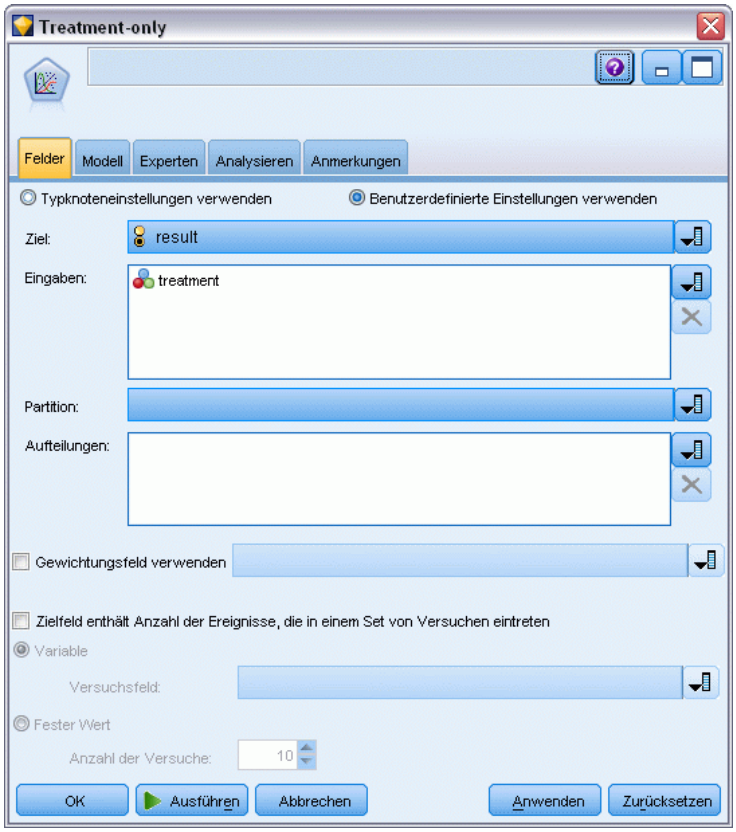
Keiner der Modelleffekte ist statistisch signifikant. Alle beobachtbaren Unterschiede in den Behandlungseffekten sind jedoch von klinischem Interesse, sodass hier ein verkürztes Modell nur mit “Behandlung” als Modellterm angepasst wird.

Anpassen des Modells “Nur Behandlung”

- Klicken Sie im GenLin-Knoten auf der Registerkarte “Felder” auf Benutzerdefinierte Einstellungen verwenden.
- Wählen Sie *result* als Ziel aus.

▶ Wählen Sie *treatment* als einzige Eingabe aus.

Abbildung 23-8
Auswählen von Feldoptionen



▶ Führen Sie den Stream aus und öffnen Sie das resultierende Modell-Nugget.

Wählen Sie auf dem Modell-Nugget die Registerkarte Erweitert und scrollen Sie zum Ende.

Parameter-Schätzer

Abbildung 23-9
Parameterschätzer für das Modell "Nur Behandlung"

Parameter	RegressionskoeffizientB	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	-1,442	,5012	-2,425	-,460	8,282	1	,004
[treatment=1]	,378	,6288	-,855	1,610	,361	1	,548
[treatment=0]	0(a)	.	.	.	.	.	.
(Skala)	1(b)						

Abhängige Variable: ResultModell: (Konstanter Term), treatment, Offset = 0

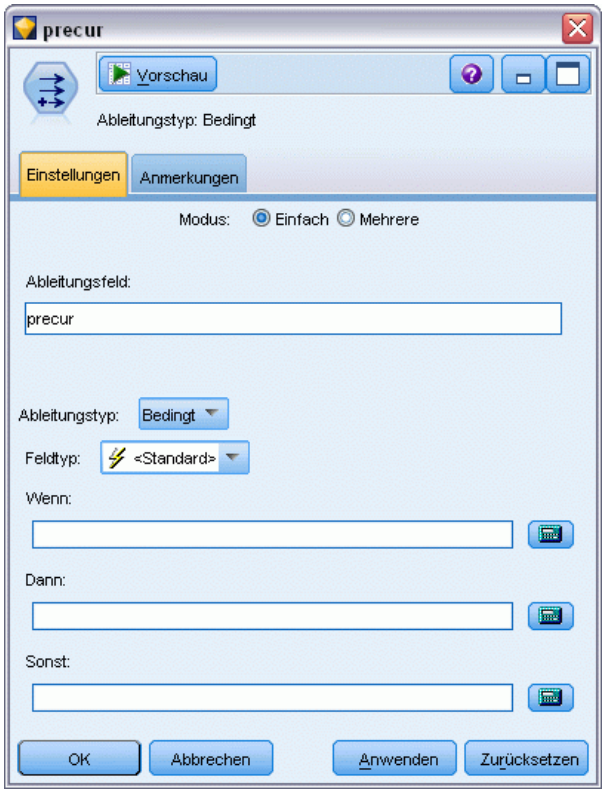
a. Auf 0 gesetzt, da dieser Parameter redundant ist.

b. Auf den angezeigten Wert festgesetzt.

Der Behandlungseffekt (der Unterschied der linearen Einflussvariablen (Prädiktor) zwischen den beiden Behandlungsebenen; d. h., der Koeffizient für *[treatment=1]*) ist auch weiterhin statistisch nicht signifikant, lässt jedoch vermuten, dass Behandlung *A [treatment=0]* möglicherweise besser ist als Behandlung *B [treatment=1]*, da der Parameterschätzer für Behandlung *B* größer ist als der für *A* und somit mit einer gesteigerten Wahrscheinlichkeit des erneuten Auftretens in den ersten 12 Monaten verbunden ist. Die lineare Einflussvariable (Anfangs- und Behandlungseffekt) ist eine Schätzung von $\text{Log}(-\text{Log}(1-\text{P}(\text{recur}_{12,t}))$, wobei $\text{P}(\text{recur}_{12,t})$ die Wahrscheinlichkeit des erneuten Auftretens innerhalb der ersten 12 Monate für Behandlung t(=*A* oder *B*) ist. Diese vorhergesagten Wahrscheinlichkeiten werden für alle Beobachtungen im Datensatz erstellt.

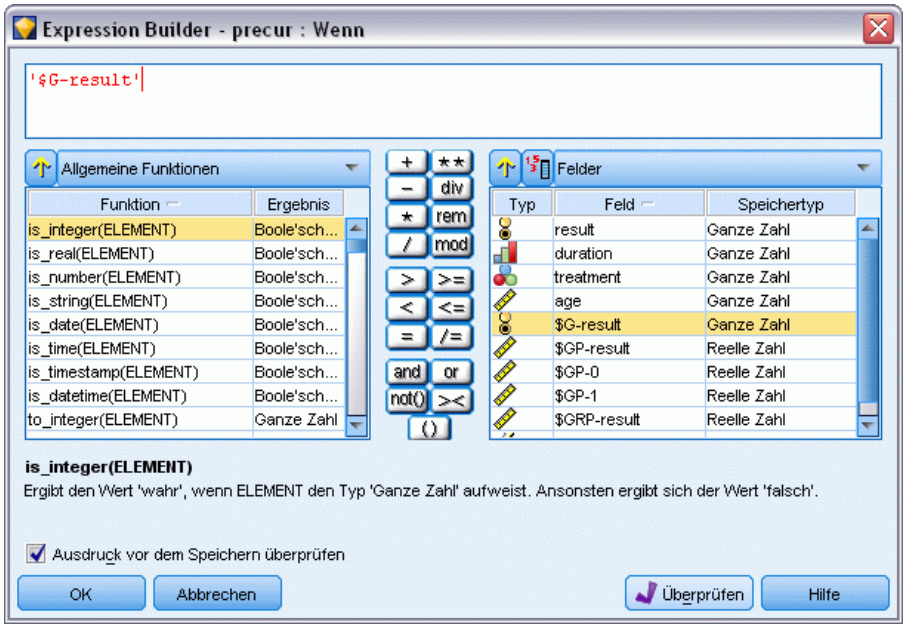
Vorhergesagtes erneutes Auftreten und Überlebenswahrscheinlichkeiten

Abbildung 23-10
Einstellungsoptionen für Ableitungsknoten



- Das Modell scort für jeden Patienten das vorhergesagte Ergebnis und die Wahrscheinlichkeit dieses vorhergesagten Ergebnisses. Um die vorhergesagten Wahrscheinlichkeiten für ein erneutes Auftreten anzuzeigen, kopieren Sie das erstellte Modell in die Palette und fügen Sie einen Ableitungsknoten hinzu.
- Geben Sie auf der Registerkarte “Einstellungen” precur als Ableitungsfeld ein.
- Wählen Sie für die Ableitung des Felds die Option Bedingt aus.
- Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder für die Bedingung Wenn zu öffnen.

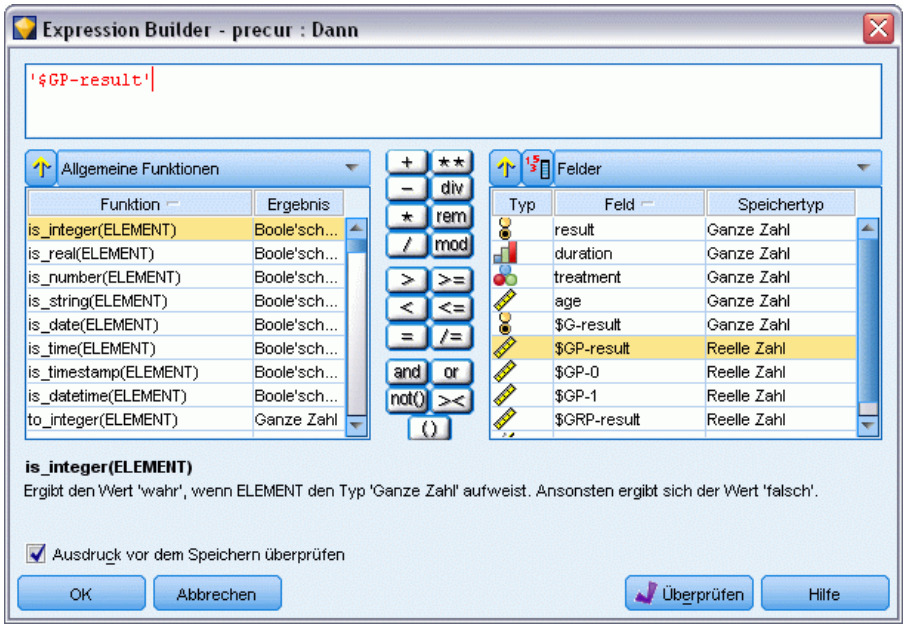
Abbildung 23-11
Ableitungsknoten: Expression Builder für Bedingung "Wenn"



- Fügen Sie das Feld *$G-result* in den Ausdruck ein.
- Klicken Sie auf OK.

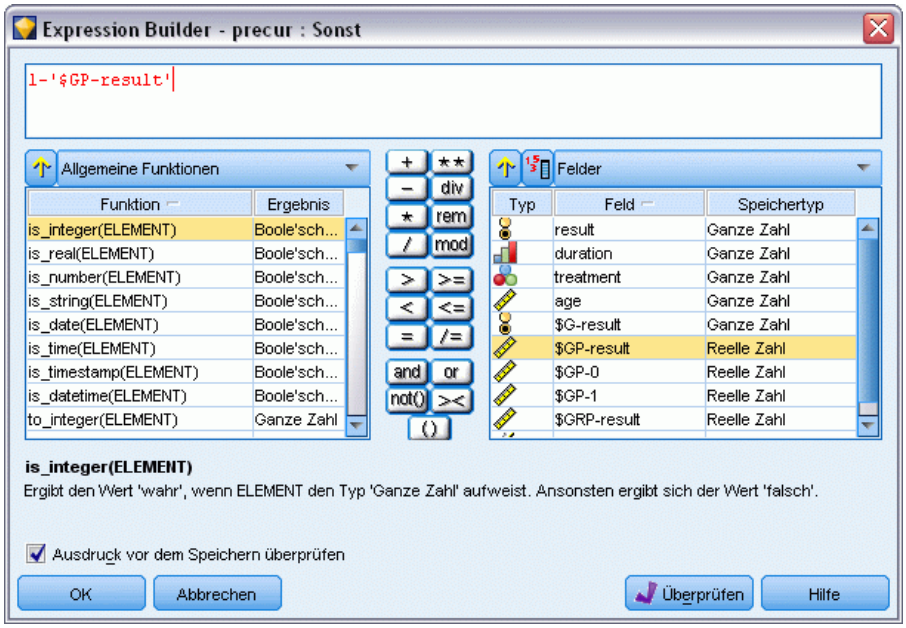
Das Ableitungsfeld *precur* nimmt den Wert des Dann-Ausdrucks an, wenn *$G-result* gleich 1 ist und den Wert des Sonst-Ausdrucks, wenn es gleich 0 ist.

Abbildung 23-12
Ableitungsknoten: Expression Builder für "Dann"-Ausdruck



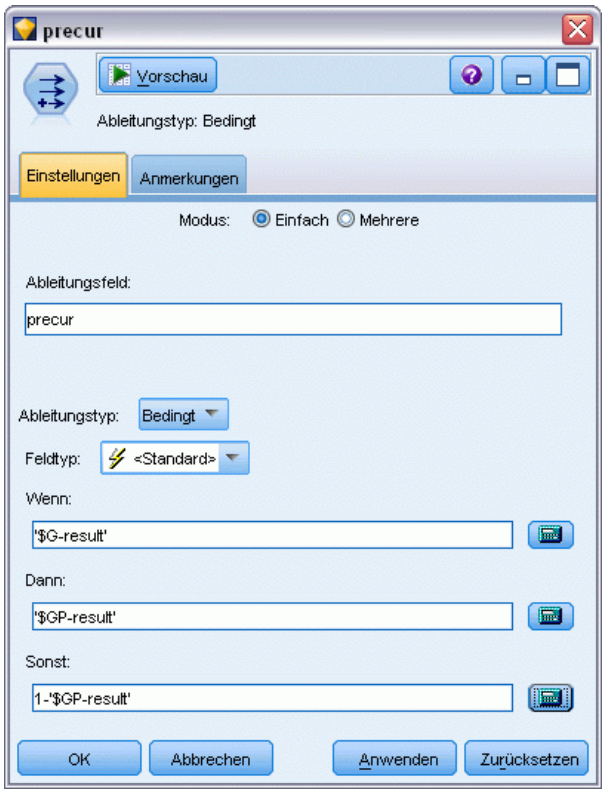
- Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder für den Dann-Ausdruck zu öffnen.
- Fügen Sie das Feld *$GP-result* in den Ausdruck ein.
- Klicken Sie auf OK.

Abbildung 23-13
Ableitungsknoten: Expression Builder für "Sonst"-Ausdruck



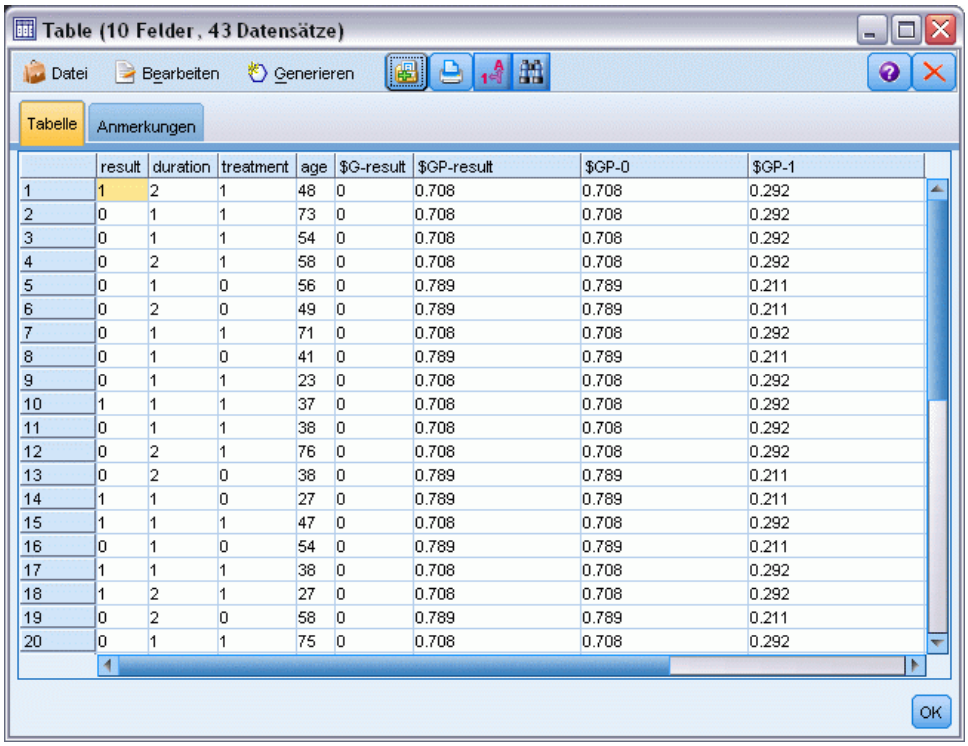
- Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder für den Sonst-Ausdruck zu öffnen.
- Geben Sie 1- in den Ausdruck ein und fügen Sie anschließend das Feld *$GP-result* in den Ausdruck ein.
- Klicken Sie auf OK.

Abbildung 23-14
Einstellungsoptionen für Ableitungsknoten



- Gliedern Sie einen Tabellenknoten an den Ableitungsknoten an und führen Sie ihn aus.

Abbildung 23-15
Geschätzte Wahrscheinlichkeiten



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

Es ist eine geschätzte Wahrscheinlichkeit von 0,211 gegeben, dass bei Patienten, die Behandlung *A* zugewiesen sind, die Krankheit in den ersten 12 Monaten erneut auftritt. Der Wert für Behandlung *B* ist 0,292. Beachten Sie, dass $1-P(recur_{12, t})$ die Überlebenswahrscheinlichkeit nach 12 Monaten ist und damit den interessanteren Wert für Überlebensanalysten darstellt.

Modellieren der Wahrscheinlichkeit eines erneuten Auftretens nach Zeitraum

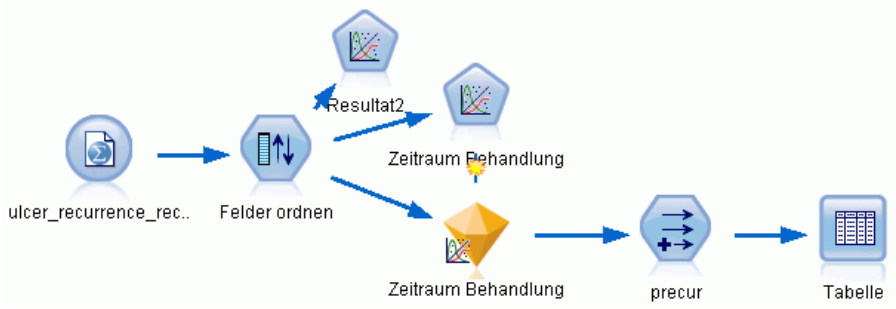
Ein Problem mit dem aktuellen Modell besteht darin, dass es die Informationen, die bei der ersten Untersuchung erfasst wurden, nicht beachtet, nämlich dass bei vielen Patienten während der ersten sechs Monate die Krankheit nicht erneut aufgetreten ist. Ein “besseres” Modell würde eine binäre Antwort modellieren, die für jedes Intervall aufzeichnet, ob das Ereignis aufgetreten ist. Für die Anpassung des Modells ist eine Rekonstruktion des ursprünglichen Daten-Sets erforderlich. Dieses Daten-Set finden Sie in der Datei *ulcer_recurrence_recoded.sav*. Für weitere Informationen siehe Thema Ordner “Demos” in Kapitel 1 in *IBM SPSS Modeler 14.2-Benutzerhandbuch*. Diese Datei enthält zwei zusätzliche Variablen:

- *Period* (Zeitraum) zeichnet auf, ob der Fall dem ersten oder dem zweiten Untersuchungszeitraum entspricht.
- *Result by period* (Ergebnis nach Zeitraum) zeichnet auf, ob in dem entsprechenden Zeitraum beim betreffenden Patienten ein erneutes Auftreten beobachtet wurde.

Jeder ursprüngliche Fall (Patient) trägt einen Fall pro Intervall bei, in dem er im Risikodatensatz bleibt. So trägt beispielsweise Patient 1 zwei Fälle bei – einen für den ersten Untersuchungszeitraum, in dem kein erneutes Auftreten beobachtet wurde, und einen für den zweiten Untersuchungszeitraum, in dem ein erneutes Auftreten beobachtet wurde. Patient 10 dagegen trägt nur einen einzigen Fall bei, da im ersten Zeitraum ein erneutes Auftreten beobachtet wurde. Die Patienten 16, 28 und 34 verließen die Studie nach sechs Monaten und tragen so nur einen einzigen Fall zum neuen Daten-Set bei.

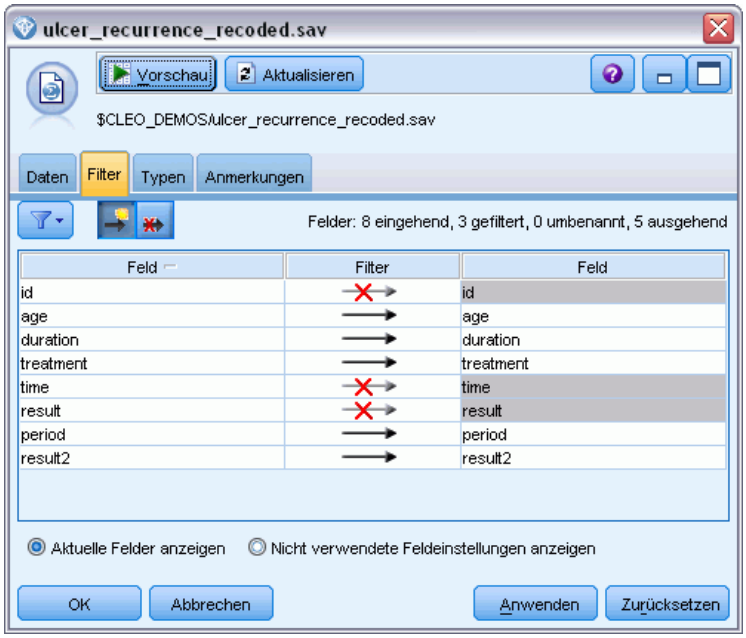
▶ Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *ulcer_recurrence_recoded.sav* im Ordner *Demos* verweist.

Abbildung 23-16
Beispiel-Stream zur Vorhersage des erneuten Auftretens von Geschwüren



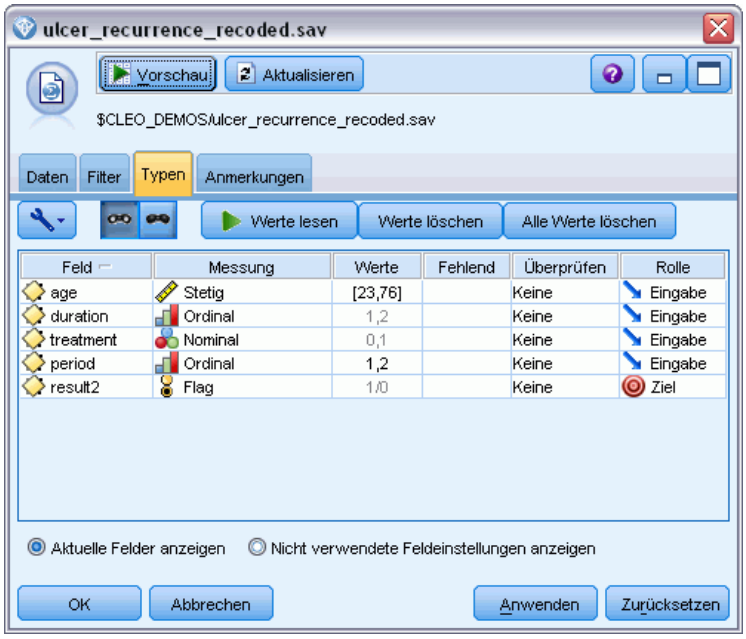
▶ Filtern Sie auf der Registerkarte “Filter” des Quellenknotens *id*, *time* und *result* heraus.

Abbildung 23-17
Filtern von unerwünschten Feldern



- Setzen Sie auf der Registerkarte “Typen” des Quellenknotens die Rolle für das Feld *result2* auf Ziel und setzen Sie das Messniveau auf Flag. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

Abbildung 23-18
Festlegen der Feldrolle



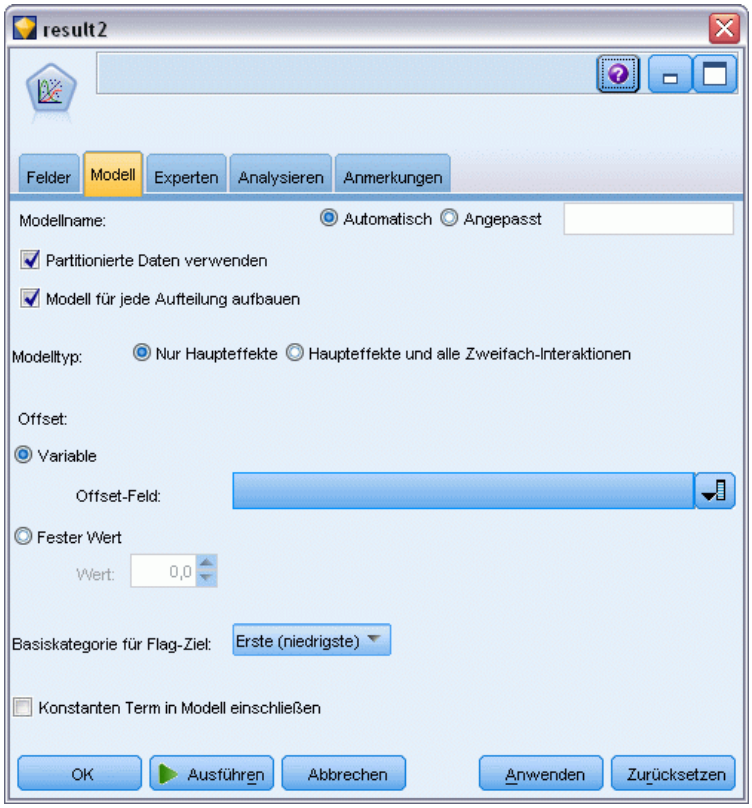
- ügen Sie einen Knoten vom Typ “Felder ordnen” hinzu und legen Sie *period*, *duration*, *treatment* und *age* als Eingabereihenfolge fest. Dadurch, dass *period* die erste Eingabe ist (und dass Sie

den konstanten Term aus dem Modell ausschließen), können Sie ein vollständiges Set von Dummy-Variablen anpassen, um die Zeitraumeffekte zu erfassen.

Abbildung 23-19
Neuordnung von Feldern für die gewünschte Eingabe in das Modell

▶ Klicken Sie im GenLin-Knoten auf die Registerkarte Modell.

Abbildung 23-20
Auswählen Modelloptionen

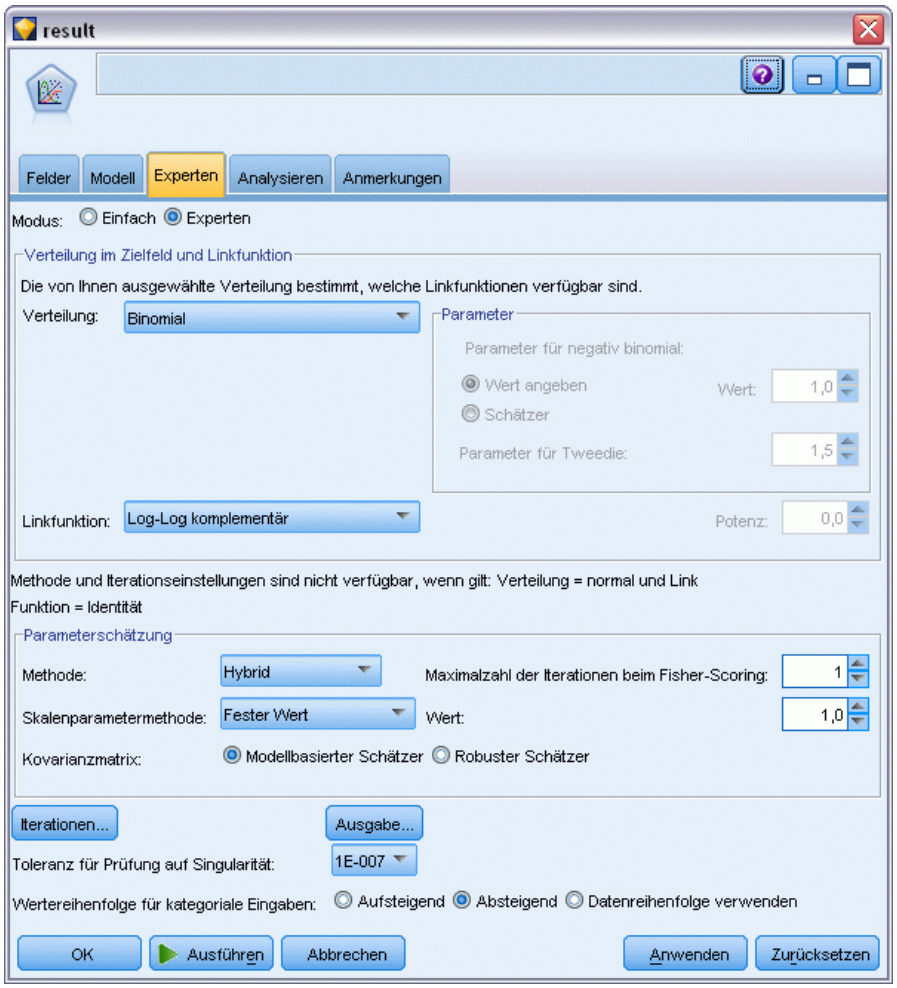


▶ Wählen Sie Erste (niedrigster Wert) als Referenzkategorie für das Ziel. Dies gibt an, dass die zweite Kategorie das relevante Ereignis ist, und ihr Effekt auf das Modell liegt in der Interpretation der Parameterschätzer.

▶ Deaktivieren Sie die Option Konstanten Term in Modell einschließen.

▶ Klicken Sie auf die Registerkarte Experten und wählen Sie Experten, um die Expertenmodellierungsoptionen zu aktivieren.

Abbildung 23-21
Auswählen von Expertenoptionen



▶ Wählen Sie Binomial als Verteilung und Log-Log komplementär als Verknüpfungsfunktion (Linkfunktion).

▶ Wählen Sie Fester Wert als Methode zur Schätzung des Skalenparameters und behalten Sie den Standardwert 1,0 bei.

▶ Wählen Sie Absteigend als Reiheinfolge der Kategorien für Faktoren. Dadurch wird angegeben, dass die erste Kategorie jedes Faktors als dessen Referenzkategorie dient; der Effekt dieser Auswahl auf das Modell besteht in der Interpretation von Parameterschätzern.

▶ Führen Sie den Stream aus, um das Modell-Nugget zu generieren; dieses wird dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die Modell-Details anzuzeigen, können Sie mit der rechten Maustaste auf das Nugget klicken und Bearbeiten oder Durchsuchen auswählen.

Tests der Modelleffekte

Abbildung 23-22
Tests für Modelleffekte für das Haupteffektmodell

Quelle	Typ III		
	Wald-Chi-Quadrat	df	Sig.
period	,464	1	,496
duration	,000	1	,988
treatment	,117	1	,732
age	,314	1	,575
Abhängige Variable: Result by periodModell: period, duration, treatment, age			

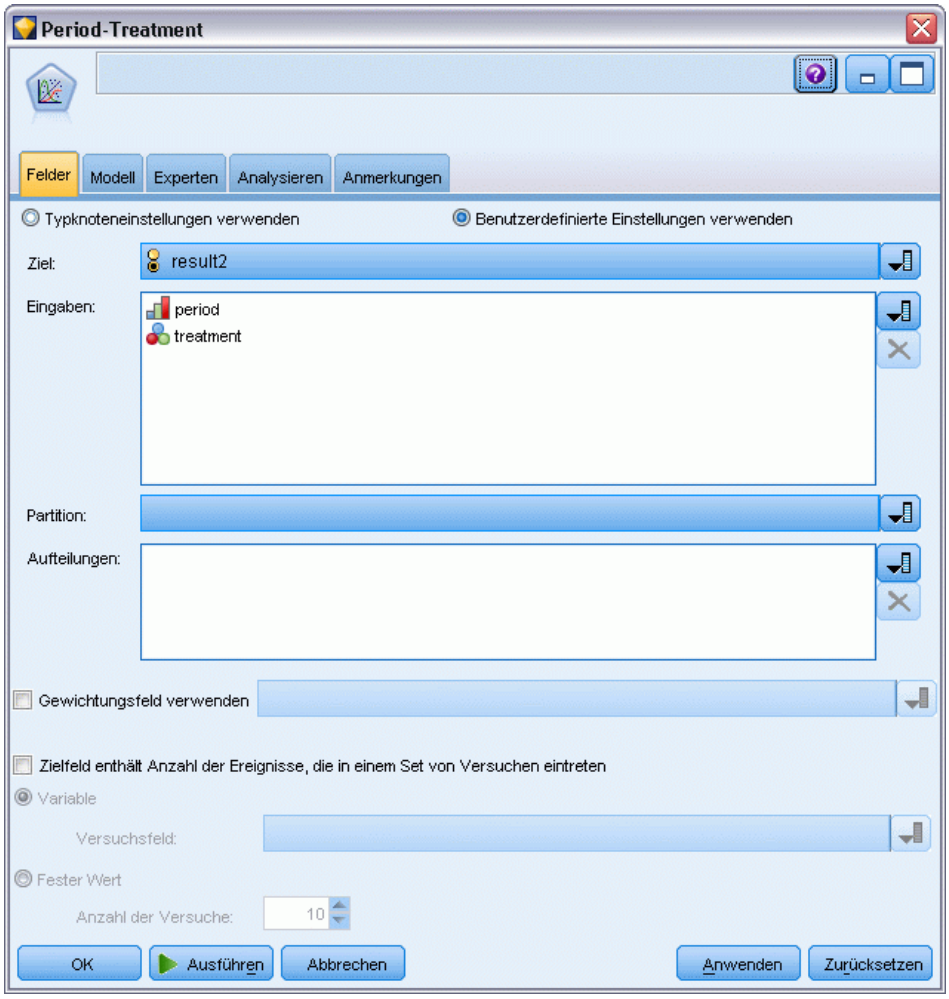
Keiner der Modelleffekte ist statistisch signifikant. Alle beobachtbaren Unterschiede in den Zeitraum- und Behandlungseffekten sind jedoch von klinischem Interesse, sodass hier ein verkürztes Modell mit nur diesen Modelltermen angepasst wird.

Anpassen des verkürzten Modells

- Klicken Sie im GenLin-Knoten auf der Registerkarte “Felder” auf Benutzerdefinierte Einstellungen verwenden.
- Wählen Sie *result2* als Ziel aus.

▶ Wählen Sie *period* und *treatment* als Eingaben aus.

Abbildung 23-23
Auswählen von Feldoptionen



▶ Führen Sie den Knoten aus und durchsuchen Sie das erstellte Modell. Kopieren Sie anschließend das erstellte Modell in die Palette, fügen Sie einen Tabellenknoten hinzu und führen Sie ihn aus.

Parameter-Schätzer

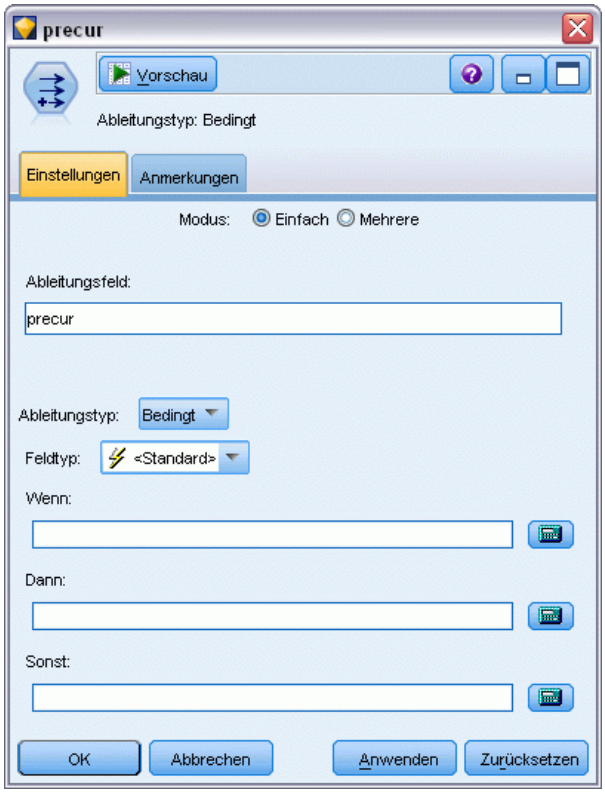
Abbildung 23-24
Parameterschätzer für das Modell "Nur Behandlung"

Parameter	RegressionskoeffizientB	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
[period=2]	-1,794	,5792	-2,929	-,659	9,597	1	,002
[period=1]	-2,206	,5912	-3,365	-1,047	13,926	1	,000
[treatment=1]	,195	,6279	-1,035	1,426	,097	1	,756
[treatment=0]	0(a)	.	.	.	.	.	.
(Skala)	1(b)						
Abhängige Variable: Result by periodModell: (Konstanter Term), period, treatment, Offset = 0							
a. Auf 0 gesetzt, da dieser Parameter redundant ist.							
b. Auf den angezeigten Wert festgesetzt.							

Der Behandlungseffekt ist auch weiterhin statistisch nicht signifikant, weist jedoch darauf hin, dass Behandlung *A* möglicherweise besser ist als Behandlung *B*, da die Parameterschätzung für Behandlung *B* mit einer gesteigerten Wahrscheinlichkeit eines erneuten Auftretens in den ersten 12 Monaten verbunden ist. Die Zeitraumwerte sind statistisch signifikant vershieden von 0. Dies ist jedoch der Fall, weil ein konstanter Term nicht angepasst ist. Der Zeitraumeffekt (der Unterschied zwischen den Werten der linearen Einflussvariablen für *[period=1]* und *[period=2]*) ist nicht statistisch signifikant, wie in den Tests der Modelleffekte deutlich wird. Die lineare Einflussvariable (Zeitraum- und Behandlungseffekt) ist ein Schätzer von $\text{Log}(-\text{Log}(1-P(\text{recur}_{p,t}))$, wobei $P(\text{recur}_{p,t})$ die Wahrscheinlichkeit des erneuten Auftretens im Zeitraum p(=1 oder 2, entsprechend 6 bzw. 12 Monaten) mit Behandlung t(=*A* oder *B*) ist. Diese vorhergesagten Wahrscheinlichkeiten werden für alle Beobachtungen im Datensatz erstellt.

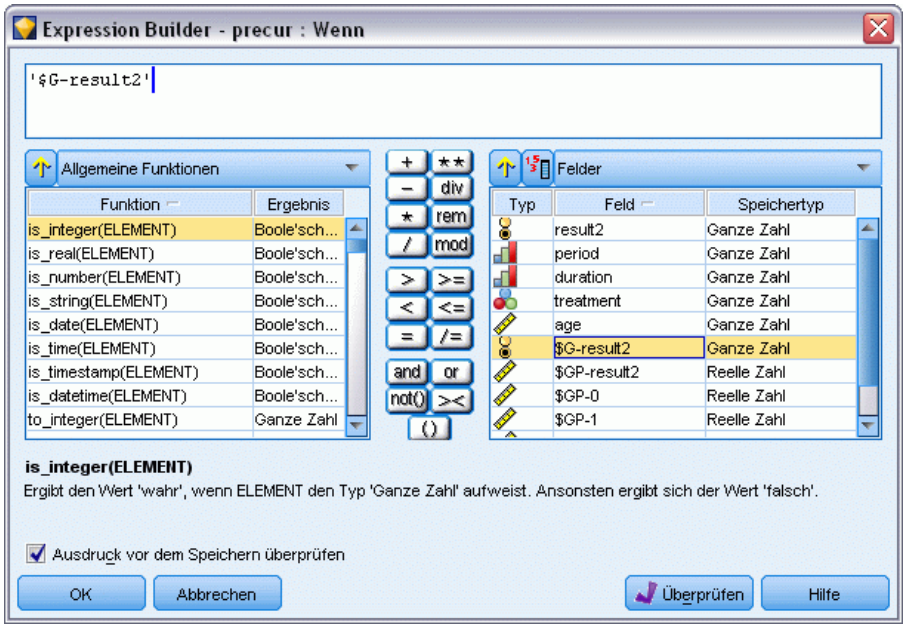
Vorhergesagtes erneutes Auftreten und Überlebenswahrscheinlichkeiten

Abbildung 23-25
Einstellungsoptionen für Ableitungsknoten



- Das Modell scort für jeden Patienten das vorhergesagte Ergebnis und die Wahrscheinlichkeit dieses vorhergesagten Ergebnisses. Um die vorhergesagten Wahrscheinlichkeiten für ein erneutes Auftreten anzuzeigen, kopieren Sie das erstellte Modell in die Palette und fügen Sie einen Ableitungsknoten hinzu.
- Geben Sie auf der Registerkarte “Einstellungen” precur als Ableitungsfeld ein.
- Wählen Sie für die Ableitung des Felds die Option Bedingt aus.
- Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder für die Bedingung Wenn zu öffnen.

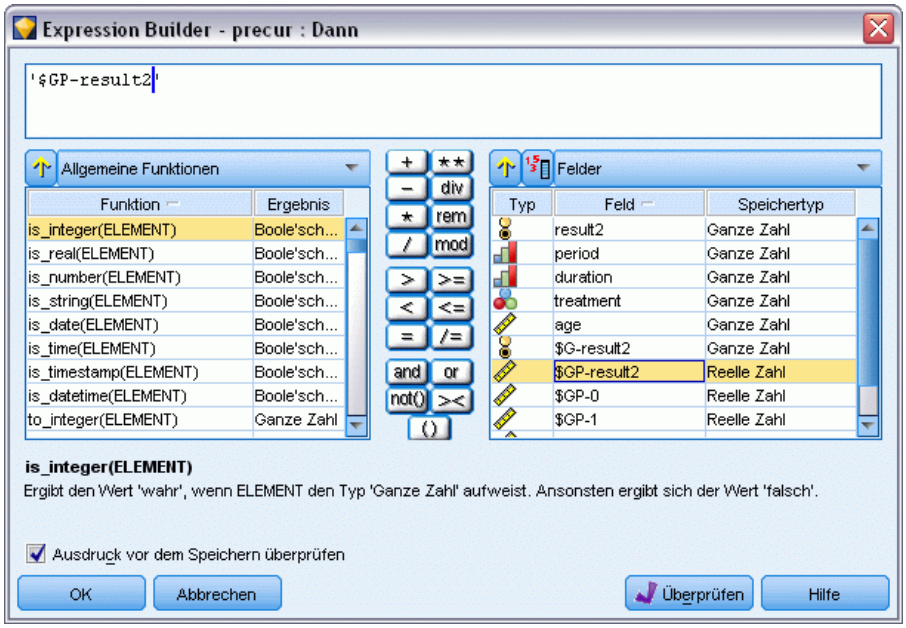
Abbildung 23-26
Ableitungsknoten: Expression Builder für Bedingung "Wenn"



- Fügen Sie das Feld *$G-result2* in den Ausdruck ein.
- Klicken Sie auf OK.

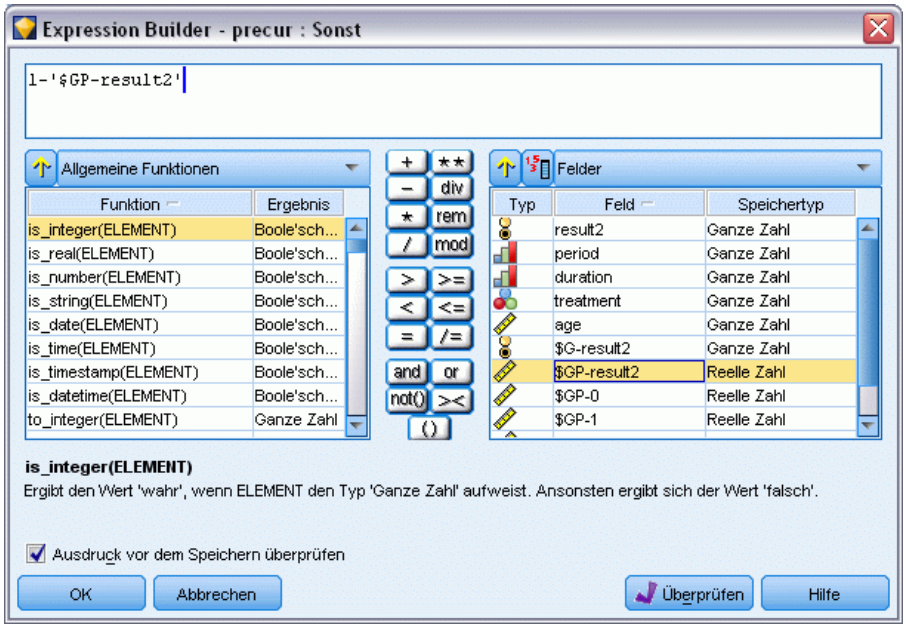
Das Ableitungsfeld *precur* nimmt den Wert des Dann-Ausdrucks an, wenn *$G-result2* gleich 1 ist, und den Wert des Sonst-Ausdrucks, wenn es gleich 0 ist.

Abbildung 23-27
Ableitungsknoten: Expression Builder für "Dann"-Ausdruck



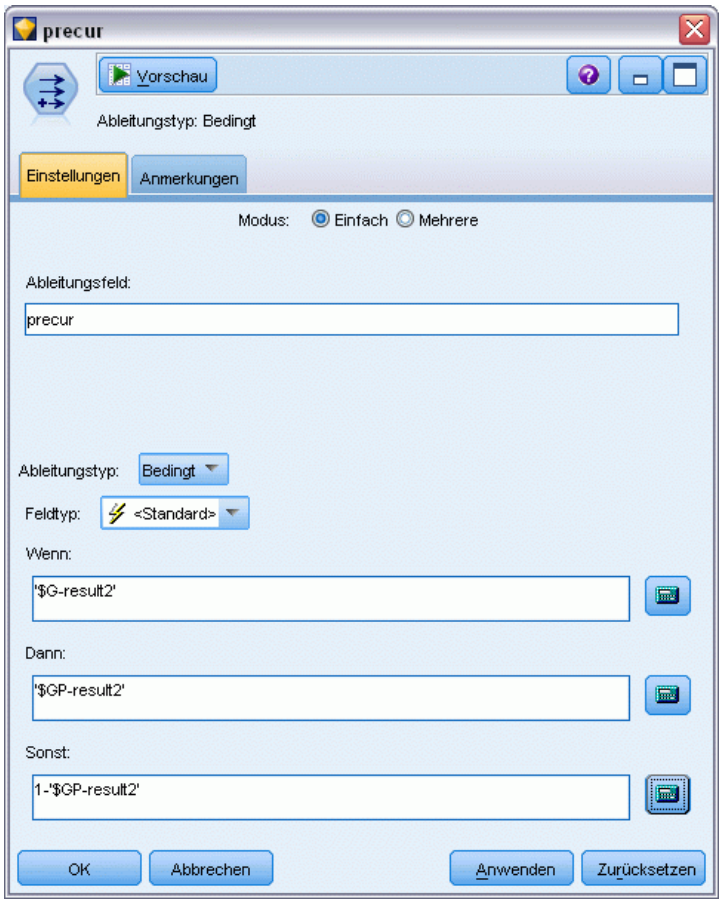
- Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder für den Dann-Ausdruck zu öffnen.
- Fügen Sie das Feld *$GP-result2* in den Ausdruck ein.
- Klicken Sie auf OK.

Abbildung 23-28
Ableitungsknoten: Expression Builder für "Sonst"-Ausdruck



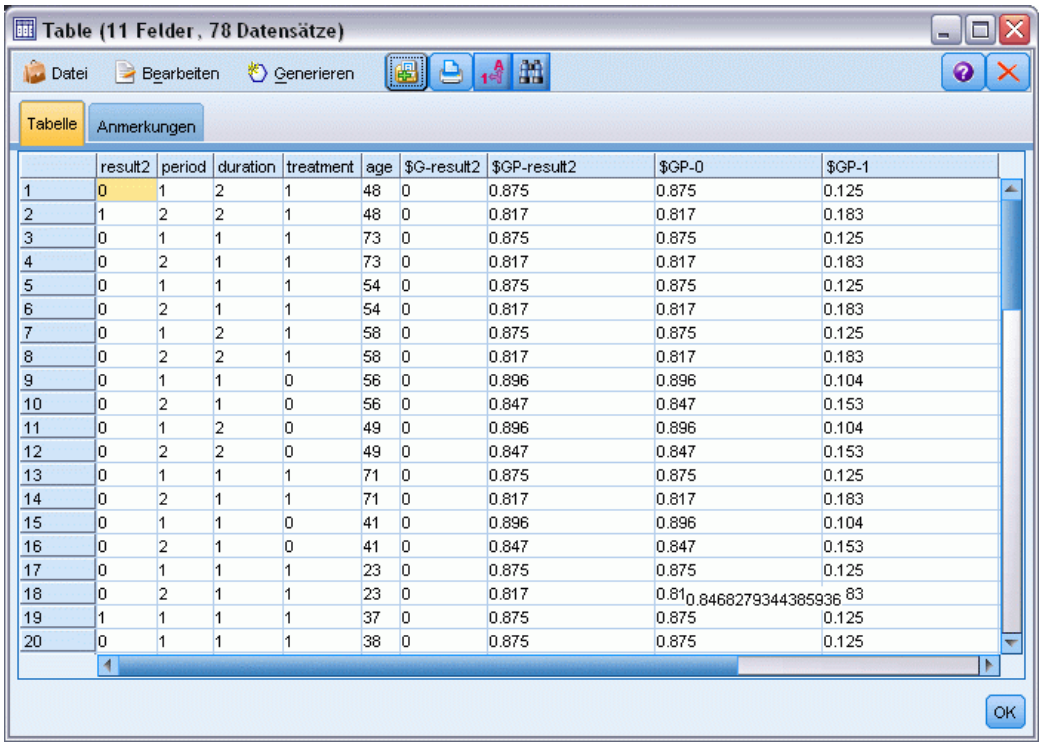
- Klicken Sie auf die Taschenrechner-Schaltfläche, um Expression Builder für den Sonst-Ausdruck zu öffnen.
- Geben Sie 1- in den Ausdruck ein und fügen Sie anschließend das Feld *$GP-result2* in den Ausdruck ein.
- Klicken Sie auf OK.

Abbildung 23-29
Einstellungsoptionen für Ableitungsknoten



- Gliedern Sie einen Tabellenknoten an den Ableitungsknoten an und führen Sie ihn aus.

Abbildung 23-30
Geschätzte Wahrscheinlichkeiten



	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.81$_{0.8468279344385936}$83	
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

Die geschätzten Wahrscheinlichkeiten für ein erneutes Auftreten lassen sich wie folgt zusammenfassen:

Behandlung	**6 Monate**	**12 Monate**
D	0.104	0.153
K	0.125	0.183

Hieraus kann die Überlebenswahrscheinlichkeit für 12 Monate geschätzt werden als $1-(P(recur_{1,t}) + P(recur_{2,t})\times(1-P(recur_{1,t})))$; d. h. für jede Behandlung:

A: $1 - (0.104 + 0.153*0.896) = 0.759$

B: $1 - (0.125 + 0.183*0.875) = 0.715$

wodurch sich erneut allerdings auch hier wieder ohne statistisch signifikante Unterstützung zeigt, dass Behandlung *A* die bessere Behandlung ist.

Zusammenfassung

Sie haben mithilfe der verallgemeinerten linearen Modelle eine Reihe von komplementären Log-Log-Regressions-Modellen für intervallzensierte Überlebensdaten angepasst. Auch wenn einiges dafür spricht, Behandlung *A* zu wählen, sollte für statistisch signifikante Ergebnisse eine

umfangreichere Studie durchgeführt werden. Die vorhandenen Daten können jedoch für weitere Berechnungen genutzt werden.

- Es könnte sich lohnen, das Modell mit Interaktionseffekten erneut anzupassen, insbesondere zwischen *Period* (Zeitraum) und *Treatment group* (Behandlungsgruppe).

Erläuterungen der mathematischen Grundlagen der in IBM® SPSS® Modeler verwendeten Modellierungsverfahren sind im *SPSS Modeler-Algorithmushandbuch* aufgeführt.

Kapitel

24

Verwenden der Poisson-Regression für die Analyse von Schiffsschadensraten (Verallgemeinerte lineare Modelle)

Generalisierte lineare Modelle können zur Anpassung einer Poisson-Regression für die Analyse von Häufigkeitsdaten verwendet werden. So bezieht sich beispielsweise ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe der Werte der Einflussvariablen gemäß einer Poisson-Rate modelliert werden. Anhand des so entstandenen Modells kann ermittelt werden, welche Schiffstypen am havarieanfälligsten sind.

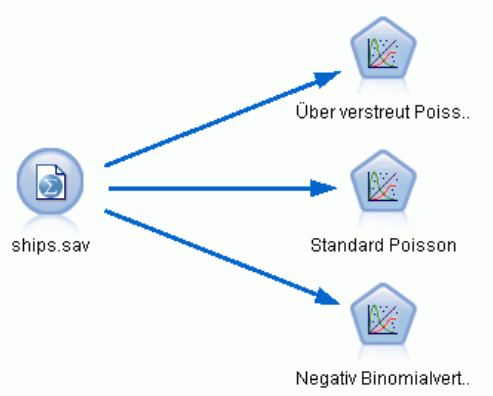
In diesem Beispiel wird der Stream *ships_genlin.str* verwendet, der auf die Datendatei *ships.sav* verweist. Die Datendatei befindet sich im Ordner *Demos* und die Stream-Datei im Unterordner *streams*. Für weitere Informationen siehe Thema Ordner "Demos" in Kapitel 1 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Die Modellierung der Rohzellenhäufigkeiten kann in dieser Situation zu falschen Ergebnissen führen, da *Aggregate months of service* (Aggregat der Betriebsmonate) je nach Schiffstyp variiert. Variablen wie diese, die die Höhe der Risiken messen, werden im verallgemeinerten linearen Modell als Offset-Variablen behandelt. Zudem wird in einer Poisson-Regression angenommen, dass der Logarithmus der abhängigen Variablen in den Einflussvariablen linear ist. Um also verallgemeinerte lineare Modelle zu nutzen, um eine Poisson-Regression an die Unfallraten anzupassen, müssen Sie *Logarithm of aggregate months of service* (Logarithmus des Aggregats der Betriebsmonate) verwenden.

Anpassen einer Poisson-Regression mit Überdispersion

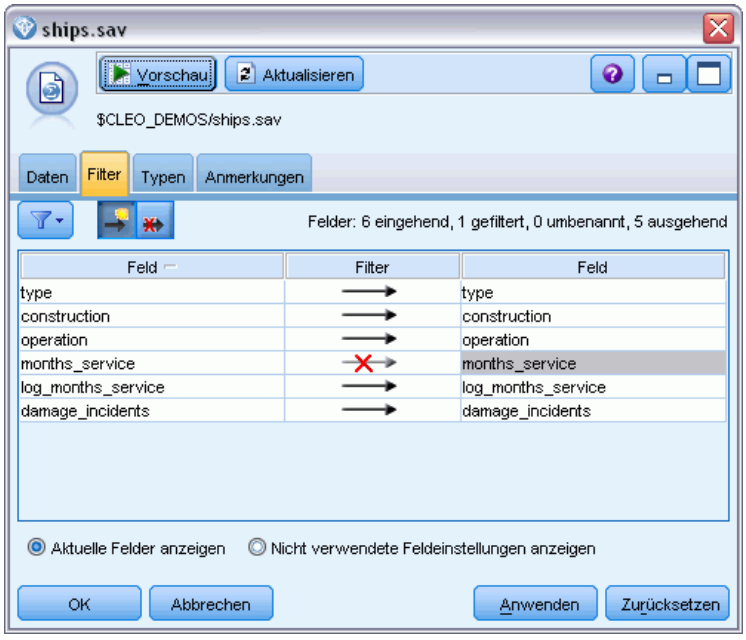
- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *ships.sav* im Ordner *Demos* verweist.

Abbildung 24-1
Beispiel-Stream für die Analyse von Schadensraten



- Schließen Sie auf der Registerkarte "Filter" des Quellenknotens das Feld *months_service* aus. Die Log-transformierten Werte dieser Variablen befinden sich im Feld *log_months_service*, das für die Analyse verwendet wird.

Abbildung 24-2
Filtern eines nicht benötigten Felds

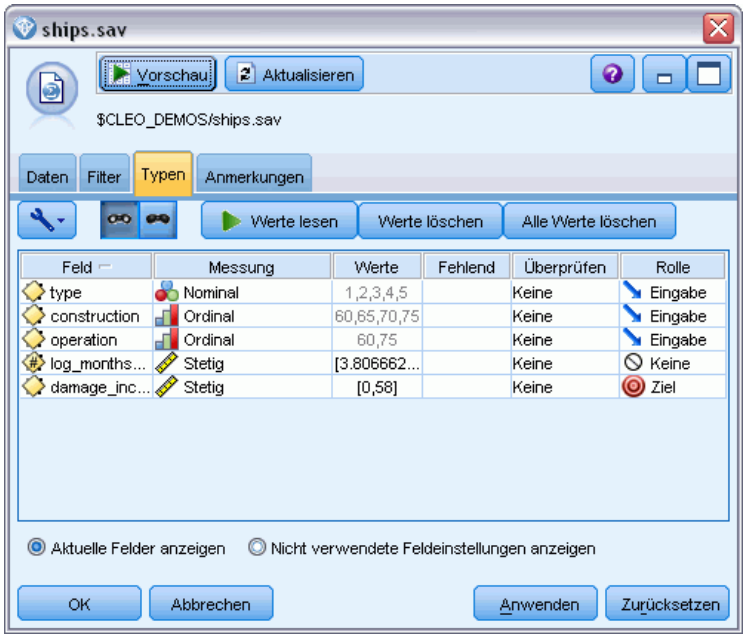


(Alternativ können Sie die Rolle für dieses Feld auf der Registerkarte "Typen" in Keine ändern, anstatt es auszuschließen, oder die gewünschten Felder im Modellierungsknoten auswählen.)

- Setzen Sie auf der Registerkarte "Typen" des Quellenknotens die Rolle für das Feld *damage_incidents* auf Ziel. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

▶ Klicken Sie auf Werte lesen, um die Daten zu instanziieren.

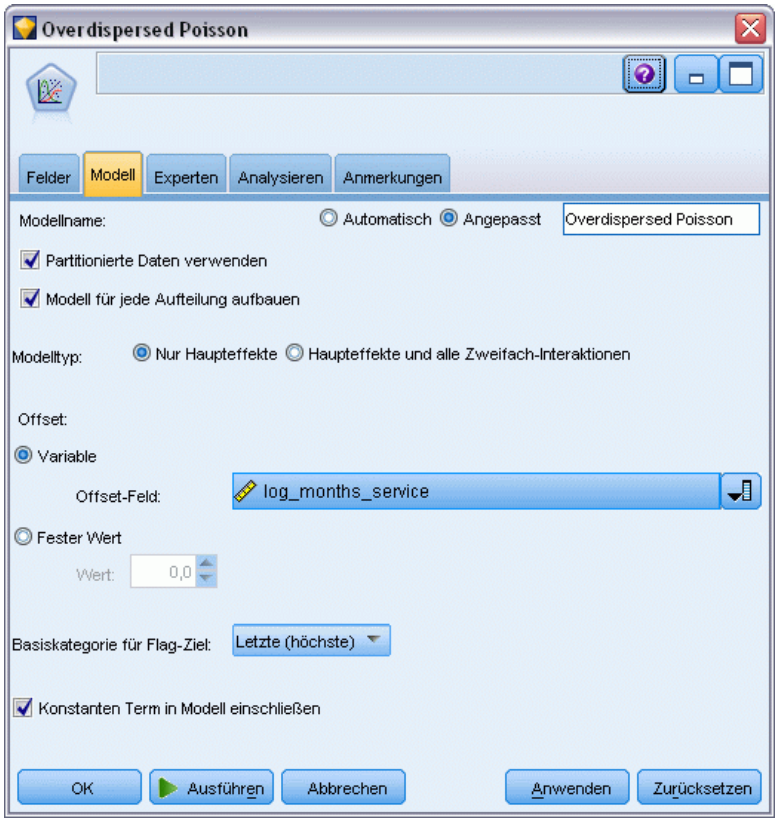
Abbildung 24-3
Festlegen der Feldrolle



▶ Fügen Sie dem Quellenknoten einen Genlin-Knoten hinzu. Klicken Sie im Genlin-Knoten auf die Registerkarte Modell.

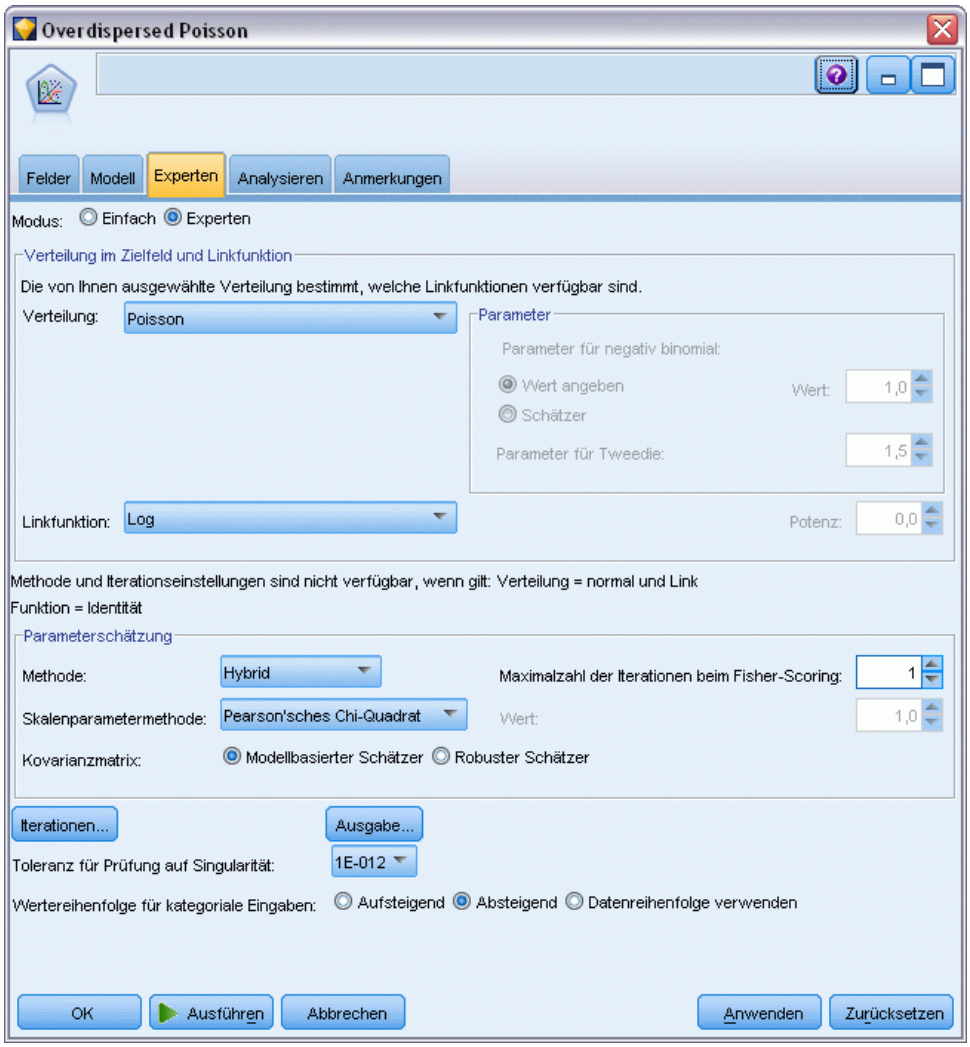
- Wählen Sie *log_months_service* als Offset-Variable.

Abbildung 24-4
Auswählen Modelloptionen



- Klicken Sie auf die Registerkarte Experten und wählen Sie Experten, um die Expertenmodellierungsoptionen zu aktivieren.

Abbildung 24-5
Auswählen von Expertenoptionen



- Wählen Sie Poisson als Verteilung für die Antwort und Log als Verknüpfungsfunktion.
- Wählen Sie Pearson-Chi-Quadrat als Methode zur Schätzung des Skalenparameters. Der Skalienparameter wird in einer Poisson-Regression üblicherweise mit 1 angegeben, doch McCullagh und Nelder nutzen die Pearson-Chi-Quadratschätzung, um konservativere Varianzschätzer und Signifikanzniveaus zu erhalten.
- Wählen Sie Absteigend als Reiheinfolge der Kategorien für Faktoren. Dadurch wird angegeben, dass die erste Kategorie jedes Faktors als dessen Referenzkategorie dient; der Effekt dieser Auswahl auf das Modell besteht in der Interpretation von Parameterschätzern.
- Klicken Sie auf Ausführen, um das Modell-Nugget zu erstellen; dieses wird dem Stream-Zeichenbereich und der Modellpalette in der rechten oberen Ecke hinzugefügt. Zum

Anzeigen der Modelldetails klicken Sie mit der rechten Maustaste auf das Nugget und wählen Bearbeiten oder Durchsuchen. Anschließend klicken Sie auf die Registerkarte Erweitert.

Statistik für Anpassungsgüte

Abbildung 24-6
Statistik für Anpassungsgüte

	Wert	df	Wert/df
Abweichung	38,695	25	1,548
Skalierte Abweichung	22,883	25	
Pearson-Chi-Quadrat	42,275	25	1,691
Skaliertes Pearson-Chi-Quadrat	25,000	25	
Log-Likelihood(b,c)	-68,281		
Log-Likelihood (angepasst)(d)	-40,379		
Akaike-Informations-Kriterium (AIC)	154,562		
AIC mit Korrektur für endliche Stichproben (AICC)	162,062		
Bayes-Informationskriterium (BIC)	168,299		
Konsistentes AIC (CAIC)	177,299		
Abhängige Variable: Number of damage incidentsModell: (Konstanter Term), type, construction, operation, Offset = log_months_service			
a. Die Informationskriterien liegen in einem möglichst kleinem Format vor.			
b. Die vollständige Log-Likelihood-Funktion wird angezeigt und bei der Berechung der Informationskriterien verwendet.			

Die Tabelle zur Statistik der Anpassungsgüte dient als Hilfe für den Vergleich von konkurrierenden Modellen. Zudem stellt *Wert/df* für die Abweichungs- und die Pearson-Chi-Quadratstatistik entsprechende Schätzer für den Skalenparameter zur Verfügung. Diese Werte sollten für eine Poisson-Regression nahe 1,0 liegen. Die Tatsache, dass sie über 1,0 liegen, gibt an, dass die Anpassung des Modells mit Überdispersion brauchbar sein kann.

Omnibus-Test

Abbildung 24-7
Omnibus-Test

Likelihood-Quotienten-Chi-Quadrat	df	Sig.
107,633	8	,000
Abhängige Variable: Number of damage incidentsModell: (Konstanter Term), type, construction, operation, Offset = log_months_service		
a. Vergleicht das angepasste Modell mit dem Modell mit ausschließlich konstanten Termen.		

Der Omnibus-Test ist ein Likelihood-Quotient-Chi-Quadrat-Test des aktuellen Modells im Vergleich zum Null-Modell (hier: Intercept-Modell (Modell mit konstantem Term). Der Signifikanzwert von weniger als 0,05 und zeigt an, dass das aktuelle Modell besser geeignet ist als das Null-Modell.

Tests der Modelleffekte

Abbildung 24-8
Tests der Modelleffekte

Quelle	Typ III		
	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	2138,657	1	,000
type	15,415	4	,004
construction	17,242	3	,001
operation	6,249	1	,012
Abhängige Variable: Number of damage incidentsModell: (Konstanter Term), type, construction, operation, Offset = log_months_service			

Jeder Term im Modell wird darauf getestet, ob er einen Effekt hat. Terme mit Signifikanzwerten von weniger als 0,05 weisen einen erkennbaren Effekt auf. Alle der Haupteffekt-Terme tragen einen Teil zum Modell bei.

Parameter-Schätzer

Abbildung 24-9
Parameterschätzer

Parameter	RegressionskoeffizientB	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	-6,406	,2828	-6,960	-5,852	513,238	1	,000
[type=5]	,326	,3067	-,276	,927	1,127	1	,288
[type=4]	-,076	,3779	-,817	,665	,040	1	,841
[type=3]	-,687	,4279	-1,526	,151	2,581	1	,108
[type=2]	-,543	,2309	-,996	-,091	5,536	1	,019
[type=1]	0(a)	.	.	.	.	.	.
[construction=75]	,453	,3032	-,141	1,048	2,236	1	,135
[construction=70]	,818	,2208	,386	1,251	13,743	1	,000
[construction=65]	,697	,1946	,316	1,079	12,835	1	,000
[construction=60]	0(a)	.	.	.	.	.	.
[operation=75]	,384	,1538	,083	,686	6,249	1	,012
[operation=60]	0(a)	.	.	.	.	.	.
(Skala)	1,691(b)						

Abhängige Variable: Number of damage incidentsModell: (Konstanter Term), type, construction, operation, Offset = log_months_service

a. Auf 0 gesetzt, da dieser Parameter redundant ist.

b. Anhand des Pearson-Chi-Quadrats berechnet.

Die Tabelle der Parameterschätzer fasst den Effekt der einzelnen Einflussvariablen zusammen. Aufgrund des Charakters der Verknüpfungsfunktion ist die Interpretation der Koeffizienten in diesem Modell zwar schwierig, die Vorzeichen der Koeffizienten für Kovariaten und die relativen Werte der Koeffizienten für Faktorstufen können jedoch wichtige Einblicke in die Effekte der Einflussvariablen im Modell bieten.

- Bei Kovariaten weisen positive (negative) Koeffizienten auf positive (inverse) Beziehungen zwischen Einflussvariablen und Ergebnis hin. Ein steigender Wert einer Kovariaten mit einem positiven Koeffizienten entspricht einer steigenden Rate an Schadensfällen.
- Für Faktoren gibt eine Faktorstufe mit einem größeren Koeffizienten eine größere Schadenshäufigkeit an. Das Vorzeichen eines Koeffizienten für eine Faktorstufe hängt von dem Effekt der betreffenden Faktorstufe in Bezug zur Referenzkategorie ab.

Auf der Grundlage der Parameterschätzer sind folgende Interpretationen möglich:

- Schiffstyp *B [type=2]* weist eine statistisch signifikant (*p*-Wert: 0,019) niedrigere Schadensrate (geschätzter Koeffizient: –0,543) als Schiffstyp *A [type=1]* (Referenzkategorie) auf. Der geschätzten Parameter von Typ *C [type=3]* ist niedriger als der von Typ *B*, aber die

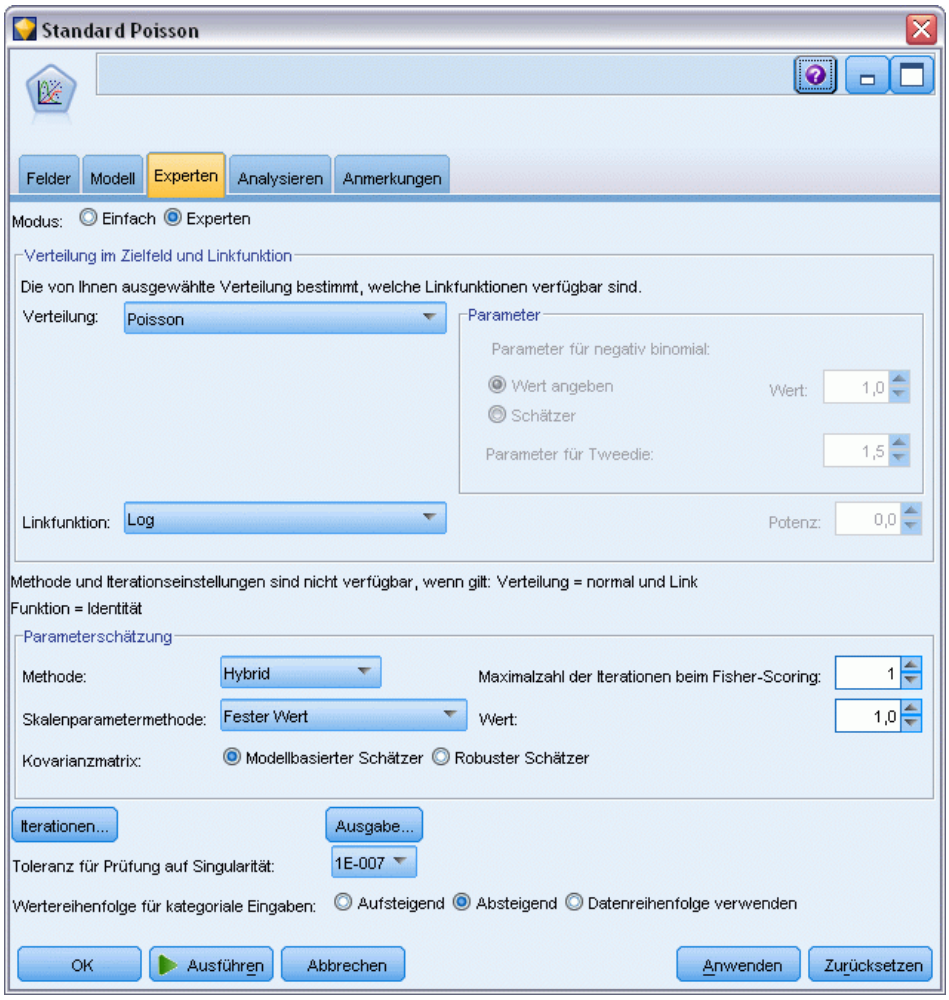
Variabilität in der Schätzung für *C* verwischen diesen Effekt. Weitere Informationen zu den Beziehungen zwischen Faktorstufen erhalten Sie über die geschätzten Randmittel.

- Schiffe, die in den Zeiträumen 1965–69 *[construction=65]* und 1970–74 *[construction=70]* gebaut wurden, haben statistisch signifikant (*p*-Werte <0,001) höhere Schadensraten (geschätzte Koeffizienten: 0,697 bzw. 0,818) als Schiffe, die im Zeitraum 1960–64 *[construction=60]* (Referenzkategorie) gebaut wurden. Weitere Informationen zu den Beziehungen zwischen Faktorstufen erhalten Sie über die geschätzten Randmittel.
- Schiffe, die im Zeitraum 1975–79 *[operation=75]* in Betrieb waren, haben statistisch signifikant (*p*-Wert: 0,012) höhere Schadensraten (geschätzter Koeffizient: 0,384) als Schiffe, die im Zeitraum 1960–1974 *[operation=60]* in Betrieb waren.

Anpassen alternativer Modelle

Ein Problem mit der "überdispersierten" Poisson-Regression besteht darin, dass es keinen formellen Weg gibt, sie im Vergleich zur "standardmäßigen" Poisson-Regression zu testen. Ein Vorschlag für einen formellen Test für die Feststellung von Überdispersion ist jedoch ein Likelihood-Quotienten-Test zwischen einer "standardmäßigen" Poisson-Regression und einer negativen binomialen Regression, wobei alle sonstigen Einstellungen gleich bleiben. Wenn keine Überdispersion in der Poisson-Regression vorliegt, sollte die Statistik −2×(Log-Likelihood für Poisson-Modell − Log-Likelihood für negatives binomiales Modell) eine gemischte Verteilung mit der Hälfte der Wahrscheinlichkeitsmasse bei 0 und dem Rest in einer Chi-Quadrat-Verteilung mit 1 Freiheitsgrad aufweisen.

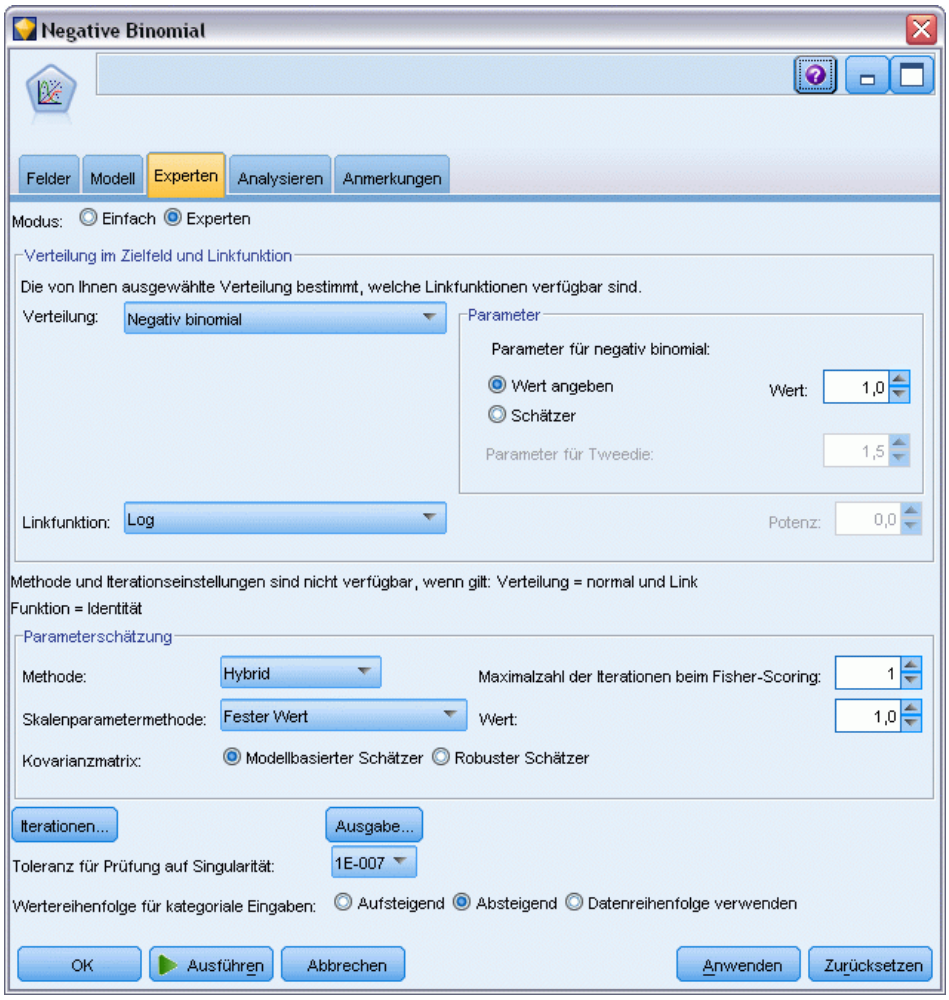
Abbildung 24-10
Registerkarte "Experten"



Um die "standardmäßige" Poisson-Regression anzupassen, kopieren Sie den geöffneten Genlin-Knoten, gliedern ihn an den Quellenknoten an, öffnen den neuen Knoten und klicken auf die Registerkarte Experten.

- Wählen Sie Fester Wert als Methode zur Schätzung des Skalenparameters. Standardmäßig ist dieser Wert 1.

Abbildung 24-11
Registerkarte "Experten"



- Um die negative binominale Regression anzupassen, kopieren Sie den geöffneten Genlin-Knoten, gliedern ihn an den Quellenknoten an, öffnen den neuen Knoten und klicken auf die Registerkarte Experten.
- Wählen Sie Negativ binomial als Verteilung. Behalten Sie den Standardwert 1 für den Hilfsparameter bei.
- Führen Sie den Stream aus und durchsuchen Sie die Registerkarte "Erweitert" auf den neu erstellten Modell-Nuggets.

Statistik für Anpassungsgüte

Abbildung 24-12
Statistiken der Anpassungsgüte für die standardmäßige Poisson-Regression

	Wert	df	Wert/df
Abweichung	38,695	25	1,548
Skalierte Abweichung	38,695	25	
Pearson-Chi-Quadrat	42,275	25	1,691
Skaliertes Pearson-Chi-Quadrat	42,275	25	
Log-Likelihood(b)	-68,281		
Akaike-Informations-Kriterium (AIC)	154,562		
AIC mit Korrektur für endliche Stichproben (AICC)	162,062		
Bayes-Informationskriterium (BIC)	168,299		
Konsistentes AIC (CAIC)	177,299		
Abhängige Variable: Number of damage incidentsModell: (Konstanter Term), type, construction, operation, Offset = log_months_service			
a. Die Informationskriterien liegen in einem möglichst kleinem Format vor.			
b. Die vollständige Log-Likelihood-Funktion wird angezeigt und bei der Berechung der Informationskriterien verwendet.			

Die ausgegebene Log-Likelihood für die standardmäßige Poisson-Regression ist –68,281. Vergleichen Sie dies mit dem negativen binomialen Modell.

Abbildung 24-13
Statistiken der Anpassungsgüte für die negative binomiale Regression

	Wert	df	Wert/df
Abweichung	11,145	25	,446
Skalierte Abweichung	11,145	25	
Pearson-Chi-Quadrat	8,815	25	,353
Skaliertes Pearson-Chi-Quadrat	8,815	25	
Log-Likelihood(b)	-83,725		
Akaike-Informations-Kriterium (AIC)	185,450		
AIC mit Korrektur für endliche Stichproben (AICC)	192,950		
Bayes-Informationskriterium (BIC)	199,187		
Konsistentes AIC (CAIC)	208,187		
Abhängige Variable: Number of damage incidentsModell: (Konstanter Term), type, construction, operation, Offset = log_months_service			
a. Die Informationskriterien liegen in einem möglichst kleinem Format vor.			
b. Die vollständige Log-Likelihood-Funktion wird angezeigt und bei der Berechung der Informationskriterien verwendet.			

Die ausgegebene Log-Likelihood für die negative binomiale Regression ist –83,725. Dieser Wert ist *kleiner* als die Log-Likelihood für die Poisson-Regression, was bedeutet (auch ohne Likelihood-Quotienten-Test), dass diese negative binomiale Regression keine Verbesserung gegenüber der Poisson-Regression aufweist.

Der gewählte Wert von 1 für den Hilfsparameter der negativen Binomialverteilung ist für dieses Daten-Set jedoch möglicherweise nicht optimal. Eine andere Möglichkeit für einen Test auf Überdispersion besteht darin, ein negatives binomiales Modell mit einem Hilfsparameter von 0 anzupassen und den Lagrange-Multiplikator-Test im Ausgabedialogfeld der Registerkarte "Experten" anzufordern. Wenn der Test nicht signifikant ist, sollte die Überdispersion für dieses Daten-Set kein Problem sein.

Zusammenfassung

Mit verallgemeinerten linearen Modellen haben Sie drei verschiedene Modelle für Häufigkeitsdaten angepasst. Die negative binomiale Regression konnte keine Verbesserung gegenüber Poisson-Regression bieten. Die Poisson-Regression mit Überdispersion scheint eine vernünftige Alternative zum standardmäßigen Poisson-Modell zu sein, doch es gibt keinen formellen Test für die Wahl zwischen diesen beiden Möglichkeiten.

Erläuterungen der mathematischen Grundlagen der in IBM® SPSS® Modeler verwendeten Modellierungsverfahren sind im *SPSS Modeler-Algorithmushandbuch* aufgeführt.

Kapitel
25

Anpassen einer Gamma-Regression an Versicherungsforderungen an Kfz-Versicherungen (Verallgemeinerte lineare Modelle)

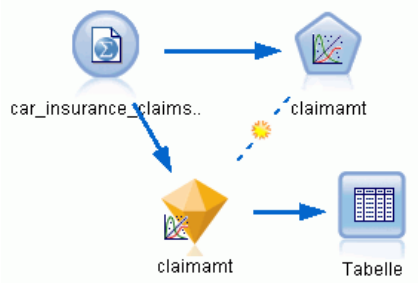
Generalisierte lineare Modelle können zur Anpassung einer Gamma-Regression für die Analyse eines positiven Datenbereichs verwendet werden. So befasst sich beispielsweise ein an anderer Stelle () vorgestelltes und analysiertes Daten-Set Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gamma-Verteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination der Einflussvariablen in Bezug zu setzen. Um die unterschiedliche Anzahl an Forderungen zu berücksichtigen, die zur Berechnung der durchschnittlichen Höhe der Schadensansprüche verwendet wurde, geben Sie *Number of claims* (Anzahl der Forderungen) als Skalierungsgewicht an.

In diesem Beispiel wird der Stream *car-insurance_genlin.str* verwendet, der auf die Datendatei *car_insurance_claims.sav* verweist. Die Datendatei befindet sich im Ordner *Demos* und die Stream-Datei im Unterordner *streams*. Für weitere Informationen siehe Thema Ordner “Demos” in Kapitel 1 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Erstellen des Modells

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *car_insurance_claims.sav* im Ordner *Demos* verweist.

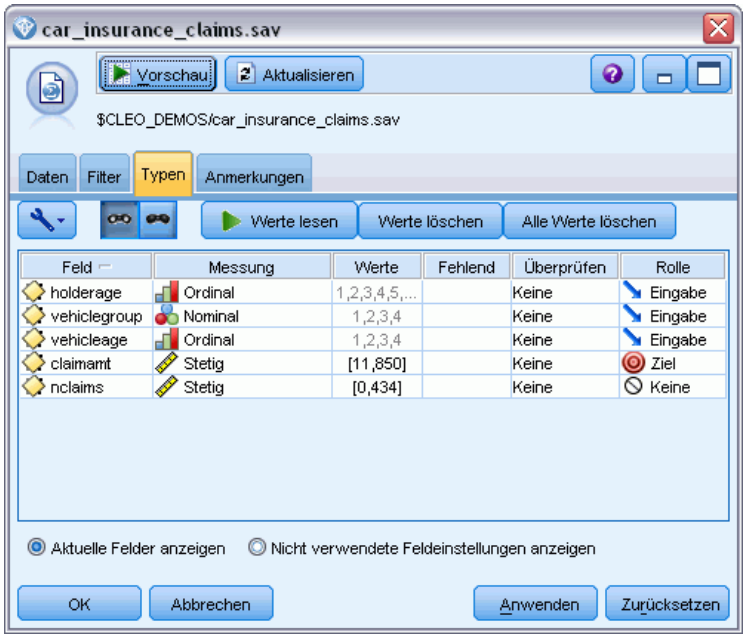
Abbildung 25-1
Beispiel-Stream für die Vorhersage von Schadensansprüchen für Autos



- Setzen Sie auf der Registerkarte “Typen” des Quellenknotens die Rolle für das Feld *claimamt* auf Ziel. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

▶ Klicken Sie auf Werte lesen, um die Daten zu instanziieren.

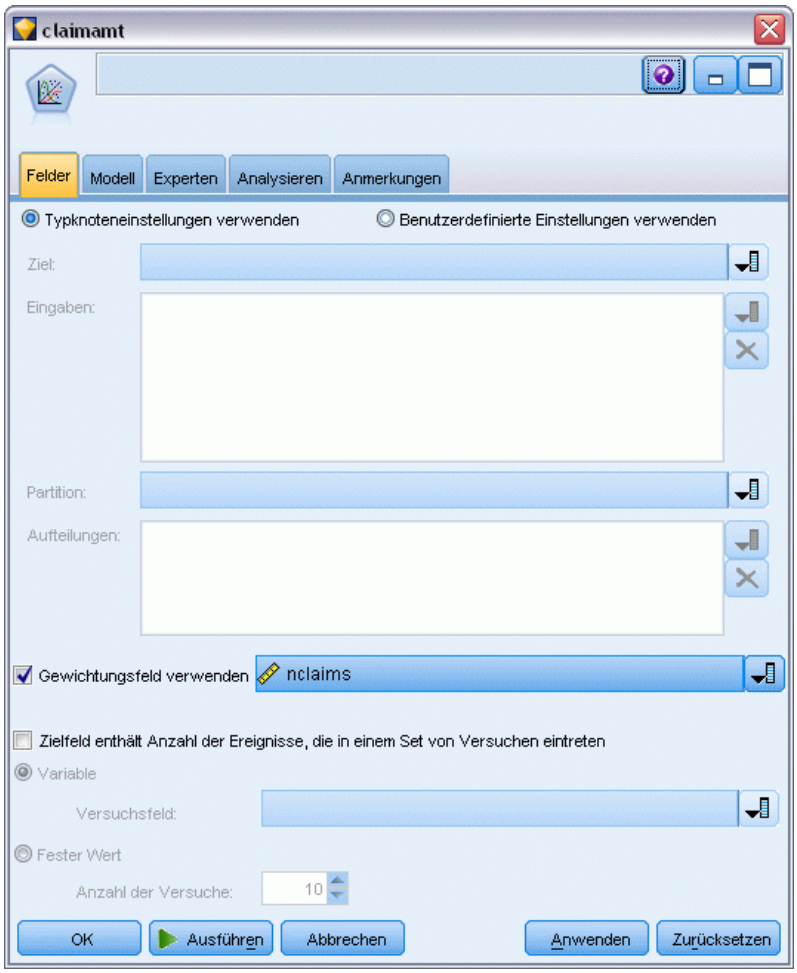
Abbildung 25-2
Festlegen der Feldrolle



▶ Fügen Sie dem Quellenknoten einen Genlin-Knoten hinzu. Klicken Sie im Genlin-Knoten auf die Registerkarte “Felder”.

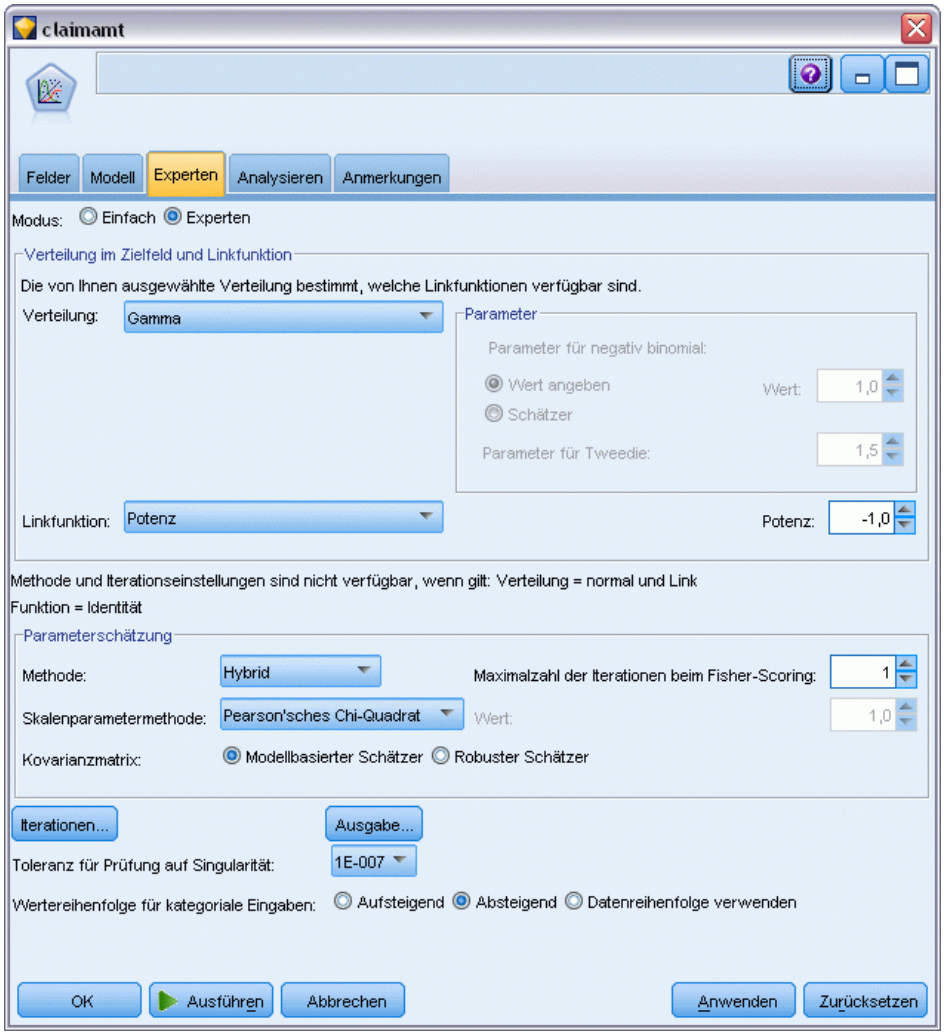
▶ Wählen Sie *nclaims* als Skalengewichtungsfeld aus.

Abbildung 25-3
Auswählen von Feldoptionen



- Klicken Sie auf die Registerkarte “Experten” und wählen Sie Experten aus, um die Expertenmodellierungsoptionen zu aktivieren.

Abbildung 25-4
Auswählen von Expertenoptionen



- Wählen Sie Gamma als Antwortverteilung.

- Wählen Sie Potenz als Verknüpfungsfunktion und geben Sie -1,0 als Exponenten der Potenzfunktion ein. Dies ist eine inverse Verknüpfung.

- Wählen Sie Pearson-Chi-Quadrat als Methode zur Schätzung des Skalenparameters. Diese Methode wird von McCullagh und Nelder verwendet. Wir folgen ihrer Methode, um ihre Ergebnisse zu reproduzieren.

- Wählen Sie Absteigend als Reiheinfolge der Kategorien für Faktoren. Dadurch wird angegeben, dass die erste Kategorie jedes Faktors als dessen Referenzkategorie dient; der Effekt dieser Auswahl auf das Modell besteht in der Interpretation von Parameterschätzern.

- Klicken Sie auf Ausführen, um das Modell-Nugget zu erstellen; dieses wird dem Stream-Zeichenbereich und der Modellpalette in der rechten oberen Ecke hinzugefügt. Zum Anzeigen der Modelldetails klicken Sie mit der rechten Maustaste auf das Modell-Nugget und wählen Bearbeiten oder Durchsuchen. Anschließend wählen Sie die Registerkarte "Erweitert" aus.

Parameter-Schätzer

Abbildung 25-5
Parameterschätzer

Parameter	RegressionskoeffizientB	Standardfehler	95% Wald-Konfidenzintervall		Hypothesentest		
			Unterer Wert	Oberer Wert	Wald-Chi-Quadrat	df	Sig.
(Konstanter Term)	,003	,0004	,003	,004	66,593	1	,000
[holderage=8]	,001	,0004	,000	,002	4,898	1	,027
[holderage=7]	,001	,0004	,000	,002	5,046	1	,025
[holderage=6]	,001	,0004	,000	,002	5,740	1	,017
[holderage=5]	,001	,0004	,001	,002	10,682	1	,001
[holderage=4]	,000	,0004	,000	,001	1,268	1	,260
[holderage=3]	,000	,0004	,000	,001	,720	1	,396
[holderage=2]	,000	,0004	-,001	,001	,054	1	,816
[holderage=1]	0(a)	.	.	.	.	.	.
[vehiclegroup=4]	-,001	,0002	-,002	-,001	61,883	1	,000
[vehiclegroup=3]	-,001	,0002	-,001	,000	13,039	1	,000
[vehiclegroup=2]	3,77E-005	,0002	,000	,000	,050	1	,823
[vehiclegroup=1]	0(a)	.	.	.	.	.	.
[vehicleage=4]	,004	,0004	,003	,005	88,175	1	,000
[vehicleage=3]	,002	,0002	,001	,002	53,013	1	,000
[vehicleage=2]	,000	,0001	,000	,001	13,191	1	,000
[vehicleage=1]	0(a)	.	.	.	.	.	.
(Skala)	1,209(b)						

Abhängige Variable: Average cost of claimsModell: (Konstanter Term), holderage, vehiclegroup, vehicleage

a. Auf 0 gesetzt, da dieser Parameter redundant ist.

b. Anhand des Pearson-Chi-Quadrats berechnet.

Der Omnibus-Test und Tests der Modelleffekte (nicht dargestellt) zeigen, dass das Modell besser als das Null-Modell funktioniert und dass alle Haupteffektterme einen Beitrag zum Modell leisten. Die Tabelle mit den Parameterschätzungen zeigt dieselben Werte an, die auch McCullagh und Nelder für die Faktorenebenen und den Skalierungsparameter erhalten haben.

Zusammenfassung

Mit verallgemeinerten linearen Modellen haben Sie Gammaregression an die Forderungsdaten angepasst. Beachten Sie: In diesem Modell wurde zwar die die kanonische Linkfunktion für die Gamma-Verteilung verwendet, eine Log-Verknüpfung würde jedoch ebenfalls zu brauchbaren Ergebnissen führen. Im Allgemeinen ist es schwierig bis unmöglich, Modelle mit unterschiedlichen Verknüpfungsfunktionen direkt zu vergleichen. Die Log-Verknüpfung ist jedoch ein Sonderfall der Potenzverknüpfung, wobei der Exponent 0 ist. Auf diese Weise können Sie die Abweichungen eines Modells mit einer Log-Verknüpfung und eines Modells mit einer Potenzverknüpfung vergleichen, um festzustellen, welches Modell zur besseren Anpassung führt (siehe z. B. Abschnitt 11.3 bei McCullagh und Nelder).

Erläuterungen der mathematischen Grundlagen der in IBM® SPSS® Modeler verwendeten Modellierungsverfahren sind im *SPSS Modeler-Algorithmushandbuch* aufgeführt.

Kapitel
26

Klassifikation von Zellproben (SVM)

Support Vector Machine (SVM) ist ein Klassifikations- und Regressionsverfahren, das sich besonders für umfangreiche Daten-Sets eignet. Umfangreiche Daten-Sets sind Daten-Sets mit einer großen Anzahl an Prädiktoren, wie sie beispielsweise im Bereich der Bioinformatik (der Anwendung der Informationstechnologie auf biochemische und biologische Daten) zu finden sind.

Ein medizinischer Forscher hat ein Daten-Set 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 dieser Zelleneigenschaften in Proben von anderen Patienten verwenden kann, um eine Frühindikation dafür abzugeben, ob die Proben vermutlich gutartig oder bösartig sind.

Für dieses Beispiel wird der Stream *svm_cancer.str* verwendet, der im Ordner *Demos* unter dem Unterordner *streams* verfügbar ist. Die verwendete Datendatei ist *cell_samples.data*. Für weitere Informationen siehe Thema Ordner "Demos" in Kapitel 1 in *IBM SPSS Modeler 14.2-Benutzerhandbuch*.

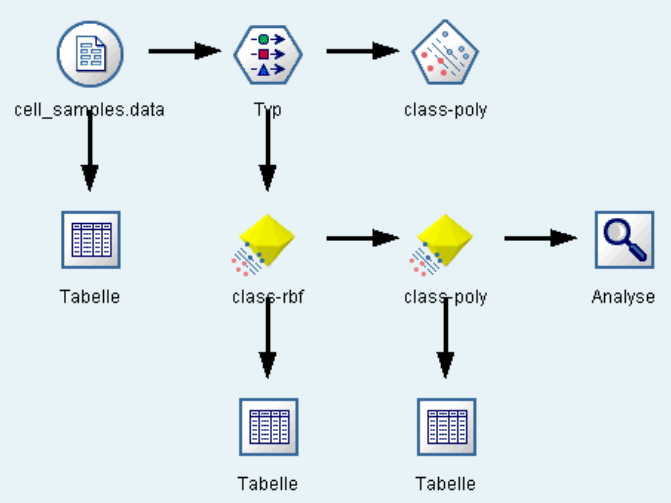
Das Beispiel beruht auf einem Daten-Set das im UCI Machine Learning Repository öffentlich zugänglich ist (Asuncion als auch Newman, 2007). Das Daten-Set besteht aus mehreren Datensätzen zu Proben menschlicher Zellen, die jeweils die Werte eines Sets von Zelleneigenschaften enthalten. Die Datensätze enthalten jeweils folgende Felder:

Feldname	Beschreibung
ID	Patienten-ID
Clump	Klumpendicke
UnifSize	Einheitlichkeit der Zellgröße
UnifShape	Einheitlichkeit der Zellform
MargAdh	Randhaftung
SingEpiSize	Größe einzelner Epithelzellen
BareNuc	Nackte Zellkerne
BlandChrom	Homogenes Chromatin
NormNucl	Normale Kernkörperchen
Mit	Mitose
Class	Gutartig oder bösartig

Für dieses Beispiel verwenden wir ein Daten-Set mit einer relativ kleinen Anzahl an Prädiktoren in jedem Datensatz.

Erstellen des Streams

Abbildung 26-1
Beispiel-Stream zur Veranschaulichung der SVM-Modellierung



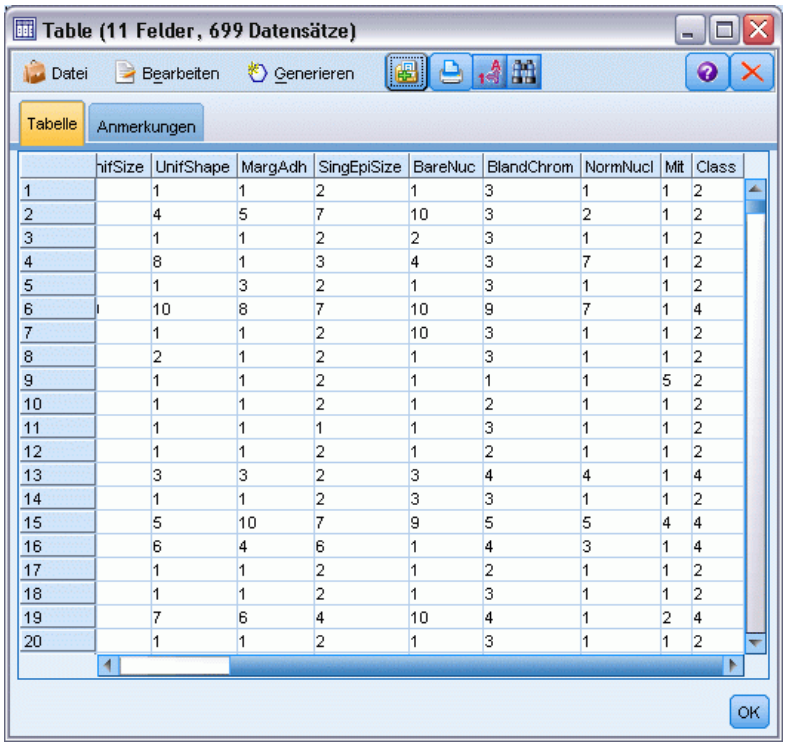
- Erstellen Sie einen neuen Stream und fügen Sie einen Quellenknoten für variable Dateien hinzu, der auf die Datei *cell_samples.data* im Verzeichnis *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist.

  Betrachten wir die Daten in der Quelldatei.

- Fügen Sie einen Tabellenknoten zum Stream hinzu.

- Gliedern Sie den Tabellenknoten mit dem Knoten für variable Dateien an und führen Sie den Stream aus.

Abbildung 26-2
Quellendaten für SVM



	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

Das Feld *ID* enthält die Patienten-IDs. Die Eigenschaften der Zellproben der einzelnen Patienten sind in den Feldern *Clump* bis *Mit* dokumentiert. Für die Werte gibt es die Stufen 1 bis 10, wobei 1 am meisten für Gutartigkeit spricht.

Das Feld *Class* enthält die Diagnose, die durch gesonderte medizinische Verfahren bestätigt wurde und angibt, ob die Proben gutartig (Wert = 2) oder bösartig (Wert = 4) sind.

Abbildung 26-3
Typknoteneinstellungen

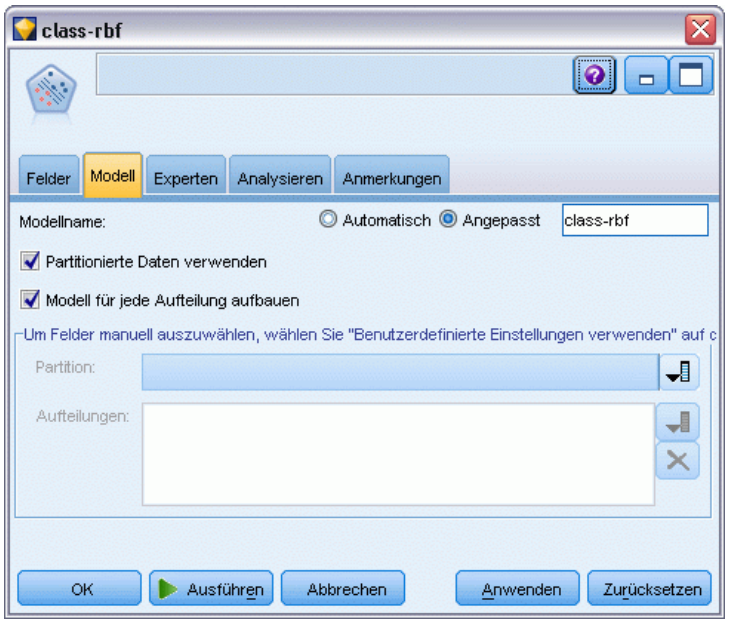
- Fügen Sie einen Typknoten hinzu und gliedern Sie ihn an den Knoten für variable Dateien an.
- Öffnen Sie den Typknoten.

  Das Modell soll den Wert von *Class* vorhersagen, also gutartig (=2) oder bösartig (=4)). Da dieses Feld nur zwei mögliche Werte annehmen kann, müssen wir sein Messniveau entsprechend ändern.

- Klicken Sie in der Spalte Messung für das Feld *Class* (das letzte in der Liste) auf den Wert Stetig und ändern Sie ihn in Flag.
- Klicken Sie auf Werte lesen.
- Legen Sie in der Spalte Rolle als Rolle für *ID* (Patienten-ID) Keine fest, da diese Variable nicht als Prädiktor oder Ziel für das Modell verwendet werden soll.
- Legen Sie als Rolle für das Ziel *Class* den Wert Ziel fest und behalten Sie für alle anderen Felder (die Prädiktoren) den Wert Eingabe bei.
- Klicken Sie auf OK.

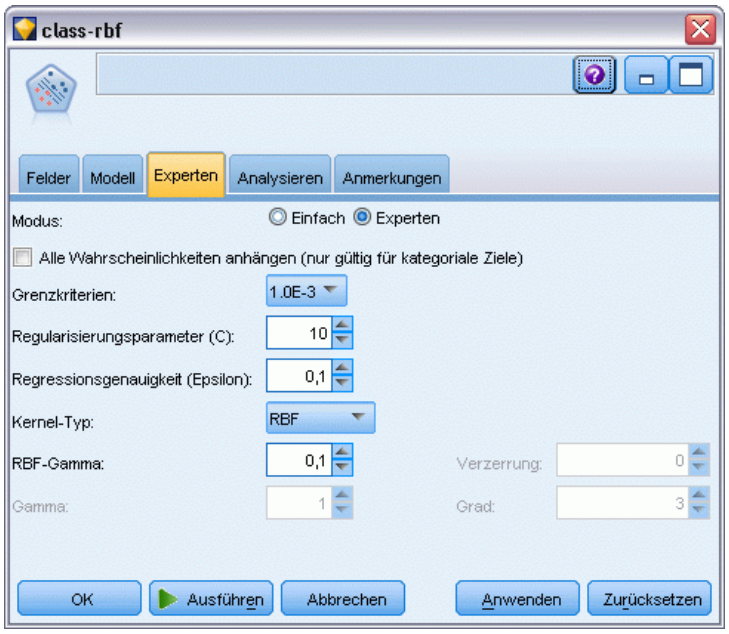
  Der SVM-Knoten bietet eine Auswahl an Kernel-Funktionen zur Durchführung der Verarbeitung. Da sich nicht im Voraus sagen lässt, welche Funktion mit einem bestimmten Daten-Set am besten funktioniert, wählen wir abwechselnd verschiedene Funktionen aus und vergleichen die Ergebnisse. Beginnen wir mit der Standardvorgabe, “RBF” (Radial Basis Function).

Abbildung 26-4
Einstellungen auf der Registerkarte "Modell"



- Gliedern Sie auf der Modellierungspalette einen SVM-Knoten an den Typknoten an.
- Öffnen Sie den SVM-Knoten. Klicken Sie auf der Registerkarte Modell unter Modellname auf die Option Angepasst und geben Sie in das angrenzende Textfeld den Ausdruck *class-rbf* ein.

Abbildung 26-5
Registerkarte "Experten" – Standardeinstellungen



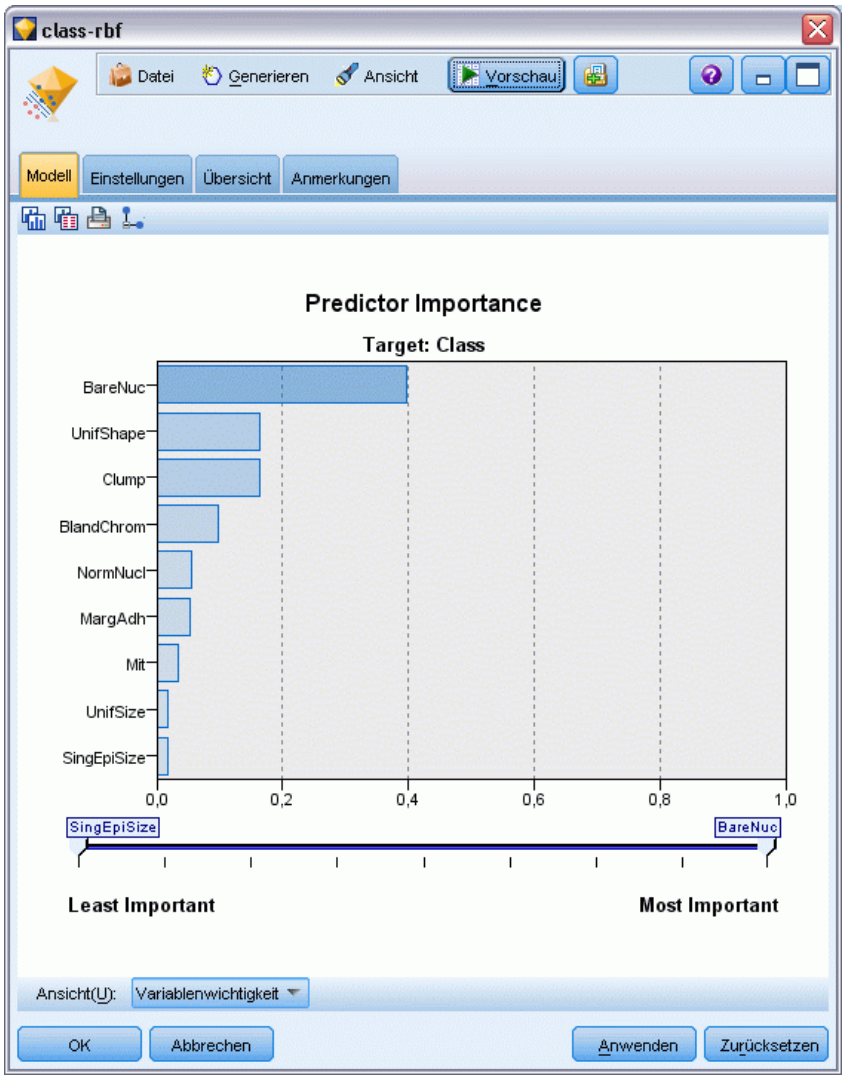
- Setzen Sie auf der Registerkarte Experten den Wert von Modus auf Experten, um eine bessere Lesbarkeit zu erzielen. Belassen Sie aber alle Standardoptionen wie vorhanden. Beachten Sie, dass Kernel-Typ standardmäßig auf RBF eingestellt ist. Im Modus "Einfach" sind alle Optionen abgeblendet.

Abbildung 26-6
Einstellungen auf der Registerkarte "Analysieren"

- Aktivieren Sie auf der Registerkarte Analysieren das Kontrollkästchen Bedeutsamkeit der Variablen berechnen.
- Klicken Sie auf Ausführen. Das Modell-Nugget wird in den Stream und in der Modellpalette oben rechts im Fenster platziert.
- Doppelklicken Sie auf das Modell-Nugget im Stream.

Untersuchen der Daten

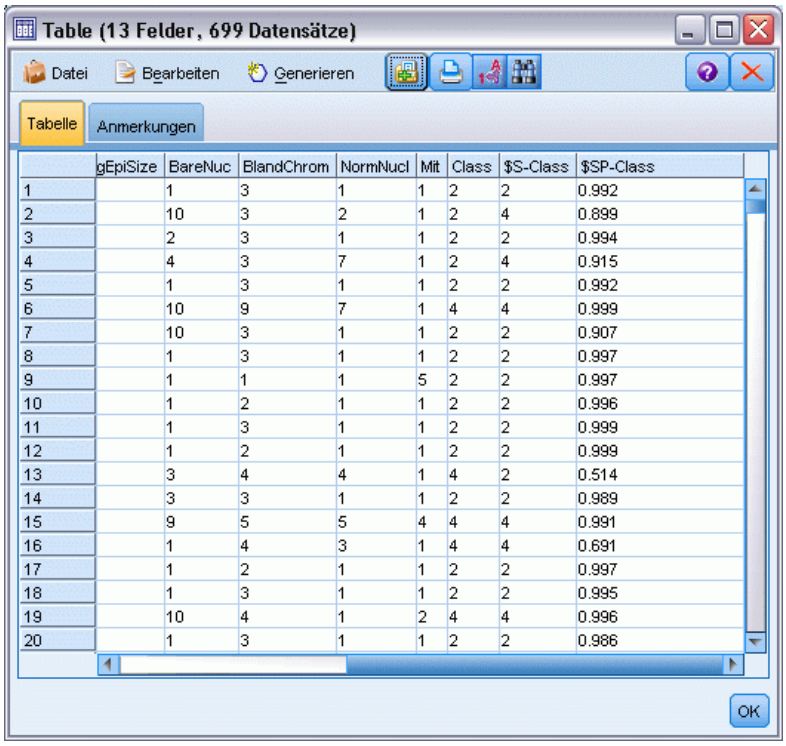
Abbildung 26-7
Diagramm für die Bedeutsamkeit des Prädiktors



Auf der Registerkarte “Modell” zeigt das Diagramm für die Bedeutsamkeit des Prädiktors den relativen Effekt der verschiedenen Felder in der Vorhersage an. Dies zeigt uns, dass *BareNuc* bei weitem den größten Effekt hat und *UnifShape* sowie *Clump* ebenfalls recht signifikant sind.

- Klicken Sie auf OK.
- Gliedern Sie einen Tabellenknoten an das Modell-Nugget *class-rbf* an.
- Öffnen Sie den Tabellenknoten und klicken Sie auf Ausführen.

Abbildung 26-8
Für Vorhersage und Konfidenzwert hinzugefügte Felder



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

▶ Das Modell hat zwei zusätzliche Felder erstellt. Führen Sie für die Tabellenausgabe einen Bildlauf nach rechts durch, um sie anzuzeigen:

Neuer Feldname	**Beschreibung**
$S-Class	Der vom Modell für *Class* vorhergesagte Wert.
$SP-Class	Neigungs-Score für diese Vorhersage (die Likelihood, dass diese Vorhersage wahr ist, ein Wert im Bereich von 0,0 bis 1,0)

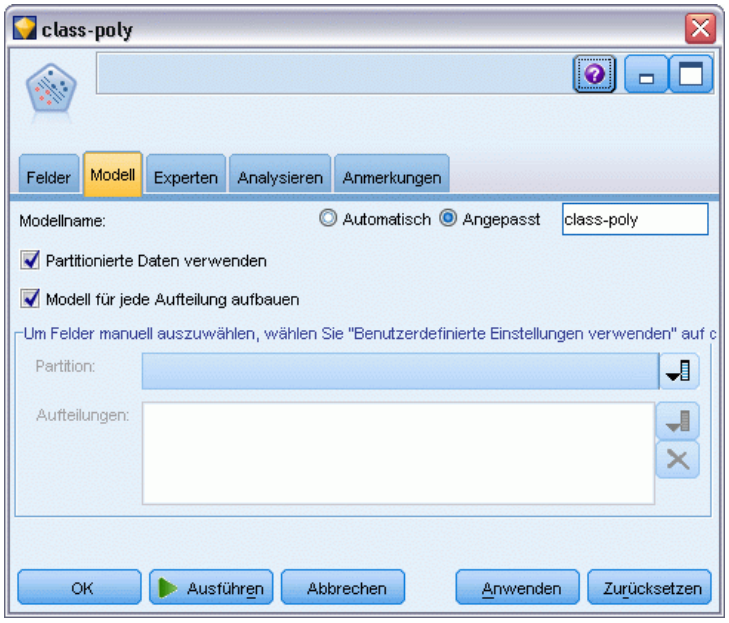
Allein durch Betrachtung der Tabelle können wir sehen, dass die Neigungsscores (in der Spalte *$SP-Class*) für die meisten Datensätze relativ hoch sind.

Es gibt jedoch einige bedeutsame Ausnahmen, beispielsweise den Datensatz für Patient 1041801 in Zeile 13, bei dem der Wert 0,514 unannehmbar niedrig ist. Beim Vergleich zwischen *Class* und *$S-Class* wird ebenfalls deutlich, dass dieses Modell eine Reihe von falschen Vorhersagen erstellt hat, auch wenn der Neigungs-Score relativ hoch war (z. B. Zeilen 2 und 4).

Untersuchen wir, ob sich durch Auswahl eines anderen Funktionstyps ein besseres Ergebnis erzielen lässt.

Versuch mit einer anderen Funktion

Abbildung 26-9
Festlegung eines neuen Namens für das Modell



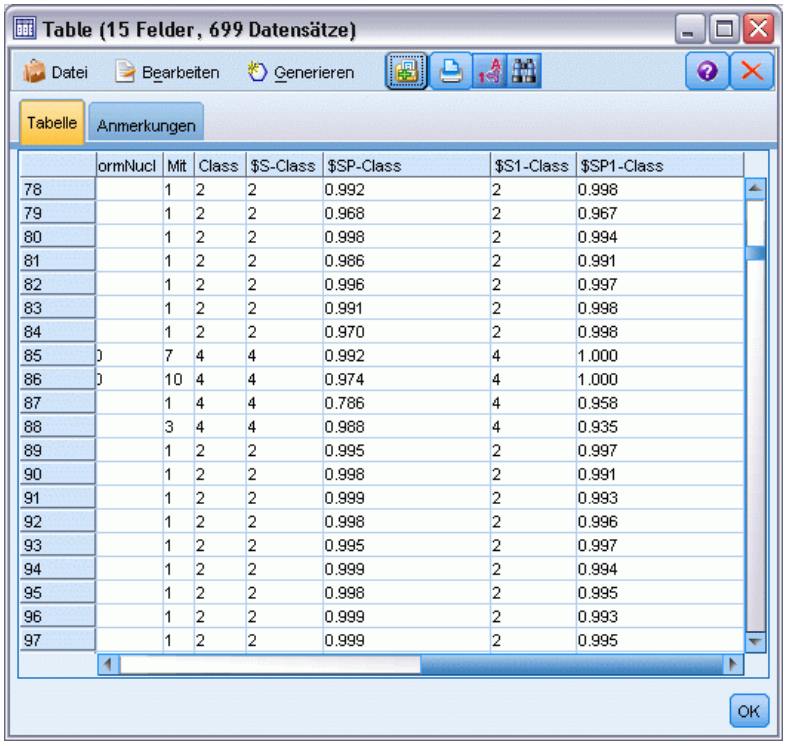
- Schließen Sie das Tabellenausgabefenster.
- Hängen Sie einen SVM-Modellierungsknoten an den Typknoten an.
- Öffnen Sie den neuen SVM-Knoten.
- Wählen Sie in der Registerkarte Modell die Option “Benutzerdefiniert” und geben Sie *class-poly* als Modellnamen ein.

Abbildung 26-10
Registerkarte "Experten" – Einstellungen für den "Polynomial"

- Setzen Sie auf der Registerkarte Experten den Wert von Modus auf Experten.
- Setzen Sie Kernel-Typ auf Polynomial und klicken Sie auf Ausführen. Das Modell-Nugget *class-poly* wird in den Stream und in der Modellpalette oben rechts im Fenster platziert.
- Verbinden Sie das Modell-Nugget *class-rbf* mit dem Modell-Nugget *class-poly* (wählen Sie Ersetzen im Warnungsdialog).
- Gliedern Sie einen Tabellenknoten an das Nugget *class-poly* an.
- Öffnen Sie den Tabellenknoten und klicken Sie auf Ausführen.

Vergleichen der Ergebnisse

Abbildung 26-11
Für Polynomialfunktion hinzugefügte Felder



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
78		1	2	2	0.992	2	0.998
79		1	2	2	0.968	2	0.967
80		1	2	2	0.998	2	0.994
81		1	2	2	0.986	2	0.991
82		1	2	2	0.996	2	0.997
83		1	2	2	0.991	2	0.998
84		1	2	2	0.970	2	0.998
85	)	7	4	4	0.992	4	1.000
86	)	10	4	4	0.974	4	1.000
87		1	4	4	0.786	4	0.958
88		3	4	4	0.988	4	0.935
89		1	2	2	0.995	2	0.997
90		1	2	2	0.998	2	0.991
91		1	2	2	0.999	2	0.993
92		1	2	2	0.998	2	0.996
93		1	2	2	0.995	2	0.997
94		1	2	2	0.999	2	0.994
95		1	2	2	0.998	2	0.995
96		1	2	2	0.999	2	0.993
97		1	2	2	0.999	2	0.995

- Führen Sie für die Tabellenausgabe einen Bildlauf nach rechts durch, um die neu hinzugefügten Felder anzuzeigen.

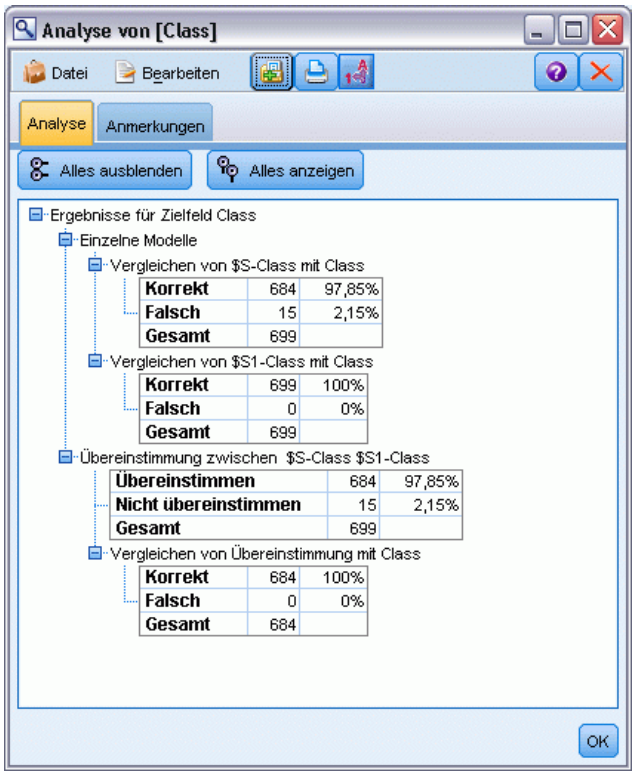
  Die generierten Felder für den Funktionstyp “Polinomial” lauten *$S1-Class* und *$SP1-Class*.

  Die Ergebnisse für “Polynomial” sehen deutlich besser aus. Viele der Neigungs-Scores liegen bei 0,995 oder höher, was sehr ermutigend ist.

- Um die Verbesserung des Modells zu bestätigen, gliedern Sie einen Analyseknoten an das Modell-Nugget *class-poly* an.

  Öffnen Sie den Analyseknoten und klicken Sie auf Ausführen.

Abbildung 26-12
Analyseknoten



Durch dieses Verfahren mit dem Analyseknoten können Sie zwei oder mehr Modell-Nuggets desselben Typs vergleichen. Die Ausgabe aus dem Analyseknoten zeigt, dass die Funktion "RBF" 97,85 % der Fälle korrekt vorhersagt, was noch recht gut ist. Die Ausgabe zeigt jedoch, dass die Polynomialfunktion die Diagnose in jedem einzelnen Fall korrekt vorhersagte. In der Praxis werden Sie kaum eine 100%ige Genauigkeit erreichen, Sie können jedoch mithilfe des Analyseknotens bestimmen, ob das Modell für Ihre spezielle Anwendung über eine ausreichende Genauigkeit verfügt.

Tatsächlich erbringt auch keiner der beiden anderen Funktionstypen ("Sigmoid" und "Linear") für dieses konkrete Daten-Set eine so gute Leistung wie die Funktion "Polynomial" Bei einem anderen Daten-Set könnten die Ergebnisse jedoch durchaus anders sein, sodass es sich immer lohnt, das gesamte Optionsspektrum auszuschöpfen.

Zusammenfassung

Sie haben mithilfe verschiedener Typen von SVM-Kernel-Funktionen eine Klassifikation aus einer Reihe von Attributen vorhergesagt. Sie haben gesehen, dass unterschiedliche Kernel bei demselben Daten-Set zu unterschiedlichen Ergebnissen führen, und festgestellt, wie Sie die Qualitätsunterschiede zwischen den Modellen messen können.

Kapitel
27

Verwenden der Cox-Regression zur Modellierung der Zeit bis zur Kundenabwanderung

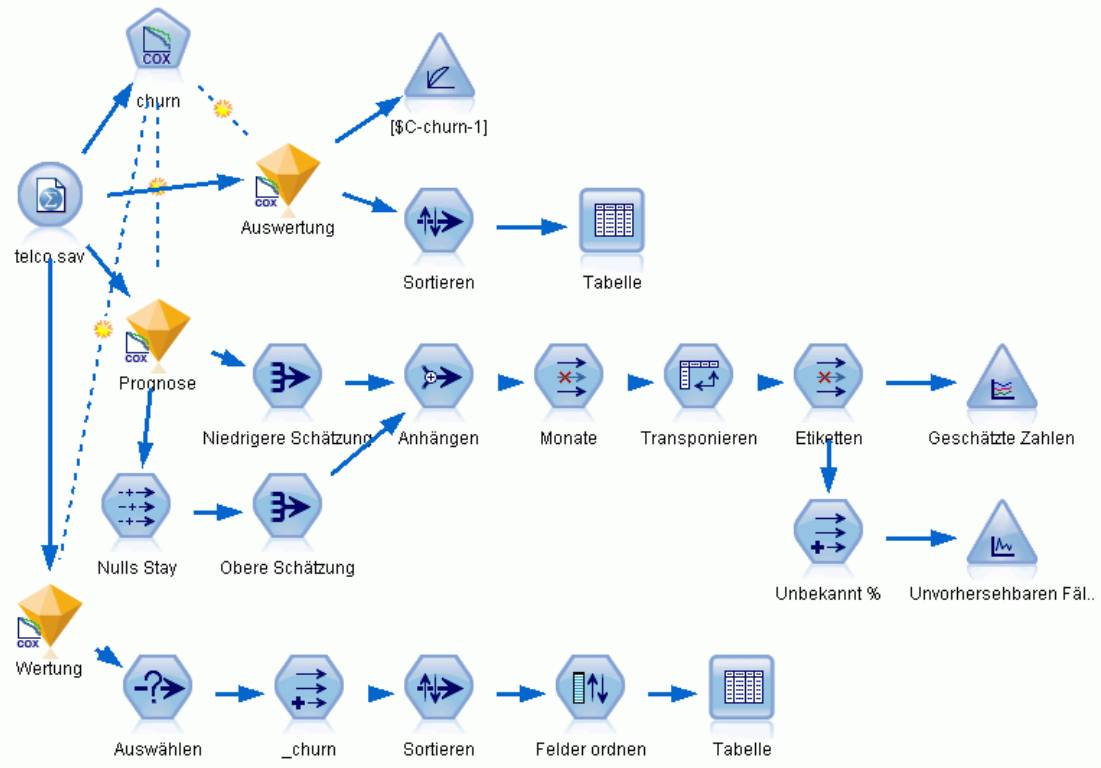
Im Rahmen seiner Bemühungen zur Reduzierung der Kundenabwanderung ist ein Telekommunikationsunternehmen daran interessiert, die "Zeit bis zur Abwanderung" zu modellieren, um die Faktoren zu ermitteln, die für Kunden gelten, die rasch zu einem anderen Dienst wechseln. Dazu wird eine Zufallsstichprobe der Kunden ausgewählt und die Dauer des Kundenverhältnisses, ob sie noch immer aktive Kunden sind und verschiedene andere Felder werden aus der Datenbank gezogen.

In diesem Beispiel wird der Stream *telco_coxreg.str* verwendet, der sich auf die Datendatei *telco.sav* bezieht. Die Datendatei befindet sich im Ordner *Demos* und die Stream-Datei im Unterordner *streams*. Für weitere Informationen siehe Thema Ordner "Demos" in Kapitel 1 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Erstellen eines geeigneten Modells

- Fügen Sie einen Statistikdatei-Quellenknoten hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

Abbildung 27-1
Beispielstream zur Analyse der Zeit bis zur Abwanderung



- Schließen Sie auf der Registerkarte Filter des Quellenknotens die Felder *region*, *income*, *longten* bis *wireten* und *loglong* bis *logwire* aus.

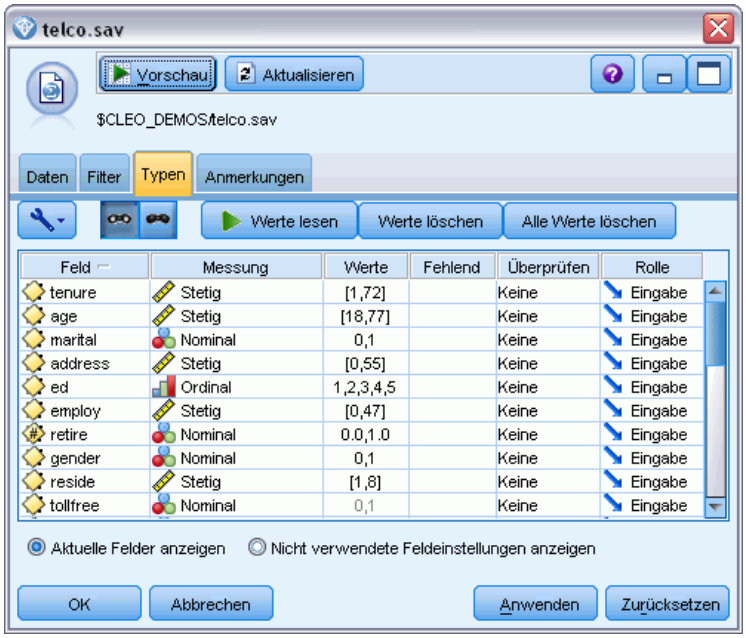
Abbildung 27-2
Filtern nicht benötigter Felder

(Alternativ können Sie die Rolle für diese Felder auf der Registerkarte “Typen” in Keine ändern, anstatt sie auszuschließen, oder die gewünschten Felder im Modellierungsknoten auswählen.)

- Setzen Sie auf der Registerkarte “Typen” des Quellenknotens die Rolle für das Feld *churn* auf Ziel und setzen Sie das Messniveau auf Flag. Für alle anderen Felder sollte als Rolle Eingabe festgelegt sein.

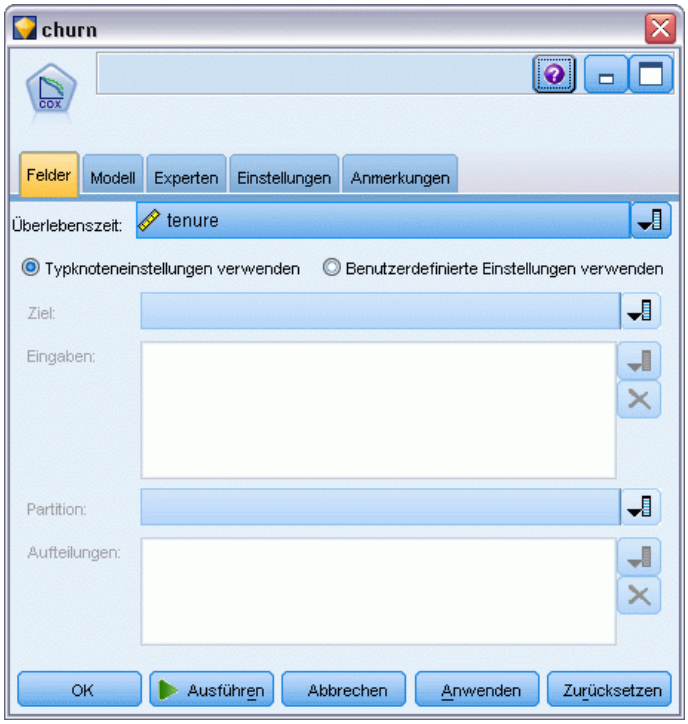
▶ Klicken Sie auf Werte lesen, um die Daten zu instanziieren.

Abbildung 27-3
Festlegen der Feldrolle



- Gliedern Sie einen Cox-Knoten an den Quellenknoten an; wählen Sie auf der Registerkarte Felder den Eintrag *tenure* als Variable für die Überlebenszeit aus.

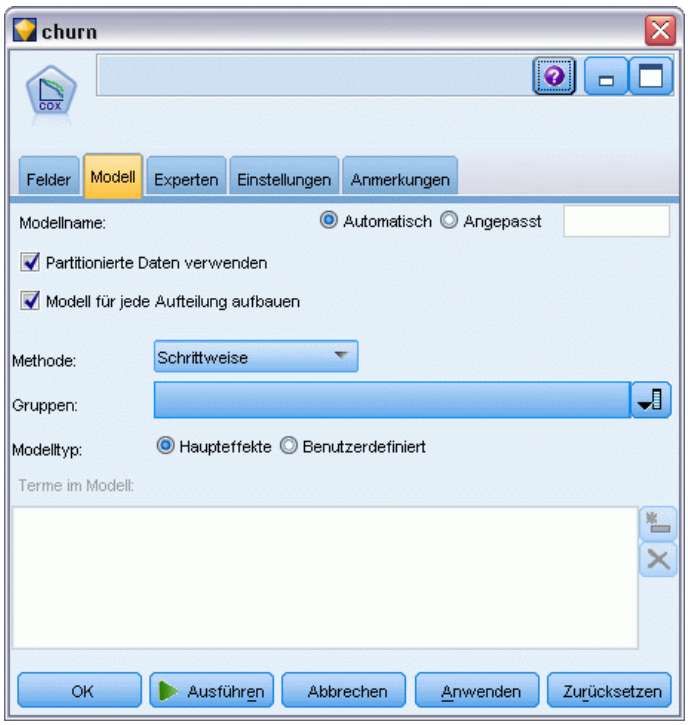
Abbildung 27-4
Auswählen von Feldoptionen



- Klicken Sie auf die Registerkarte Modell.

▶ Wählen Sie Schrittweise als Methode für die Auswahl der Variablen aus.

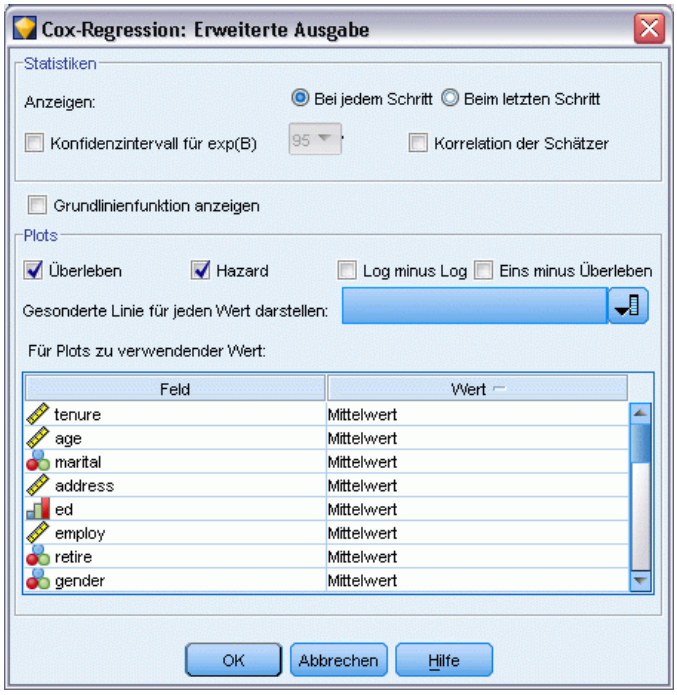
Abbildung 27-5
Auswählen Modelloptionen



▶ Klicken Sie auf die Registerkarte Experten und wählen Sie Experten, um die Expertenmodellierungsoptionen zu aktivieren.

▶ Klicken Sie auf Ausgabe.

Abbildung 27-6
Auswahl der erweiterten Ausgabeoptionen



▶ Wählen Sie die Optionen Überleben und Hazard als zu erstellende Diagramme aus und klicken Sie dann auf OK.

▶ Klicken Sie auf Ausführen, um das Modell-Nugget zu erstellen; dieses wird dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die zugehörigen Details anzuzeigen, können Sie auf das Nugget im Stream doppelklicken. Betrachten Sie zunächst die Registerkarte “Erweiterte Ausgabe”.

Zensierte Fälle

Abbildung 27-7
Zusammenfassung der Fallverarbeitung

		N	Prozent
Für Analyse verfügbare Fälle	**Ereignis(a)**	274	27,4%
	Zensiert	726	72,6%
	Insgesamt	1000	100,0%
Nicht verwendete Fälle	**Fälle mit fehlenden Werten**	0	,0%
	Fälle mit negativer Zeit	0	,0%
	Zensierte Fälle vor dem frühesten Ereignis in einer Schicht	0	,0%
	Insgesamt	0	,0%
Insgesamt		1000	100,0%

a. Abhängige Variable: Months with service

Die Statusvariable gibt an, ob das Ereignis für einen bestimmten Fall eingetreten ist. Wenn das Ereignis nicht eingetreten ist, spricht man von einem “zensierten” Fall. Zensierte Fälle werden bei der Berechnung der Regressionskoeffizienten nicht verwendet, aber zur Berechnung der Basis-Hazard-Rate. Die Auswertung der Fallverarbeitung zeigt, dass 726 Fälle zensiert wurden. Hierbei handelt es sich um die Kunden, die nicht abgewandert sind.

Kodierungen für kategoriale Variablen

Abbildung 27-8
Categorical variable codings

		Häufigkeit	(1)(s)	(2)	(3)	(4)
marital(t)	0=Unmarried	505	1			
	1=Married	495	0			
ed(t)	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire(t)	0=No	953	1			
	1=Yes	47	0			
gender(t)	0=Male	483	1			
	1=Female	517	0			
tollfree(t)	0=No	526	1			
	1=Yes	474	0			
equip(t)	0=No	614	1			
	1=Yes	386	0			
callcard(t)	0=No	322	1			
	1=Yes	678	0			
wireless(t)	0=No	704	1			
	1=Yes	296	0			
multline(t)	0=No	525	1			
	1=Yes	475	0			
voice(t)	0=No	696	1			
	1=Yes	304	0			
pager(t)	0=No	739	1			
	1=Yes	261	0			
internet(t)	0=No	632	1			
	1=Yes	368	0			
callid(t)	0=No	519	1			
	1=Yes	481	0			
callwait(t)	0=No	515	1			
	1=Yes	485	0			
forward(t)	0=No	507	1			
	1=Yes	493	0			
confer(t)	0=No	498	1			
	1=Yes	502	0			
ebill(t)	0=No	629	1			
	1=Yes	371	0			
custcat(t)	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	

Status *Married* (Verheiratet) zwar möglicherweise den Variablenwert “1” in der Datendatei auf, zum Zwecke der Regression werden sie jedoch als “0” kodiert.

Variablenauswahl

Abbildung 27-9
Omnibus-Tests

Schritt	-2 Log-Likelihood	Gesamt (Wert)			Änderung aus vorangegangenem Schritt			Änderung aus vorangegangenem Block		
		Chi-Quadrat	df	Signifikanz	Chi-Quadrat	df	Signifikanz	Chi-Quadrat	df	Signifikanz
1(c)	3392,536	162,303	1	,000	133,828	1	,000	133,828	1	,000
2(d)	3087,314	249,392	2	,000	305,222	1	,000	439,050	2	,000
3(e)	3027,085	328,426	3	,000	60,229	1	,000	499,279	3	,000
4(f)	2990,790	347,197	4	,000	36,294	1	,000	535,574	4	,000
5(g)	2973,790	362,673	5	,000	17,000	1	,000	552,574	5	,000
6(h)	2958,796	376,140	6	,000	14,994	1	,000	567,568	6	,000
7(i)	2945,503	384,717	7	,000	13,293	1	,000	580,861	7	,000
8(j)	2936,993	417,341	8	,000	8,510	1	,004	589,371	8	,000
9(k)	2926,000	423,911	9	,000	10,994	1	,001	600,364	9	,000
10(l)	2917,551	428,078	10	,000	8,449	1	,004	608,813	10	,000
11(m)	2913,308	436,837	11	,000	4,243	1	,039	613,056	11	,000
12(n)	2908,078	440,158	12	,000	5,230	1	,022	618,286	12	,000

a. Anfangsblocknummer 0, anfängliche Log-Likelihood-Funktion: -2 Log-Likelihood: 3526,364

b. Beginnen mit Block-Nr. 1. Methode = Vorwärts schrittweise (Likelihood Ratio)

c. Variable(n) eingegeben in Schritt Nr. 1: callcard

d. Variable(n) eingegeben in Schritt Nr. 2: longmon

e. Variable(n) eingegeben in Schritt Nr. 3: equip

f. Variable(n) eingegeben in Schritt Nr. 4: employ

g. Variable(n) eingegeben in Schritt Nr. 5: multline

h. Variable(n) eingegeben in Schritt Nr. 6: voice

i. Variable(n) eingegeben in Schritt Nr. 7: address

j. Variable(n) eingegeben in Schritt Nr. 8: equipmon

k. Variable(n) eingegeben in Schritt Nr. 9: ebill

l. Variable(n) eingegeben in Schritt Nr. 10: callid

m. Variable(n) eingegeben in Schritt Nr. 11: internet

n. Variable(n) eingegeben in Schritt Nr. 12: reside

Beim Modellerstellungsprozess wird ein Algorithmus vom Typ “Schrittweise vorwärts” verwendet. Die Omnibus-Tests sind ein Maß für die Leistungsfähigkeit des Modells. Die Chi-Quadrat-Änderung seit dem vorherigen Schritt ist die Differenz zwischen der −2

Log-Likelihood des Modells im vorherigen Schritt und der im aktuellen Schritt. Falls der Schritt im Hinzufügen einer Variablen bestand, ist die Aufnahme sinnvoll, wenn die Signifikanz der Veränderung weniger als 0,05 beträgt. Falls der Schritt im Entfernen einer Variablen bestand, ist der Ausschluss sinnvoll, wenn die Signifikanz der Veränderung mehr als 0,10 beträgt. In zwölf Schritten werden zwölf Variablen zum Modell hinzugefügt.

Abbildung 27-10
Variablen in der Gleichung (nur Schritt 12)

		B	SE	Wald	df	Signifikanz	Exp(B)
Schritt 12	address	-,035	,009	14,543	1	,000	,966
	employ	-,051	,010	25,767	1	,000	,950
	reside	-,103	,046	5,037	1	,025	,902
	equip	-1,948	,381	26,180	1	,000	,143
	callcard	,777	,151	26,451	1	,000	2,175
	longmon	-,233	,022	115,619	1	,000	,792
	equipmon	-,042	,011	15,377	1	,000	,959
	multline	,612	,145	17,854	1	,000	1,844
	voice	-,501	,157	10,197	1	,001	,606
	internet	-,362	,160	5,114	1	,024	,697
	callid	-,464	,148	9,790	1	,002	,629
	ebill	-,399	,156	6,557	1	,010	,671

Das endgültige Modell enthält die Variablen *address*, *employ*, *reside*, *equip*, *callcard*, *longmon*, *equipmon*, *multline*, *voice*, *internet*, *callid* und *ebill*. Um einen Einblick in die Effekte der einzelnen Prädiktoren zu erhalten, betrachten wir den Wert Exp(B), der als vorhergesagte Änderung an der Hazard-Rate für einen Anstieg der Einheit im Prädiktor interpretiert werden kann.

- Der Wert von Exp(B) für *address* bedeutet, dass die Abwanderungs-Hazard-Rate sich für jedes Jahr, in dem Kunde dieselbe Adresse hatte, um 100 % −(100 %×0,966)=3,4 % verringert. Die Abwanderungs-Hazard-Rate für einen Kunden, der fünf Jahre lang an derselben Adresse lebte, verringert sich um $100\ \%-(100\ \%\times 0{,}966^5)=15{,}88\ \%$.
- Der Wert von Exp(B) für *callcard* bedeutet, dass die Abwanderungs-Hazard-Rate für einen Kunden, der nicht den Telefonkarten-Service abonniert hat, 2,175-mal so hoch ist wie die eines Kunden mit dem Service. Wir erinnern uns aus den Kodierungen für die kategorialen Variablen, dass für die Regression gilt: *No* (Nein) = 1.
- Der Wert von Exp(B) für *internet* bedeutet, dass die Abwanderungs-Hazard-Rate für einen Kunden, der nicht den Internet-Service abonniert hat, 0,697-mal so hoch ist wie die eines Kunden mit dem Service. Dies ist ein wenig beunruhigend, da es nahelegt, dass Kunden, die diesen Service abonniert haben, dem Unternehmen schneller den Rücken kehren als Kunden ohne diesen Service.

Abbildung 27-11
Nicht im Modell verwendete Variablen (nur Schritt 12)

Schritt 12	**age**	,122	1	,726
	marital	,648	1	,421
	ed	6,328	4	,176
	ed(1)	,007	1	,934
	ed(2)	,203	1	,652
	ed(3)	,835	1	,361
	ed(4)	5,773	1	,016
	retire	,013	1	,908
	gender	,214	1	,644
	tollfree	3,243	1	,072
	wireless	,668	1	,414
	tollmon	,000	1	,987
	cardmon	3,163	1	,075
	wiremon	1,084	1	,298
	pager	1,808	1	,179
	callwait	,266	1	,606
	forward	2,201	1	,138
	confer	2,568	1	,109
	lninc	2,853	1	,091
	custcat	,864	3	,834
	custcat(1)	,466	1	,495
	custcat(2)	,450	1	,502
	custcat(3)	,019	1	,889

Die nicht im Modell verwendeten Variablen weisen jeweils Score-Statistiken mit Signifikanzwerten von mehr als 0,05 auf. Die Signifikanzwerte für *tollfree* und *cardmon* sind zwar nicht niedriger als 0,05, aber zumindest nicht weit von diesem Wert entfernt. Es könnte sich lohnen, diese in weiteren Studien zu untersuchen.

Mittelwerte von Kovariaten

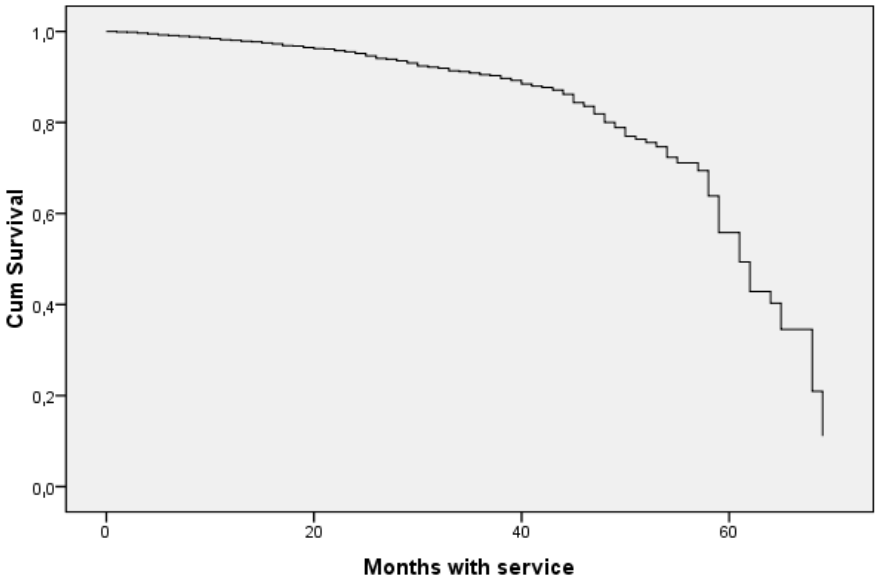
Abbildung 27-12
Mittelwerte von Kovariaten

	Mittelwert
age	41,684
marital	,505
address	11,551
ed(1)	,204
ed(2)	,287
ed(3)	,209
ed(4)	,234
employ	10,987
retire	,953
gender	,483
reside	2,331
tollfree	,526
equip	,614
callcard	,322
wireless	,704
longmon	11,723
tollmon	13,274
equipmon	14,220
cardmon	13,781
wiremon	11,584
multline	,525
voice	,696
pager	,739
internet	,632
callid	,519
callwait	,515
forward	,507
confer	,498
ebill	,629
lninc	3,957
custcat(1)	,266
custcat(2)	,217
custcat(3)	,281

Diese Tabelle zeigt den Durchschnittswert der einzelnen Prädiktorvariablen. Diese Tabelle ist eine nützliche Referenz bei der Untersuchung der Überlebensdiagramme, die für die Mittelwerte erstellt werden. Beachten Sie jedoch bei der Untersuchung der Mittelwerte der Indikatorvariablen für kategoriale Prädiktoren, dass es den "durchschnittlichen" Kunden in der Realität nicht gibt. Selbst mit allen metrischen Prädiktoren werden Sie schwerlich einen Kunden finden, dessen Kovariatenwerte alle nahe beim Mittelwert liegen. Wenn Sie die Überlebenskurve für einen bestimmten Fall anzeigen möchten, können Sie im Dialogfeld "Diagramme" die Kovariatenwerte ändern, die für die Darstellung der Überlebenskurve verwendet werden sollen. Wenn Sie die Überlebenskurve für einen bestimmten Fall anzeigen möchten, können Sie im Gruppenfeld "Diagramme" des Dialogfelds "Erweiterte Ausgabe" die Kovariatenwerte ändern, die für die Darstellung der Überlebenskurve verwendet werden sollen.

Überlebenskurve

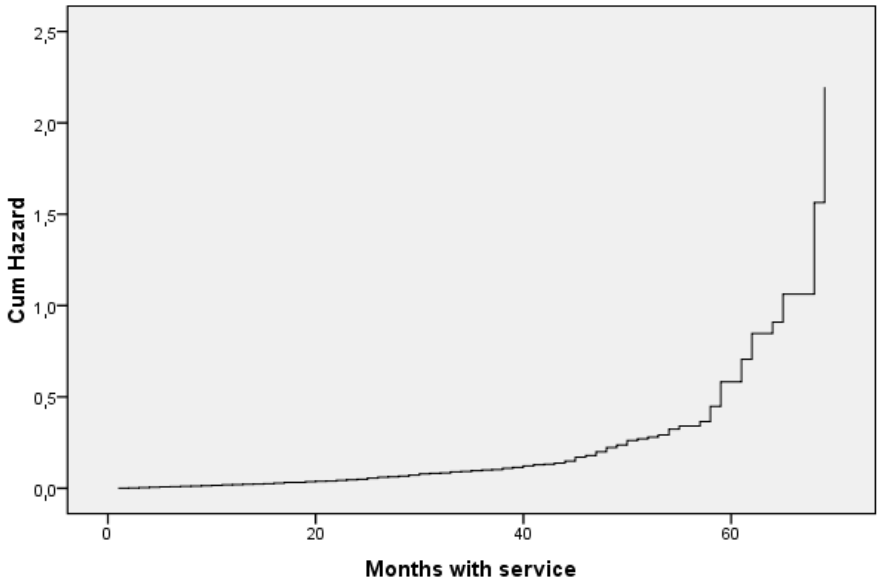
Abbildung 27-13
Überlebenskurve für den "durchschnittlichen" Kunden



Die einfache Überlebenskurve ist eine visuelle Anzeige der vom Modell vorhergesagten Zeit bis zur Abwanderung für den durchschnittlichen Kunden. Auf der horizontalen Achse wird die Zeit bis zum Eintreten des Ereignisses angezeigt. Auf der vertikalen Achse wird die Überlebenswahrscheinlichkeit angezeigt. So zeigt jeder Punkt auf der Überlebenskurve die Wahrscheinlichkeit an, dass der "durchschnittliche" Kunde nach diesem Zeitpunkt noch immer zum Kundenkreis gehört. Nach 55 Monaten wird die Überlebenskurve weniger gleichmäßig. Es gibt weniger Kunden, die so lange Zeit zum Kundenkreis des Unternehmens gehörten, sodass weniger Informationen zur Verfügung stehen. Dadurch wird die Kurve stufig.

Hazard-Kurve

Abbildung 27-14
Hazard-Kurve für den "durchschnittlichen" Kunden



Die einfache Hazard-Kurve ist eine visuelle Anzeige des kumulativen vom Modell vorhergesagten Abwanderungspotenzials für den "durchschnittlichen" Kunden. Auf der horizontalen Achse wird die Zeit bis zum Eintreten des Ereignisses angezeigt. Auf der vertikalen Achse wird die kumulative Hazard-Rate angezeigt, die gleich dem negativen Logarithmus der Überlebenswahrscheinlichkeit ist. Nach 55 Monaten wird die Hazard-Kurve (wie zuvor die Überlebenskurve und aus denselben Gründen) weniger gleichmäßig.

Evaluierung

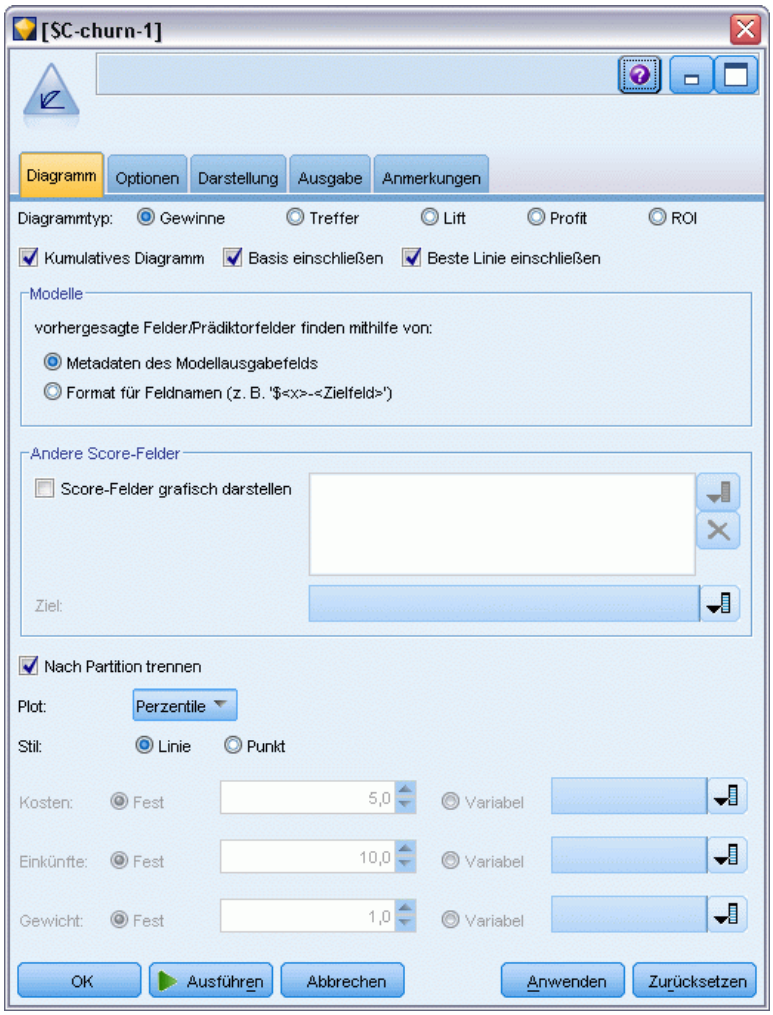
Die Methoden zur stufenweisen Auswahl gewährleisten, dass im Modell nur "statistisch signifikante" Prädiktoren enthalten sind, sie gewährleisten jedoch nicht, dass das Modell auch tatsächlich bei der Vorhersage des Ziels gute Ergebnisse liefert. Um dies zu erreichen, müssen Sie gescorte Datensätze analysieren.

Abbildung 27-15
Cox-Nugget: Registerkarte "Einstellungen"

- Platzieren Sie das Modell-Nugget im Zeichenbereich und verbinden Sie es mit dem Quellenknoten, öffnen Sie das Nugget und klicken Sie auf die Registerkarte "Einstellungen".
- Wählen Sie die Option Zeitfeld und geben Sie *tenure* an. Jeder Datensatz wird als Dauer des Kundenverhältnisses gescort.
- Wählen Sie die Option Alle Wahrscheinlichkeiten ausgeben.

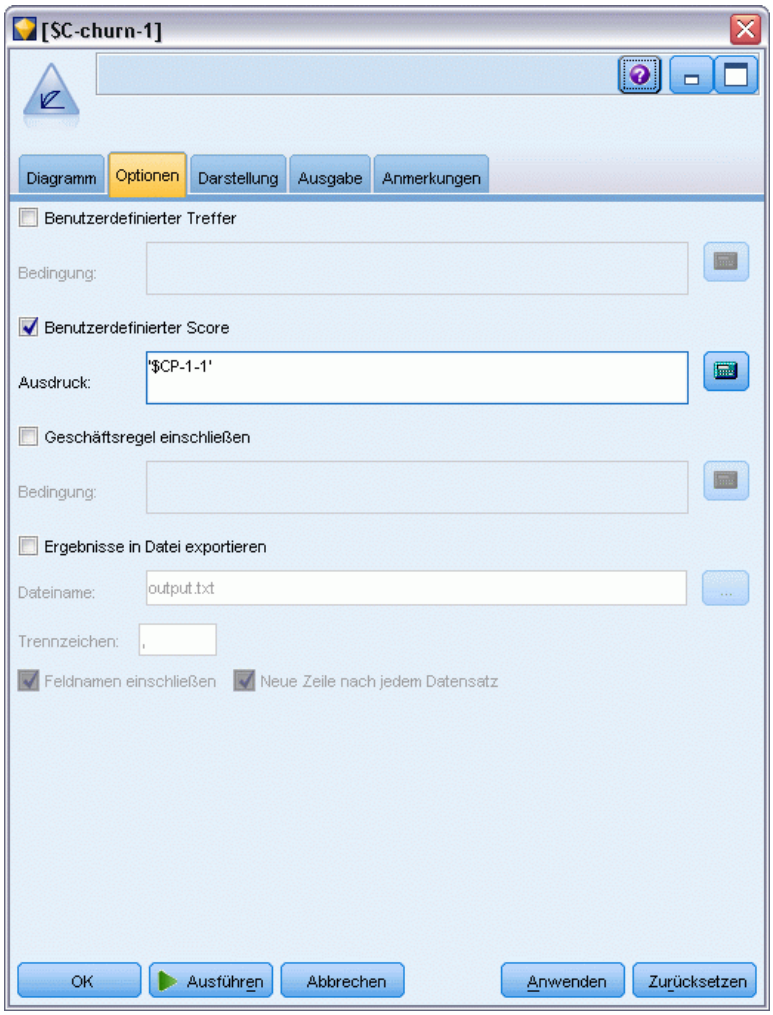
Dadurch werden Scores erstellt, wobei 0,5 als Trennwert für die Abwanderung eines Kunden verwendet wird; wenn die Abwanderungsneigung des Kunden mehr als 0,5 beträgt, wird der Kunde als "abwanderungswillig" gescort. Es muss nicht unbedingt der Wert 0,5 verwendet werden. Es ist durchaus möglich, dass ein anderer Trennwert zu wünschenswerteren Ergebnissen führt. Bei der Entscheidung über einen geeigneten Trennwert kann ein Evaluationsknoten hilfreich sein.

Abbildung 27-16
Evaluationsknoten: Registerkarte "Plot" (Diagramm)



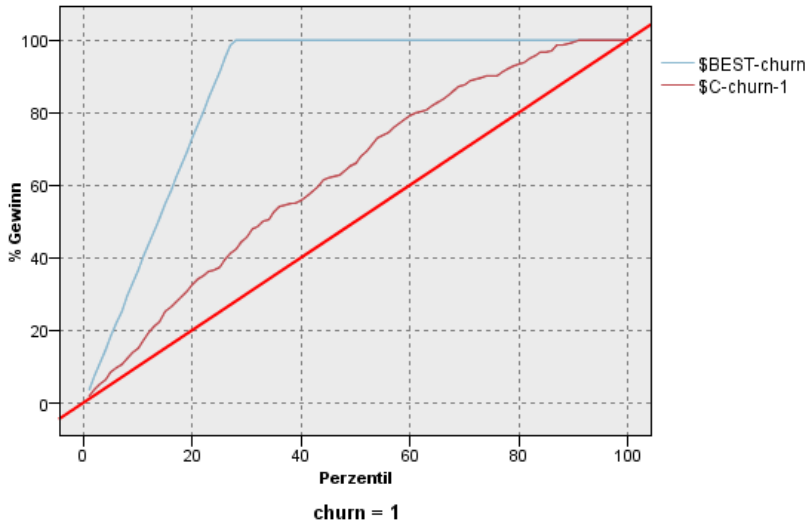
- Gliedern Sie einen Evaluationsknoten an das Modell-Nugget an. Wählen Sie auf der Registerkarte "Plot" (Diagramm) die Option Beste Linie einschließen.
- Klicken Sie auf die Registerkarte Optionen.

Abbildung 27-17
Evaluationsknoten: Registerkarte "Optionen"



- Wählen Sie die Option Benutzerdefinierter Score und geben Sie '$CP-1-1' als Ausdruck ein. Dies ist ein vom Modell generiertes Feld, das der Abwanderungsneigung entspricht.
- Klicken Sie auf Ausführen.

Abbildung 27-18
Gewinndiagramm



Das kumulative Gewinndiagramm zeigt den Prozentsatz der Fälle in einer bestimmten Kategorie, die "gewonnen" werden, indem ein bestimmter Prozentsatz der Gesamtzahl der Fälle anvisiert wird. Ein Punkt der Kurve liegt beispielsweise bei (10 %, 15 %). Dies bedeutet: Wenn Sie ein Daten-Set mit dem Modell scoren und alle Fälle nach der vorhergesagten Abwanderungsneigung sortieren, ist zu erwarten, dass die obersten 10 % der Fälle etwa 15 % der Fälle enthalten, die tatsächlich in Kategorie *1* (abwanderungswillig) fallen. Die obersten 60 % der Fälle enthalten etwa 79,2 % der abwanderungswilligen Personen. Wenn Sie 100 % des gescorten Daten-Sets auswählen, erhalten Sie alle abwanderungswilligen Personen im Daten-Set.

Die Diagonale Linie ist die "Basiskurve": Wenn Sie nach dem Zufallsprinzip 20 % der Datensätze im gescorten Daten-Set auswählen, ist zu erwarten, dass Sie ungefähr 20 % aller Datensätze "gewinnen", die tatsächlich in Kategorie *1* fallen. Je höher über der Basis eine Kurve liegt, desto größer ist der Gewinn. Die "beste" Linie zeigt die Kurve für ein "perfektes" Modell, das allen abwanderungswilligen Personen einen höheren Score für die Abwanderungsneigung zuweist als allen Personen, die nicht abwandern. Das kumulative Gewinndiagramm erleichtert die Auswahl eines Trennwerts für die Klassifizierung: Wählen Sie einen Prozentsatz aus, der dem angestrebten Gewinn entspricht, und ordnen Sie dann diesen Prozentsatz dem entsprechenden Trennwert zu.

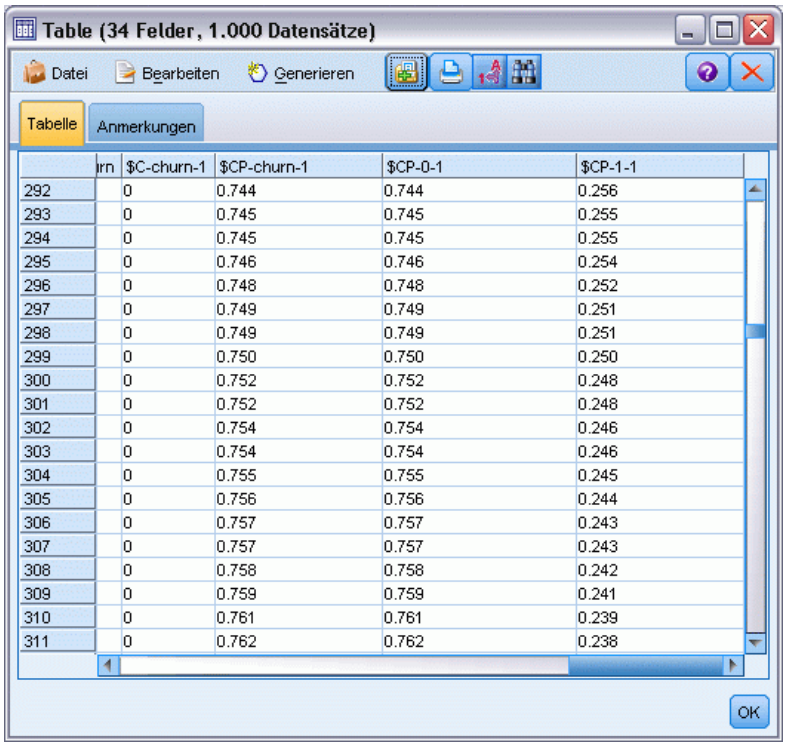
Welcher Gewinn angestrebt wird, hängt von den Kosten für Fehler erster und zweiter Art (Typ I und Typ II) ab. Die Frage ist also: Wie hoch sind die Kosten für eine fälschliche Klassifizierung einer abwanderungswilligen Person als Person, die nicht abwandert (Fehler erster Art)? Wie hoch sind die Kosten für eine fälschliche Klassifizierung einer Person, die nicht abwandert, als abwanderungswillige Person (Fehler zweiter Art)? Wenn Ihr Hauptanliegen im Halten der Kunden besteht, sollte der Fehler erster Art gesenkt werden. Im kumulativen Gewinndiagramm könnte dies einer verstärkten Pflege von Kunden entsprechen, die zu den obersten 60 % für die vorhergesagte Wahrscheinlichkeit von *1* gehören. Damit werden 79,2 % der möglichen Abwanderer erfasst, es müssen jedoch Zeit und Ressourcen aufgewendet werden, die ansonsten in die Akquise neuer Kunden investiert werden könnten. Wenn eine Senkung der Kosten für

den Erhalt des derzeitigen Kundenstamms oberste Priorität hat, sollte der Fehler zweiter Art gesenkt werden. Im Diagramm entspricht dies einer verstärkten Kundenpflege für die obersten 20 %, womit 32, 5 % der abwanderungswilligen Personen erfasst sind. Normalerweise sind beide Anliegen von Bedeutung, sodass Sie eine Entscheidungsregel für die Klassifizierung der Kunden ermitteln müssen, die die beste Mischung aus Sensitivität und Spezifität darstellt.

Abbildung 27-19
Sortierknoten: Registerkarte "Einstellungen"

- Angenommen, Sie haben sich entschieden, dass 45,6 % ein erstrebenswerter Gewinn ist. Dies entspricht der Verwendung der obersten 30 % der Datensätze. Um einen geeigneten Trennwert für die Klassifizierung zu finden, gliedern Sie einen Sortierknoten an das Modell-Nugget an.
- Legen Sie auf der Registerkarte "Einstellungen" fest, dass die Sortierung nach *$CP-1-1* in absteigender Reihenfolge vorgenommen werden soll, und klicken Sie auf OK.

Abbildung 27-20
Tabelle



	$C-churn-1	$CP-churn-1	$CP-0-1	$CP-1-1
292	0	0.744	0.744	0.256
293	0	0.745	0.745	0.255
294	0	0.745	0.745	0.255
295	0	0.746	0.746	0.254
296	0	0.748	0.748	0.252
297	0	0.749	0.749	0.251
298	0	0.749	0.749	0.251
299	0	0.750	0.750	0.250
300	0	0.752	0.752	0.248
301	0	0.752	0.752	0.248
302	0	0.754	0.754	0.246
303	0	0.754	0.754	0.246
304	0	0.755	0.755	0.245
305	0	0.756	0.756	0.244
306	0	0.757	0.757	0.243
307	0	0.757	0.757	0.243
308	0	0.758	0.758	0.242
309	0	0.759	0.759	0.241
310	0	0.761	0.761	0.239
311	0	0.762	0.762	0.238

- Verbinden Sie einen Tabellenknoten mit dem Sortierknoten.
- Öffnen Sie den Tabellenknoten und klicken Sie auf Ausführen.

Wenn Sie in der Ausgabe einen Bildlauf nach unten durchführen, sehen Sie, dass der Wert von *$CP-1-1* für den 300. Datensatz 0,248 beträgt. Bei Verwendung von 0,248 als Klassifizierungstrennwert sollten etwa 30 % der Kunden als abwanderungswillig gescort werden, wobei etwa 45 % der Gesamtzahl der Personen erfasst wird, die tatsächlich abwandern.

Verfolgung der erwarteten Anzahl an Kunden, die gehalten werden können

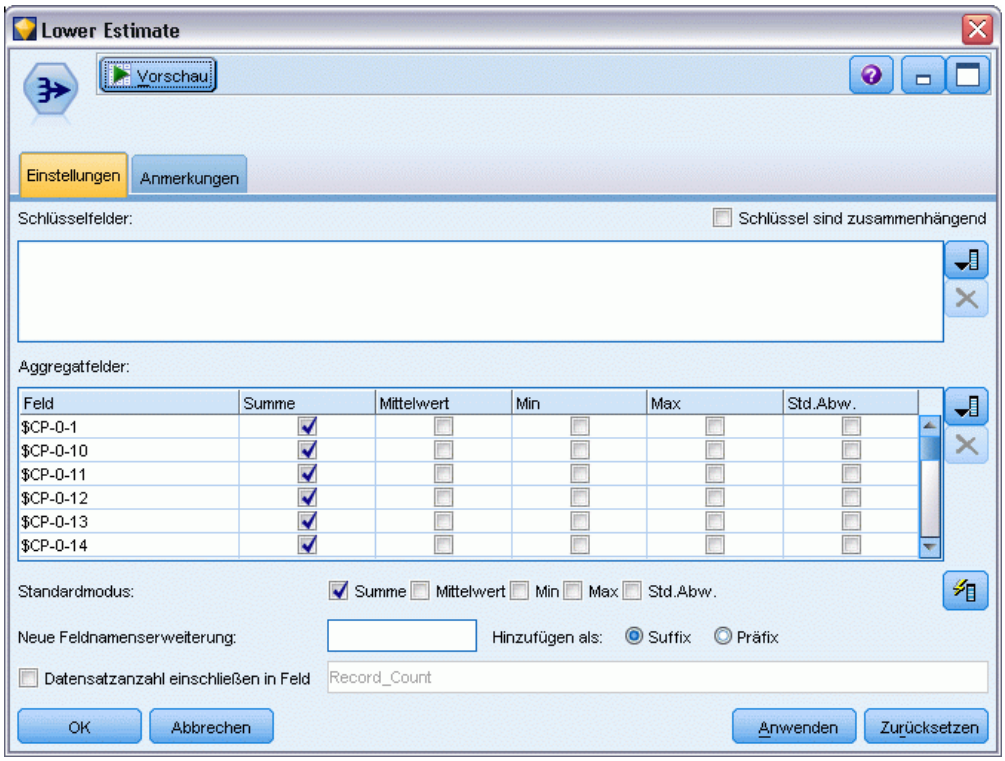
Wenn Sie mit einem Modell zufrieden sind, sollten Sie die erwartete Anzahl an Kunden im Daten-Set aufzeichnen, die in den nächsten beiden Jahren als Kunden gehalten werden können. Die Nullwerte, also Kunden, bei denen die Gesamtdauer des Kundenverhältnisses (zukünftige Zeit + *tenure*) über den Bereich der Überlebenszeiten hinausgeht, der zum Trainieren des Modells verwendet wurde, stellen eine interessante Herausforderung dar. Eine Möglichkeit für den Umgang mit diesen Kunden besteht darin, zwei Sets von Vorhersagen zu erstellen, eines, bei dem davon ausgegangen wird, dass die Kunden mit Nullwerten abgewandert sind, und ein anderes, bei

dem davon ausgegangen wird, dass sie als Kunden gehalten werden konnten. Auf diese Weise können Sie die Ober- und Untergrenze für die erwartete Anzahl der gehaltenen Kunden ermitteln.

Abbildung 27-21
Cox-Nugget: Registerkarte "Einstellungen"

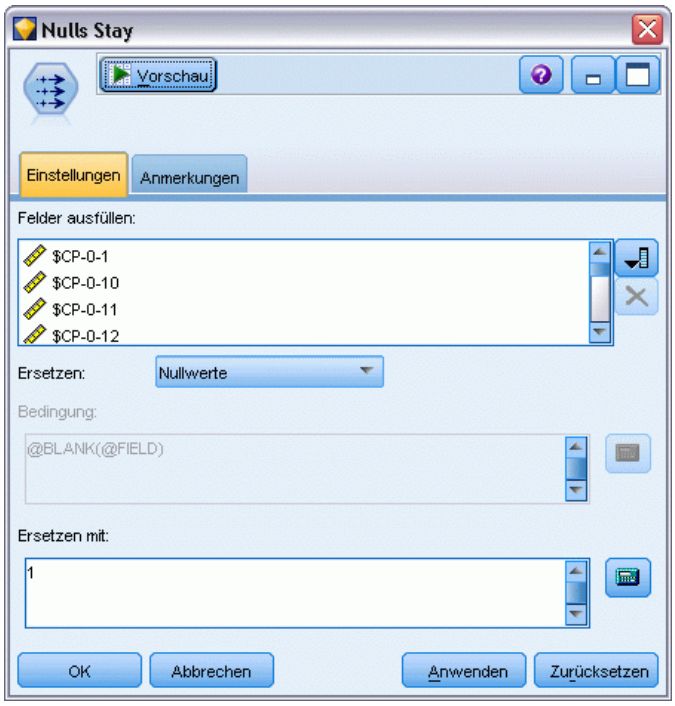
- Doppelklicken Sie auf das Modell-Nugget in der Modellpalette (oder kopieren Sie das Nugget und fügen Sie es auf der Stream-Zeichenfläche ein) und hängen Sie das neue Nugget an den Quellenknoten an.
- Öffnen Sie die Registerkarte "Einstellungen" für das Nugget.
- Vergewissern Sie sich, dass Regelmäßige Intervalle ausgewählt ist, und geben Sie 1,0 als Zeitintervall und 24 als Anzahl der zu scorenden Zeiträume ein. Dies bedeutet, dass jeder Datensatz für jeden der folgenden 24 Monate gescort wird.
- Wählen Sie *tenure* als Feld zur Angabe der vergangenen Überlebenszeit ein. Der Scoring-Algorithmus berücksichtigt die die Zeitdauer, die jede Person Kunde des Unternehmens war.
- Wählen Sie die Option Alle Wahrscheinlichkeiten ausgeben.

Abbildung 27-22
Aggregatknoten: Registerkarte "Einstellungen"



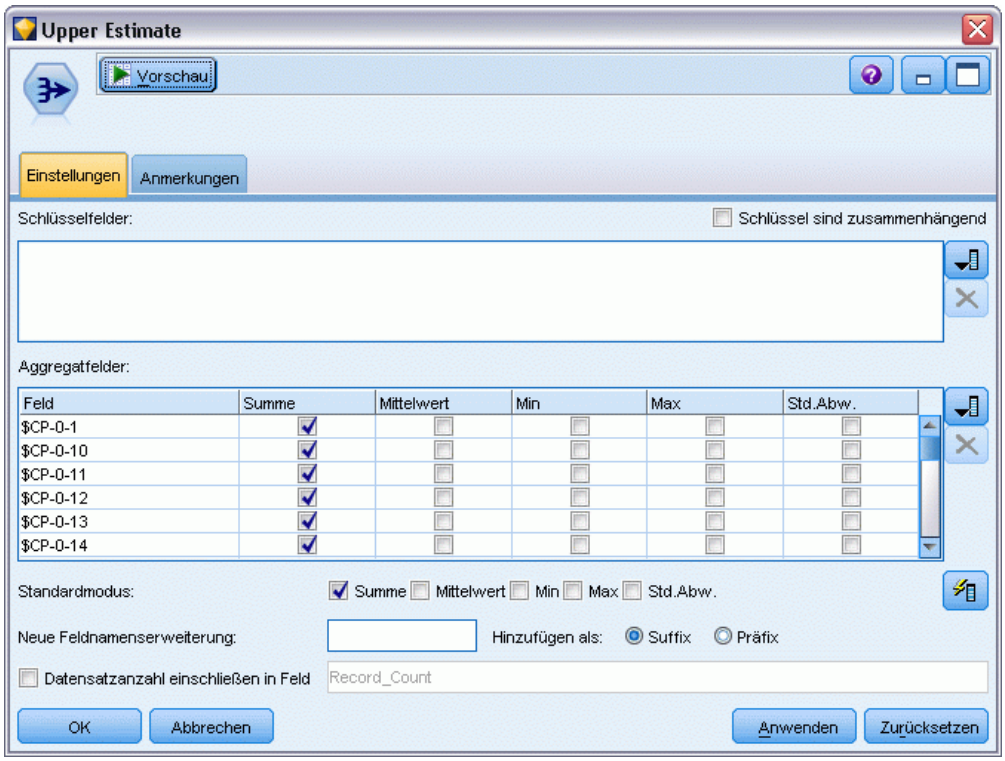
- Gliedern Sie einen Aggregatknoten an das Modell-Nugget an. Heben Sie auf der Registerkarte "Einstellungen" die Auswahl der Option Mittelwert als Standardmodus auf.
- Wählen Sie die Felder *$CP-0-1* bis *$CP-0-24*, die Felder des Formats *$CP-0-n*, als zu aggregierende Felder aus. Dies geht am einfachsten, wenn Sie auf der Registerkarte "Felder auswählen", die Felder nach Namen (also in alphabetischer Reihenfolge) sortieren.
- Heben Sie die Auswahl von Datensatzanzahl einschließen in Feld auf.
- Klicken Sie auf OK. Dieser Knoten erstellt die Vorhersagen für die "Untergrenze".

Abbildung 27-23
Füllerknoten: Registerkarte "Einstellungen"



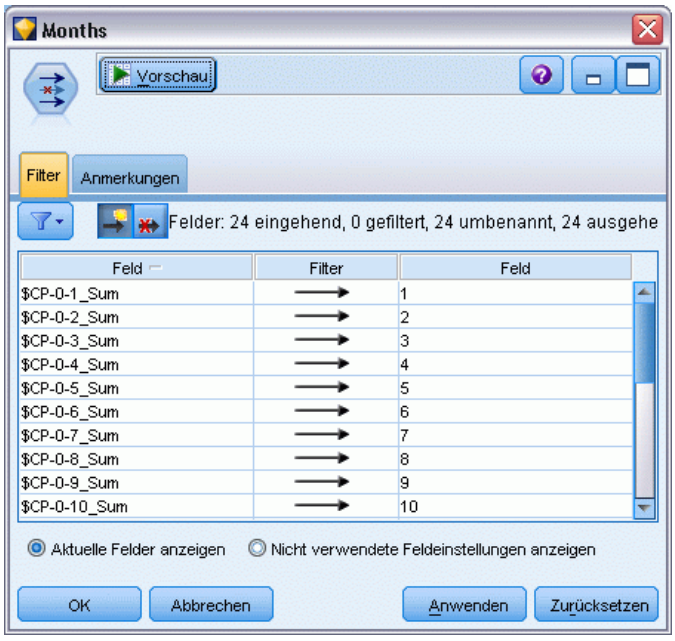
- Gliedern Sie einen Füllerknoten an das Coxreg-Nugget an, das Sie soeben an den Aggregatknoten angegliedert haben. Wählen Sie auf der Registerkarte "Einstellungen" die Felder *$CP-0-1* bis *$CP-0-24*, die Felder des Formats *$CP-0-n* als auszufüllende Felder aus. Dies geht am einfachsten, wenn Sie auf der Registerkarte "Felder auswählen", die Felder nach Namen (also in alphabetischer Reihenfolge) sortieren.
- Legen Sie fest, dass Nullwerte durch den Wert 1 ersetzt werden sollen.
- Klicken Sie auf OK.

Abbildung 27-24
Aggregatknoten: Registerkarte "Einstellungen"



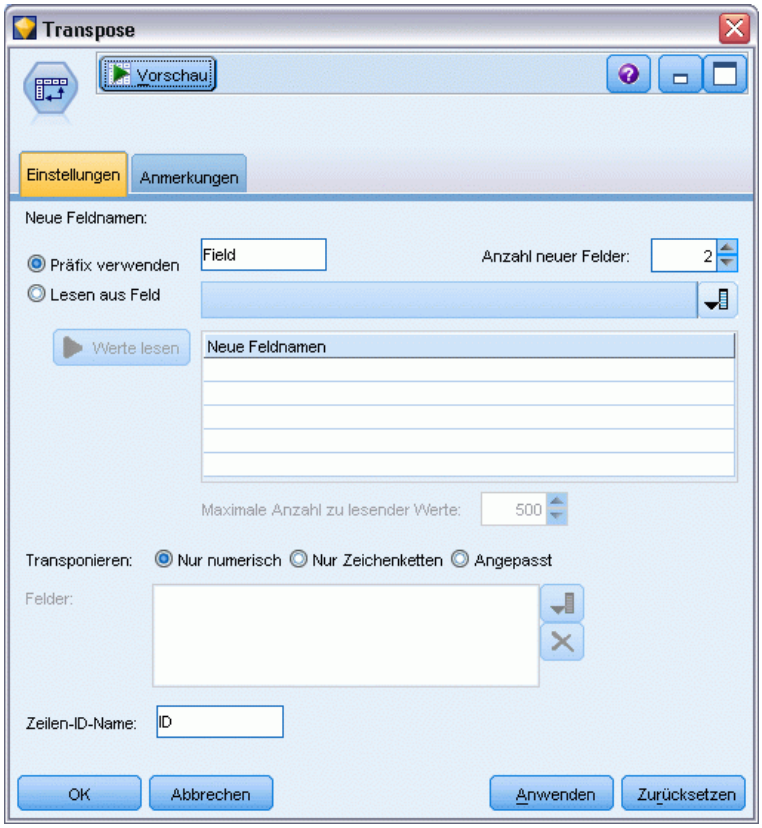
- Gliedern Sie einen Aggregatknoten an den Füllerknoten an. Heben Sie auf der Registerkarte "Einstellungen" die Auswahl der Option Mittelwert als Standardmodus auf.
- Wählen Sie die Felder *$CP-0-1* bis *$CP-0-24*, die Felder des Formats *$CP-0-n*, als zu aggregierende Felder aus. Dies geht am einfachsten, wenn Sie auf der Registerkarte "Felder auswählen", die Felder nach Namen (also in alphabetischer Reihenfolge) sortieren.
- Heben Sie die Auswahl von Datensatzanzahl einschließen in Feld auf.
- Klicken Sie auf OK. Dieser Knoten erstellt die Vorhersagen für die "Obergrenze".

Abbildung 27-25
Filterknoten: Registerkarte "Einstellungen"



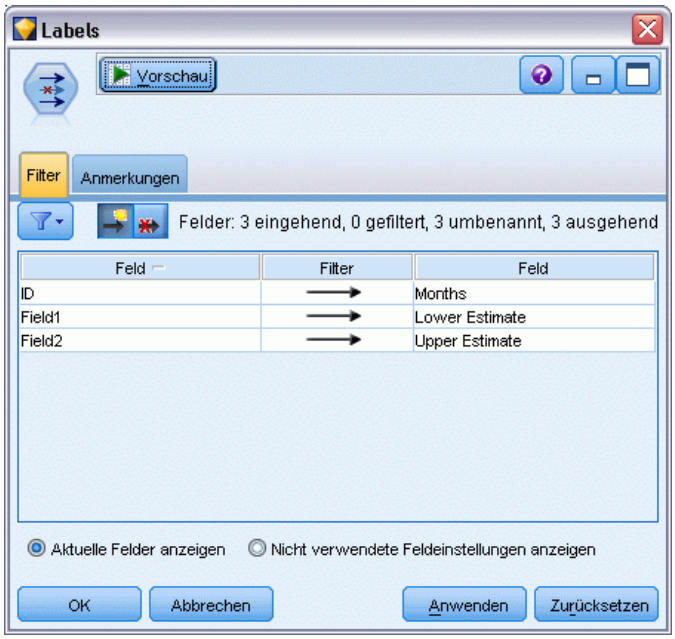
- Gliedern Sie einen Anhangknoten an die beiden Aggregatknoten an und gliedern Sie anschließend einen Filterknoten an den Anhangknoten an.
- Benennen Sie auf der Registerkarte "Einstellungen" des Filterknotens die Felder in *1* bis *24* um. Durch die Verwendung eines Transponierknotens werden diese Feldnamen zu Werten für die *x*-Achse in Diagrammen weiter unten im Stream.

Abbildung 27-26
Transponierknoten: Registerkarte "Einstellungen"



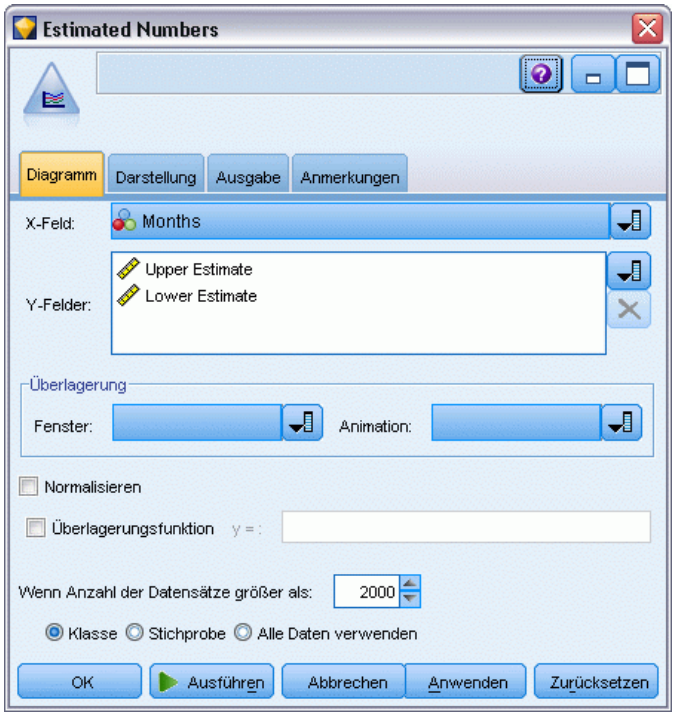
- GliedernSie einen Transponierknoten an den Filterknoten an.
- Wählen Sie 2 als Anzahl der neuen Felder aus.

Abbildung 27-27
Filterknoten: Registerkarte "Filter"



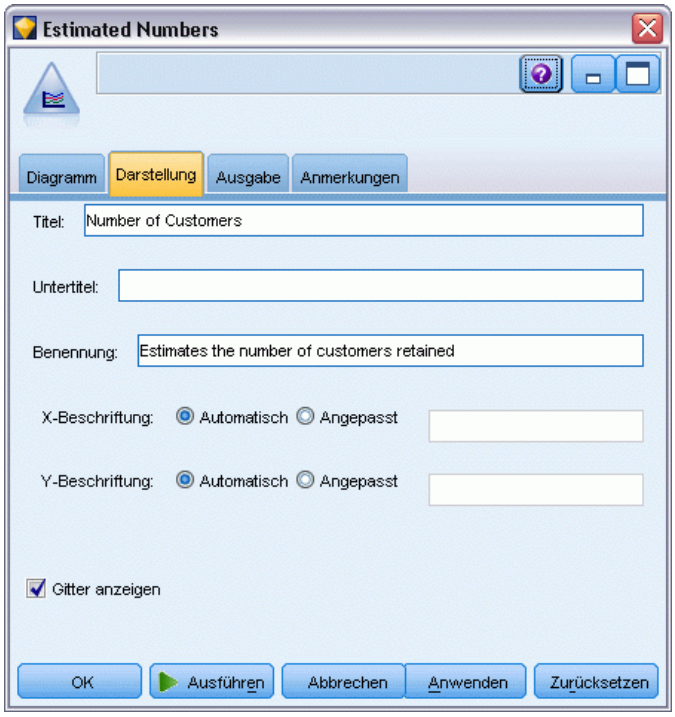
- Gliedern Sie einen Filterknoten an den Transponierknoten an.
- Benennen Sie auf der Registerkarte "Einstellungen" des Filternknotens *ID* in *Months* (Monate), *Feld1* in *Lower Estimate* (Unterer Schätzer) und *Feld2* in *Upper Estimate* (Oberer Schätzer) um.

Abbildung 27-28
Multidiagrammknoten: Registerkarte "Plot" (Diagramm)



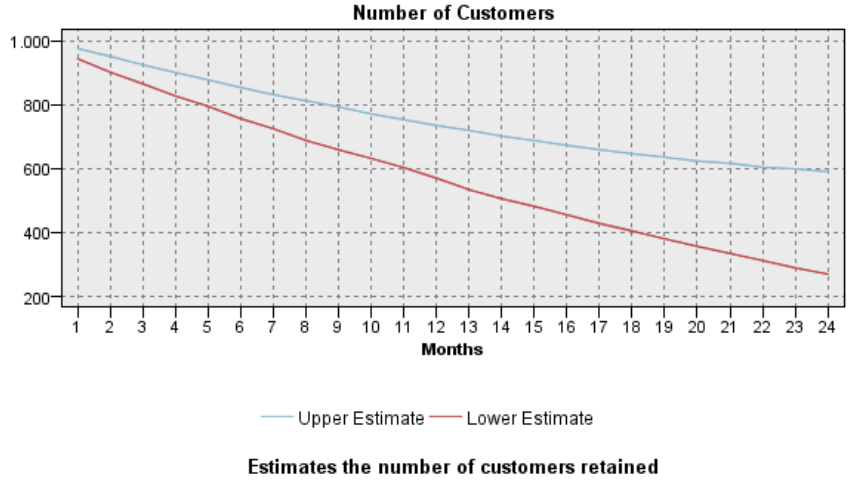
- Gliedern Sie einen Multidiagrammknoten an den Filterknoten an.
- Wählen Sie auf der Registerkarte "Plot" *Months* (Monate) als X-Feld und *Lower Estimate* (Unterer Schätzer) und *Upper Estimate* (Oberer Schätzer) als Y-Feld aus.

Abbildung 27-29
Multidiagrammknoten: Darstellung, Registerkarte



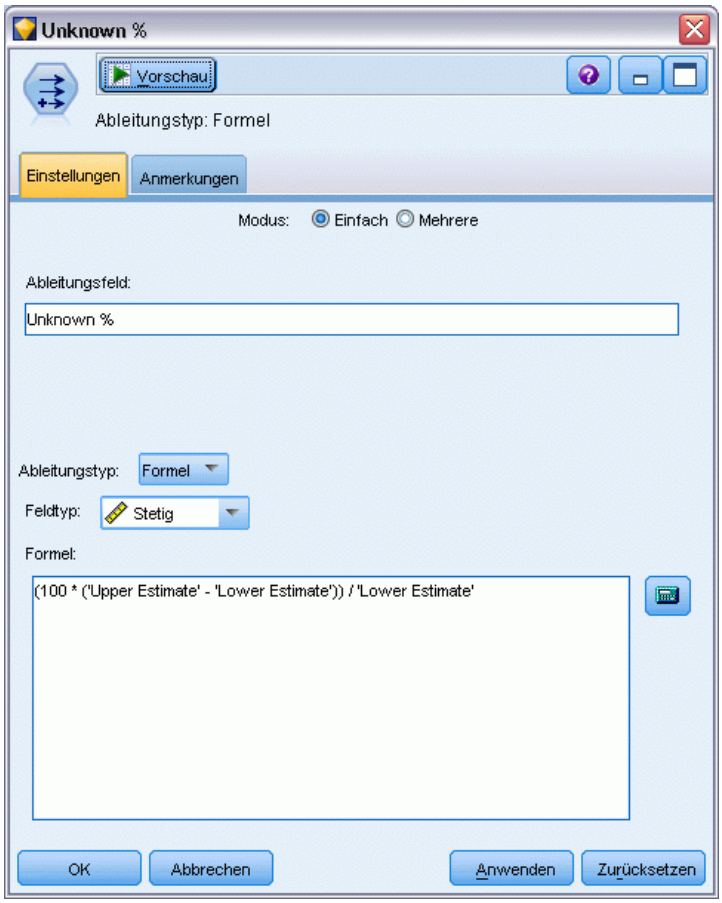
- Klicken Sie auf die Registerkarte “Darstellung”
- Geben Sie Number of Customers (Anzahl der Kunden) als Titel ein.
- Geben Sie Schätzt die Anzahl der gehaltenen Kunden Benennung ein.
- Klicken Sie auf Ausführen.

Abbildung 27-30
Multidiagramm zur Schätzung der Anzahl der gehaltenen Kunden



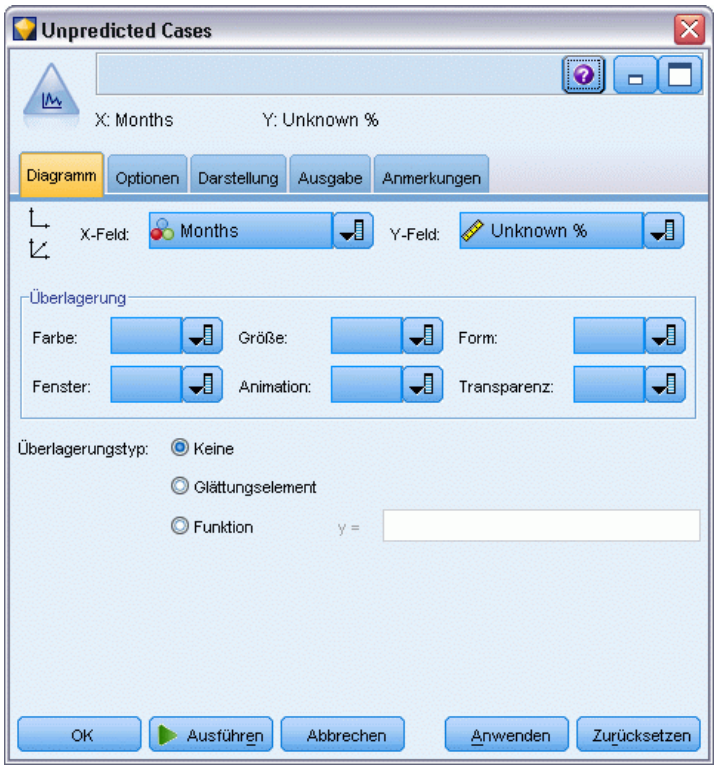
Die Ober- und die Untergrenze der geschätzten Anzahl an Kunden, die gehalten werden können, werden grafisch dargestellt. Die Differenz zwischen den beiden Linien ist die Anzahl an Kunden, die als “null” gescort wurden, deren Status also höchst ungewiss ist. Im Laufe der Zeit steigt die Anzahl dieser Kunden. Nach 12 Monaten können Sie erwarten, dass zwischen 601 und 735 der ursprünglichen Kunden im Daten-Set noch erhalten geblieben sind; nach 24 Monaten liegt dieser Wert zwischen 288 und 597.

Abbildung 27-31
Ableitungsknoten: Registerkarte "Einstellungen"



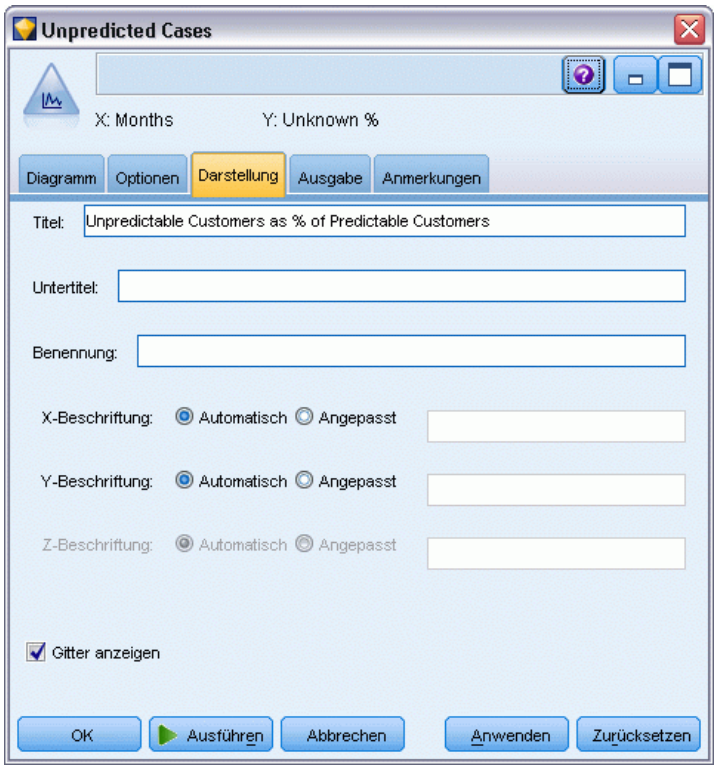
- Um einen weiteren Einblick darin zu erhalten, wie unsicher die Schätzer für die Anzahl der gehaltenen Kunden sind, gliedern Sie einen Ableitungsknoten an den Filterknoten an.
- Geben Sie auf der Registerkarte "Einstellungen" des Ableitungsknotens *Unknown %* (% unbekannt) als Ableitungsfeld ein.
- Wählen Sie Stetig als Feldtyp aus.
- Geben Sie (100 * ('Upper Estimate' - 'Lower Estimate')) / 'Lower Estimate' als Formel ein *Unknown %* (% unbekannt) gibt die Anzahl der Kunden, über die Zweifel bestehen, als Prozentsatz des unteren Schätzers an.
- Klicken Sie auf OK.

Abbildung 27-32
Plotknoten: Registerkarte "Plot" (Diagramm)



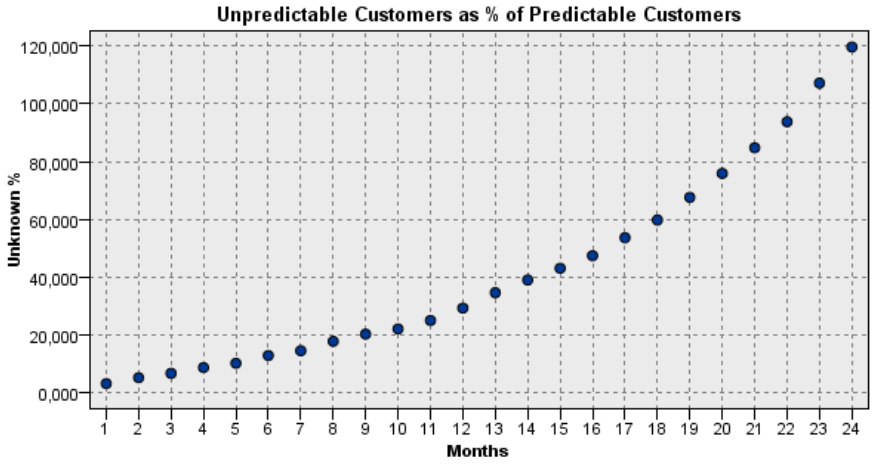
- Gliedern Sie einen Plotknoten an den Ableitungsknoten an.
- Wählen Sie auf der Registerkarte "Plot" (Diagramm) des Plotknotens *Months* (Monate) als X-Feld und *Unknown %* (% unbekannt) als Y-Feld aus.
- Klicken Sie auf die Registerkarte Darstellung.

Abbildung 27-33
Plotknoten: Darstellung, Registerkarte



- Geben Sie Unpredictable Customers as % of Predictable Customers (Unvorhersagbare Kunden als % der vorhersagbaren Kunden) als Titel ein.
- Führen Sie den Knoten aus.

Abbildung 27-34
Plot der unvorhersagbaren Kunden



Während des ersten Jahres steigt der Prozentsatz der unvorhersagbaren Kunden relativ linear an, im zweiten Jahr jedoch explodiert die Anstiegsrate, bis ab Monat 23 die Anzahl der Kunden mit Nullwerten die erwartete Anzahl gehaltener Kunden übersteigt.

Scoring

Wenn Sie mit einem Modell zufrieden sind, können Sie die Kunden (nach Quartal) scoren, um die Personen, die mit der höchsten Wahrscheinlichkeit innerhalb des nächsten Jahres abwandern, zu ermitteln.

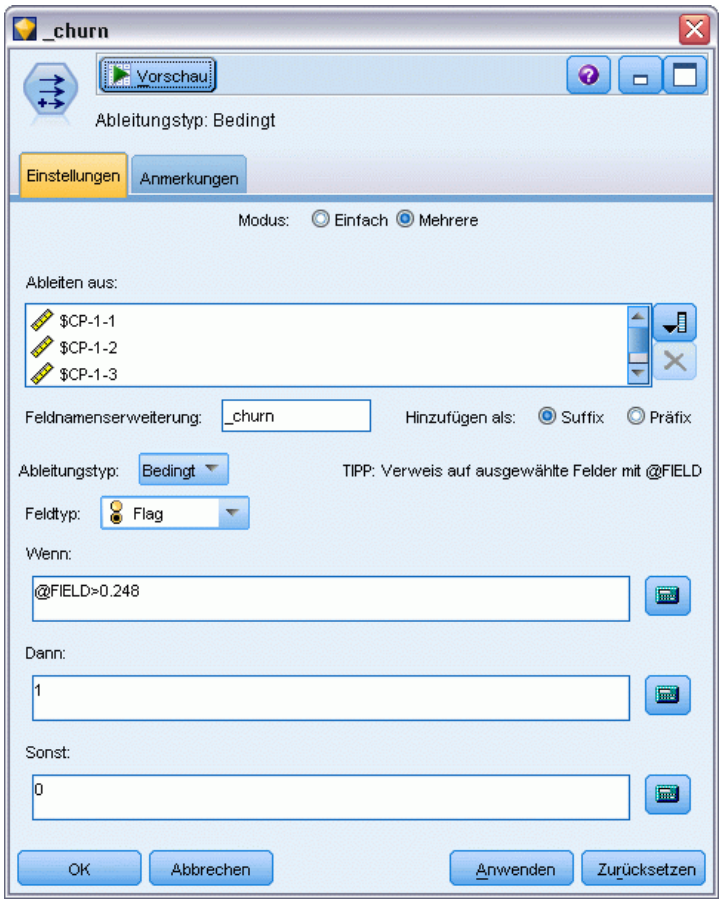
Abbildung 27-35
Coxreg-Nugget: Registerkarte "Einstellungen"

- Gliedern Sie ein drittes Modell-Nugget an Quellenknoten an und öffnen Sie das Modell-Nugget.
- Vergewissern Sie sich, dass Regelmäßige Intervalle ausgewählt ist, und geben Sie 3,0 als Zeitintervall und 4 als Anzahl der zu scorenden Zeiträume ein. Dies bedeutet, dass jeder Datensatz für die folgenden 4 Quartale gescort wird.
- Wählen Sie *tenure* als Feld zur Angabe der vergangenen Überlebenszeit ein. Der Scoring-Algorithmus berücksichtigt die die Zeitdauer, die jede Person Kunde des Unternehmens war.
- Wählen Sie die Option Alle Wahrscheinlichkeiten ausgeben. Diese zusätzlichen Felder erleichtern die Sortierung der Datensätze zur Anzeige in einer Tabelle.

Abbildung 27-36
Auswahlknoten: Registerkarte "Einstellungen"

- Gliedern Sie einen Auswahlknoten an das Modell-Nugget an. Geben Sie auf der Registerkarte "Einstellungen" churn=0 als Bedingung ein. Dadurch werden Kunden, die bereits abgewandert sind, aus der Ergebnistabelle entfernt.

Abbildung 27-37
Ableitungsknoten: Registerkarte "Einstellungen"

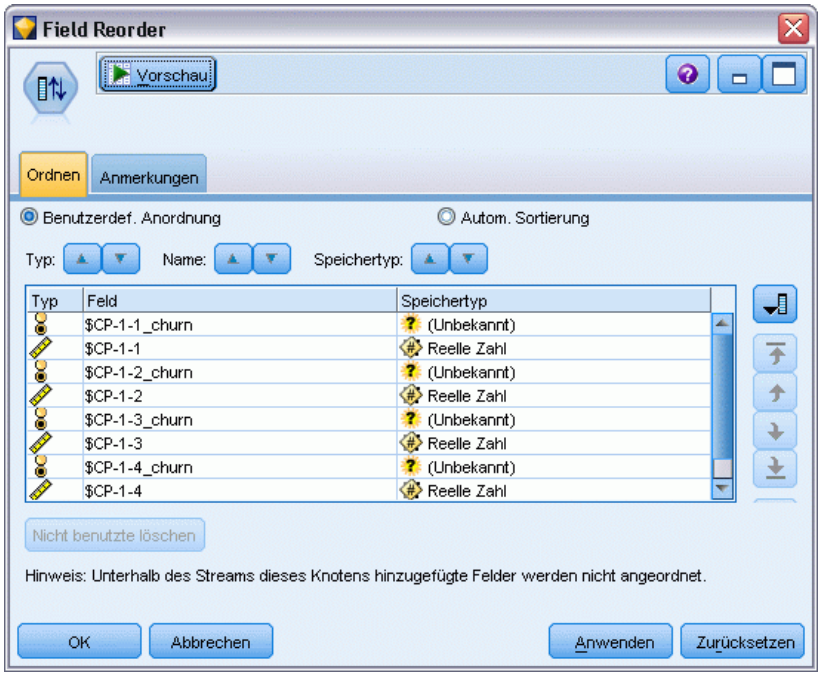


- Gliedern Sie einen Ableitungsknoten an den Auswahlknoten an. Wählen Sie auf der Registerkarte "Einstellungen" Mehrere als Modus aus.
- Legen Sie fest, dass die Ableitung aus den Feldern *$CP-1-1* bis *$CP-1-4*, den Feldern des Formats *$CP-1-n*, erfolgen soll, und geben Sie _churn als hinzuzufügendes Suffix an. Dies geht am einfachsten, wenn Sie auf der Registerkarte "Felder auswählen" die Felder nach Namen (also in alphabetischer Reihenfolge) sortieren.
- Wählen Sie für die Ableitung des Felds die Option Bedingt aus.
- Wählen Sie Flag als Messniveau aus.
- Geben Sie @FIELD>0,248 als Wenn-Bedingung ein. Wie Sie sich erinnern, war dies der während der Evaluation ermittelte Trennwert für die Klassifizierung.
- Geben Sie 1 als Dann-Ausdruck ein.
- Geben Sie 0 als Sonst-Ausdruck ein.
- Klicken Sie auf OK.

Abbildung 27-38
Sortierknoten: Registerkarte "Einstellungen"

- Gliedern Sie einen Sortierknoten an den Ableitungsknoten an. Legen Sie auf der Registerkarte "Einstellungen" fest, dass die Sortierung nach *$CP-1-1_churn* bis *$CP-1-4-churn* und anschließend nach *$CP-1-1* bis *$CP-1-4* (jeweils in absteigender Reihenfolge) erfolgen soll. Kunden, für die eine Abwanderung vorhergesagt wurde, werden oben angezeigt.

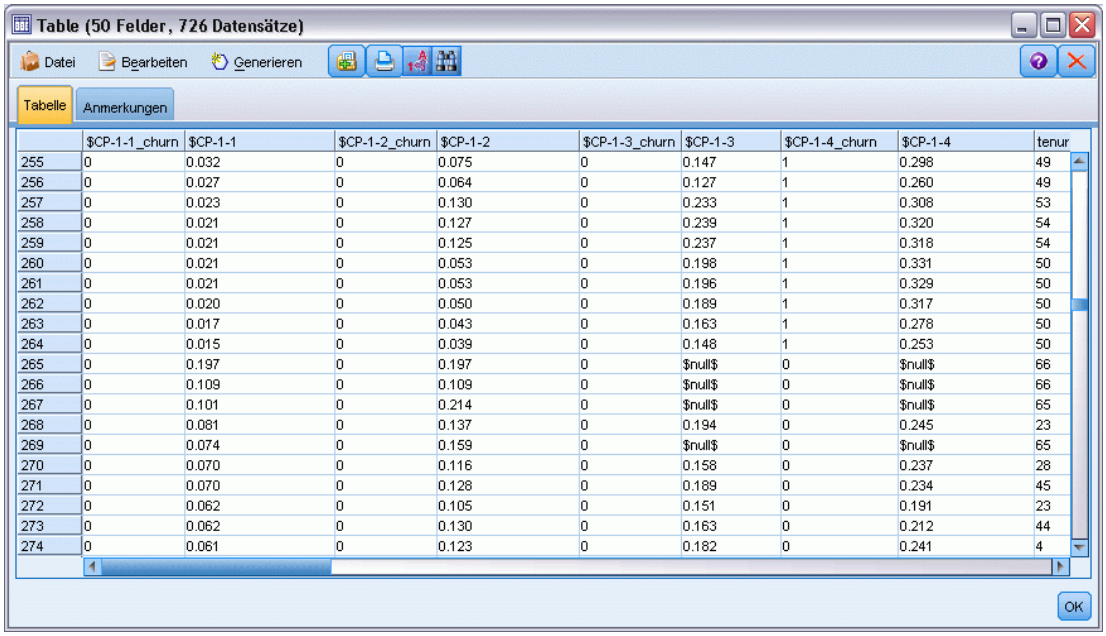
Abbildung 27-39
Knoten "Felder ordnen": Registerkarte "Ordnen"



- Gliedern Sie einen Knoten vom Typ "Felder ordnen" an den Sortierknoten an. Legen Sie auf der Registerkarte "Ordnen" fest, dass die Felder *$CP-1-1_churn* bis *$CP-1-4* vor den anderen

Feldern platziert werden sollen. Durch diesen optionalen Vorgang wird die Ergebnistabelle leichter lesbar. Sie müssen die Schaltflächen verwenden, um die Felder an die in der Abbildung gezeigte Position zu verschieben.

Abbildung 27-40
Tabelle mit Kunden-Scores

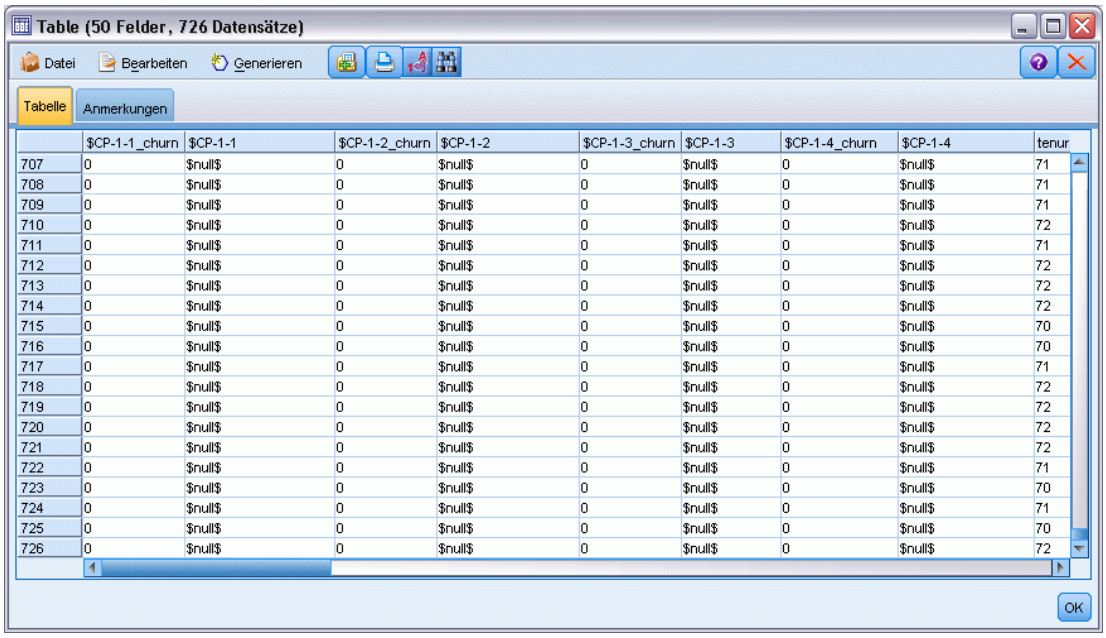


	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

- Gliedern Sie einen Tabellenknoten an den Knoten “Felder ordnen” an und führen Sie ihn aus.

Es wird erwartet, dass 264 bis zum Ende des Jahres abwandern, 184 bis zum Ende des dritten Quartals, 103 bis zum Ende des zweiten Quartals und 31 im ersten Quartal. Beachten Sie: Bei zwei Kunden weist der Kunde mit der höheren Abwanderungsneigung im ersten Quartal nicht unbedingt auch in späteren Quartalen eine höhere Abwanderungsneigung auf. Betrachten Sie beispielsweise die Datensätze 256 und 260. Der Grund hierfür liegt vermutlich in der Form der Hazard-Funktion für die Monate nach der aktuellen Dauer des jeweiligen Kundenverhältnisses. So liegt bei Kunden, die aufgrund einer Werbeaktion zum Unternehmen kamen, die Wahrscheinlichkeit einer frühen Abwanderung höher als bei Kunden, die sich aufgrund einer persönlichen Empfehlung für das Unternehmen entschieden, aber Werbeaktionskunden, die nicht frühzeitig abwandern, sind möglicherweise für den restlichen Zeitraum treuer als die Kunden mit persönlicher Empfehlung. Es kann sinnvoll sein, die Kunden neu zu sortieren, um die Kunden, die mit der größten Wahrscheinlichkeit abwandern, unter verschiedenen Gesichtspunkten zu betrachten.

Abbildung 27-41
Tabelle mit Kunden mit Nullwerten



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

Unten in der Tabelle befinden sich Kunden mit vorhergesagten Nullwerten. Hierbei handelt es sich um Kunden, bei denen die Gesamtdauer des Kundenverhältnisses (zukünftige Zeit + *tenure*) über den Bereich der Überlebenszeiten hinausgeht, der zum Trainieren des Modells verwendet wurde.

Zusammenfassung

Mithilfe der Cox-Regression haben Sie ein akzeptables Modell für die Zeit bis zur Abwanderung ermittelt, die erwartete Anzahl der gehaltenen Kunden für die nächsten zwei Jahre in einem Plot grafisch dargestellt und die einzelnen Kunden ermittelt, die mit der höchsten Wahrscheinlichkeit im nächsten Jahr abwandern. Beachten Sie, dass dies zwar ein akzeptables Modell ist, jedoch nicht unbedingt das beste. Idealerweise sollten Sie dieses Modell, das mit der Methode “Schrittweise vorwärts” erstellt wurde, zumindest mit einem Modell vergleichen, das mit der Methode “Schrittweise rückwärts” erstellt wurde.

Erläuterungen der mathematischen Grundlagen der in IBM® SPSS® Modeler verwendeten Modellierungsverfahren sind im *SPSS Modeler-Algorithmushandbuch* aufgeführt.

Kapitel
28

Warenkorbanalyse (Regelinduktion/C5.0)

In diesem Beispiel werden frei erfundene Daten verwendet, die den Inhalt von Supermarkt-Warenkörben beschreiben (d. h. eine Sammlung von gekauften Produkten), mit den verknüpften persönlichen Daten des Käufers, die über eine Treuekarte ermittelt werden können. Das Ziel besteht darin, Gruppen von Kunden zu ermitteln, deren Kaufverhalten ähnlich ist und die demografisch beschrieben werden können, z. B. nach Alter, Einkommen usw.

Dieses Beispiel stellt zwei Phasen des Data-Minings dar:

- Assoziationsregelmodellierung und eine Netzdiagramm-Anzeige, die Zusammenhänge zwischen gekauften Produkten deutlich macht.
- C5.0-Regelinduktion, die ein Profil der Käufer von bestimmten Produktgruppen erstellt.

Hinweis: Diese Anwendung verwendet die Vorhersagemodellierung nicht direkt, sodass es keine Genauigkeitsmessung für die resultierenden Modelle und keine damit verbundene Unterscheidung zwischen Training/Test im Data-Mining-Prozess gibt.

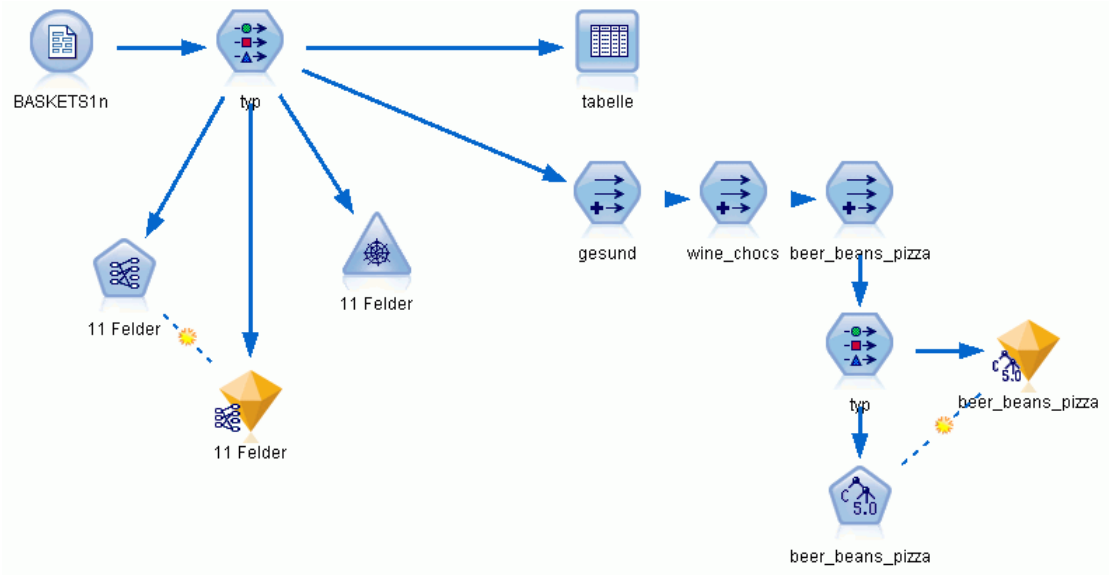
In diesem Beispiel wird der Stream *baskrule* verwendet, der Bezug nimmt auf die Datendatei *BASKETS1n*. Die Dateien stehen im Verzeichnis *Demos* der IBM® SPSS® Modeler-Installation zur Verfügung. Dieser Ordner kann über die Programmgruppe "IBM® SPSS® Modeler" im Windows-Startmenü aufgerufen werden. Die Datei *baskrule* befindet sich im Verzeichnis *streams*.

Datenzugriff

Stellen Sie unter Verwendung des Knotens "Variable Datei" eine Verbindung mit dem Daten-Set *BASKETS1n* her und wählen Sie aus, dass die Feldnamen aus der Datei gelesen werden. Verbinden Sie einen Typknoten mit der Datenquelle und verbinden Sie den Knoten dann mit einem Tabellenknoten. Setzen Sie das Messniveau des Felds *CardID* auf *Ohne Typ* (da jede Treuekarten-ID nur einmal im Daten-Set vorkommt und deshalb für die Modellierung nicht verwendet werden kann). Wählen Sie *Nominal* als Messniveau für das Feld *Geschlecht*. (Damit

wird sichergestellt, dass der A Priori-Modellierungsalgorithmus *Geschlecht* nicht als Flag behandelt wird.)

Abbildung 28-1
baskrule-Stream



Führen Sie den Stream jetzt aus, um den Typknoten zu instanziieren und die Tabelle anzuzeigen. Das Daten-Set enthält 18 Felder, wobei jeder Datensatz einen Korb darstellt.

Die 18 Felder werden in den folgenden Überschriften dargestellt.

Warenkorbübersicht:

- *CardID*. Treuekartenbezeichner für Kunden, die diesen Warenkorb kaufen.
- *Wert*. Gesamter Kaufpreis des Warenkorbs.
- *Zahlart*. Zahlungsweise für den Warenkorb.

Persönliche Informationen über den Karteninhaber:

- *Geschlecht*
- *Hausbesitzer*. Ob der Karteninhaber Hausbesitzer ist.
- *income*
- *Alter*

Warenkorbinhalt – ist in die folgenden Produktkategorien aufgeteilt:

- *Obst*
- *Fleisch*
- *Milchprod*
- *Konservengemüse*
- *Konservenfleisch*

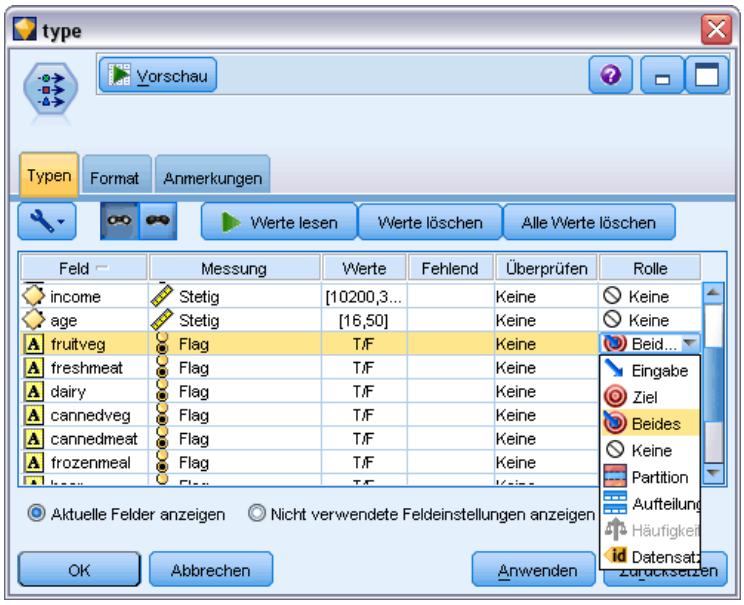
- *TK-Fertiggericht*
- *Bier*
- *Wein*
- *Softdrink*
- *Fisch*
- *Süßwaren*

Entdecken von Affinitäten beim Warenkorbinhalt

Zunächst müssen Sie einen Überblick über die Affinitäten (Assoziationen) im Warenkorbinhalt erhalten. Verwenden Sie dazu A Priori, um Assoziationsregeln zu erstellen. Wählen Sie die in diesem Modellierungsprozess zu verwendenden Felder, indem Sie den Typknoten bearbeiten, die Rolle aller Produktkategorien auf *Beides* setzen und die Rolle für alle anderen Elemente auf *Keine*. (*Beides* bedeutet, dass es sich bei dem Feld entweder um die Eingabe oder um die Ausgabe des resultierenden Modells handelt.)

Hinweis: Mit Umschalt-Klicken können Sie Optionen für mehrere Felder festlegen. So können Sie die Felder auswählen, bevor Sie eine Option aus den Spalten festlegen.

Abbildung 28-2
Auswählen von Feldern für die Modellierung



Sobald Sie Felder für die Modellierung festgelegt haben, verbinden Sie einen A Priori-Knoten mit dem Typknoten, bearbeiten diesen, wählen die Option Nur wahre Werte für Flags und führen den A Priori-Knoten aus. Das Ergebnis, ein Modell auf der Registerkarte “Modelle” oben rechts

im Manager-Fenster, enthält Assoziationsregeln, die Sie mithilfe des Kontextmenüs und unter Auswahl von Durchsuchen anzeigen können.

Abbildung 28-3
Assoziationsregeln

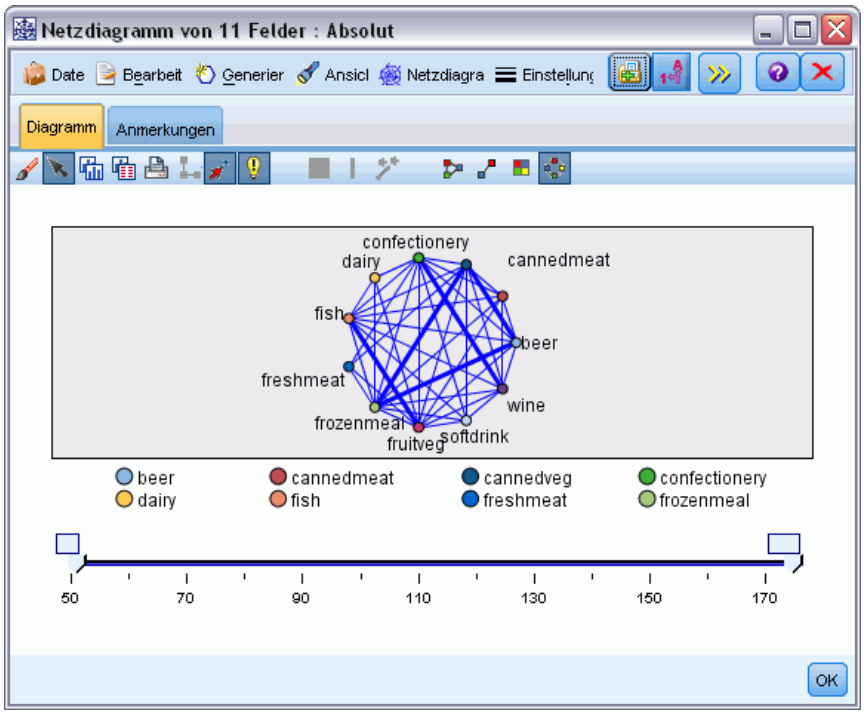
Diese Regeln zeigen eine Vielzahl von Assoziationen zwischen TK-Fertiggerichten, Konservengemüse und Bier. Das Vorhandensein von Assoziationsregeln in beide Richtungen, wie:

TK-Fertiggericht -> Bier
Bier -> TK-Fertiggericht

legt nahe, dass eine Netzdiagramm-Anzeige (die nur Verbindungen in beide Richtungen darstellt) einige der Muster in diesen Daten hervorhebt.

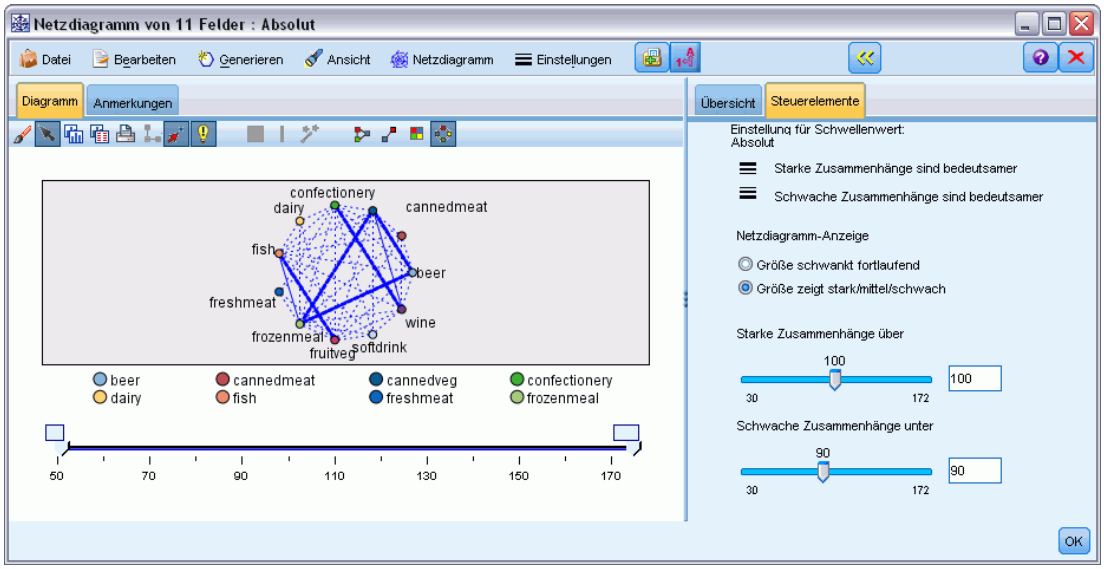
Verbinden Sie einen Netzknoten mit einem Typknoten, bearbeiten Sie den Netzknoten, wählen Sie alle Felder für den Warenkorbinhalt aus, wählen Sie Nur wahre Flags anzeigen und führen Sie den Netzknoten aus.

Abbildung 28-4
Netzdiagramm-Anzeige der Produktverbindungen



Da die meisten Kombinationen von Produktkategorien in mehreren Warenkörben auftreten, sind die starken Zusammenhänge in diesem Netz zu zahlreich, um die vom Modell vorgeschlagenen Gruppen von Kunden anzuzeigen.

Abbildung 28-5
Beschränkte Netzdiagramm-Anzeige



- Zur Angabe von schwachen und starken Verbindungen klicken Sie auf den gelben Doppelpfeil in der Symbolleiste. Auf diese Weise wird das Dialogfeld erweitert und es werden die Webausgabeübersicht und die Steuerungen angezeigt.
- Wählen Sie Größe zeigt stark/mittel/schwach.
- Legen Sie fest: Schwache Zusammenhänge unter 90.
- Legen Sie fest: Starke Zusammenhänge über 100.

In der so entstehenden Anzeige treten drei Gruppen von Kunden in den Vordergrund:

- Die Kunden, die Fisch, Obst und Gemüse kaufen und die als “gesunde Esser” bezeichnet werden können.
- Die Kunden, die Wein und Süßwaren kaufen.
- Die Kunden, die Bier, Tiefkühl-Fertiggerichte und Konservengemüse (“Bier, Bohnen und Pizza”) kaufen.

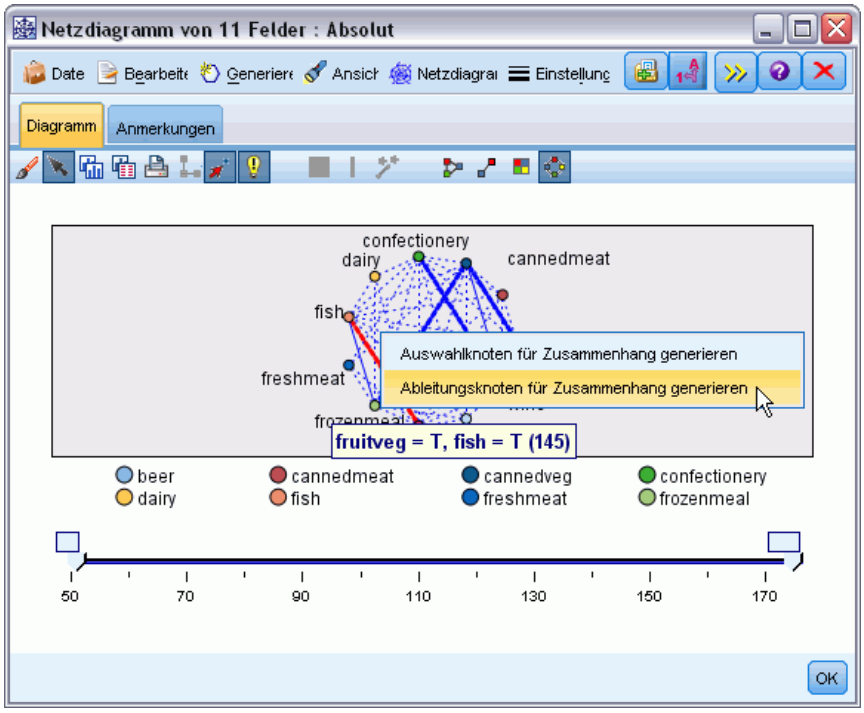
Profilerstellung der Kundengruppen

Sie haben jetzt drei Gruppen von Kunden basierend auf den Typen von Produkten, die Sie kaufen, identifiziert. Sie möchten aber auch wissen, wer diese Kunden sind, d. h. ihr demografisches Profil kennen. Dieses kann dadurch erreicht werden, dass jedem Kunden für jede dieser Gruppen

ein Flag zugewiesen werden kann und die Regelinduktion (C5.0) zum Erstellen von regelbasierten Profilen dieser Flags verwendet werden kann.

Zunächst müssen Sie für jede Gruppe ein Flag ableiten. Dies kann unter Verwendung der soeben erstellten Netzdiagramm-Anzeige automatisch generiert werden. Klicken Sie mit der rechten Maustaste auf die Verknüpfung für den Zusammenhang zwischen *Obst* und *Fisch*, um sie zu markieren, und klicken Sie dann mit der rechten Maustaste und wählen Sie Ableitungsknoten für Zusammenhang generieren.

Abbildung 28-6
Ableiten eines Flags für eine Kundengruppe



Bearbeiten Sie den resultierenden Ableitungsknoten, um den Namen des Ableitungsfelds in *gesund* zu ändern. Wiederholen Sie die Übung mit dem Zusammenhang aus *Wein* und *Süßwaren*, indem Sie das resultierende Ableitungsfeld als *Wein_Süßw* bezeichnen.

Stellen Sie für die dritte Gruppe (drei Zusammenhänge umfassend) zunächst sicher, dass keine Zusammenhänge ausgewählt sind. Wählen Sie dann alle drei Zusammenhänge im Dreieck *Konservengemüse*, *Bier* und *TK-Fertiggericht* aus, indem Sie die Umschalttaste gedrückt halten, während Sie mit der linken Maustaste klicken. (Sie müssen sich dabei im interaktiven Modus befinden und nicht im Bearbeitungsmodus.) Wählen Sie anschließend aus den Menüs für die Netzdiagramm-Anzeige Folgendes aus:

Erzeugen > Ableitungsknoten (“And”)

Ändern Sie den Namen des resultierenden Ableitungsfelds in *Bier_Bohnen_Pizza*.

Um ein Profil dieser Kundengruppen zu erstellen, verbinden Sie den vorhandenen Typknoten mit diesen drei Ableitungsknoten in Folge und fügen Sie dann einen weiteren Typknoten hinzu. Setzen Sie im neuen Typknoten die Rolle aller Felder auf *Keine* mit Ausnahme von *Wert*,

Zahlart, *Geschl*, *Hausbesitzer*, *Einkommen* und *Alter*, das auf *Eingabe* gesetzt werden muss, und die entsprechende Kundengruppe (z. B. *Bier_Bohnen_Pizza*), die auf *Ziel* gesetzt werden muss. Fügen Sie einen C5.0-Knoten hinzu, setzen Sie den Ausgabetyp auf Regelmenge und führen Sie den Knoten aus. Das resultierende Modell (für *Bier_Bohnen_Pizza*) enthält ein klares demografisches Profil für diese Kundengruppe:

Regel 1 für T:
falls Geschlecht = M
und Einkommen <= 16.900,
dann T

Dieselbe Methode kann für die anderen Kundengruppenflags angewendet werden, indem sie als Ausgabe im zweiten Typknoten ausgewählt werden. In diesem Zusammenhang kann ein breiterer Bereich von alternativen Profilen unter Verwendung von A Priori anstelle von C5.0 generiert werden; A Priori kann auch verwendet werden, um ein Profil aller Kundengruppenflags gleichzeitig zu erstellen, da keine Beschränkung auf ein einzelnes Ausgabefeld vorhanden ist.

Zusammenfassung

Dieses Beispiel zeigt, wie IBM® SPSS® Modeler zur Ermittlung von Affinitäten bzw. Zusammenhängen in einer Datenbank verwendet werden kann, sowohl durch die Modellierung (mit A Priori) als auch die Visualisierung (mit Netzdiagramm-Anzeige). Diese Zusammenhänge entsprechen den Gruppierungen von Fällen in den Daten und diese Gruppen können detailliert untersucht und anhand einer Modellierung (mit C5.0-Regelmengen) erstellt werden.

Im Einzelhandel können derartige Kundengruppierungen z. B. für Sonderangebote verwendet werden, um die Reaktionsgeschwindigkeit auf direkte Mailing-Aktionen zu verbessern oder um die in einer Zweigstelle auf Lager vorhandene Produktpalette so anzupassen, dass sie den Anforderungen der demografischen Kundenbasis entspricht.

Kapitel

29

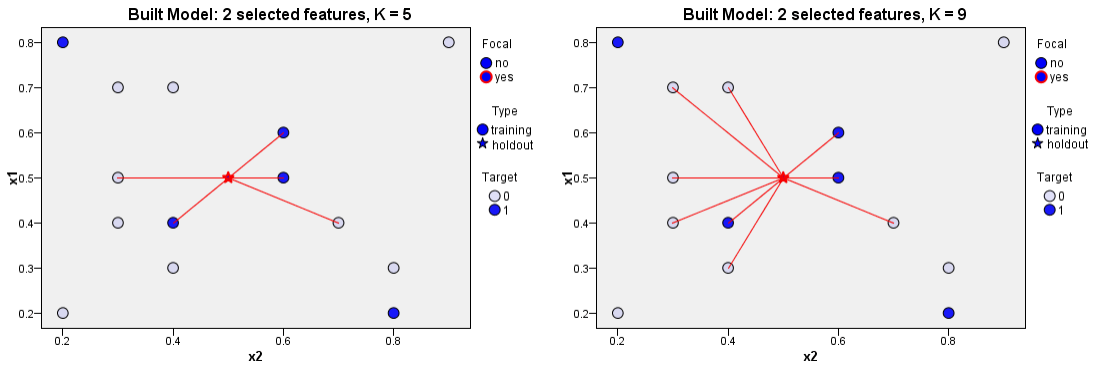
Beurteilen neuer Fahrzeugangebote (KNN)

Die Nächste-Nachbarn-Analyse ist eine Methode zur Klassifizierung von Fällen anhand ihrer Ähnlichkeit zu anderen Fällen. Im Maschinenlernen wurde es entwickelt, um Datenmuster zu erkennen, ohne dass eine exakte Übereinstimmung mit gespeicherten Mustern oder Fällen benötigt wird. Ähnliche Fälle befinden sich nahe beeinander und unterschiedliche Fälle sind voneiander 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 Zahl der zu untersuchenden nächsten Nachbarn festlegen; dieser Wert wird *k* genannt. Die Abbildungen zeigen, wie ein neuer Fall mit Hilfe von zwei verschiedenen Werten von *k* klassifiziert würde. Ist *k* = 5, wird der neue Fall der Kategorie *1* zugeordnet, da der Großteil der nächsten Nachbarn der Kategorie *1* angehört. Ist jedoch *k* = 9, wird der neue Fall der Kategorie *0* zugeordnet, da der Großteil der nächsten Nachbarn der Kategorie *0* angehört.

Abbildung 29-1
Auswirkungen der Änderung von "k" bei der Klassifizierung



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.

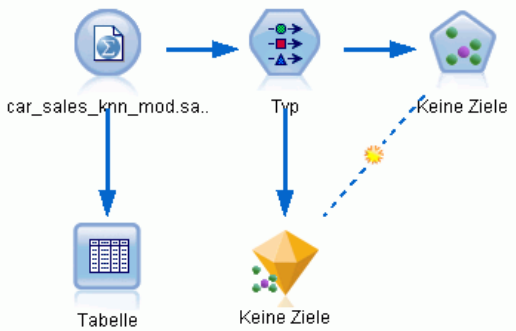
Ein Automobilhersteller hat Prototypen für zwei neue Fahrzeuge entwickelt: einen Personenwagen und einen LKW. Vor der Einführung der neuen Modelle in sein Angebot möchte der Hersteller festsellen, welche bestehenden Fahrzeuge auf dem Markt den Prototypen am ähnlichsten sind, d. h. welche Fahrzeuge ihre "nächsten Nachbarn" und damit die Modelle sind, mit denen sie im Wettbewerb stehen.

Der Hersteller verfügt über eine Datensammlung zu bestehenden Modellen unter einer Reihe von Kategorien, denen er die Details seiner Prototypen hinzugefügt hat. Die Kategorien, unter denen die Modelle verglichen werden sollen, umfassen den Preis in Tausendern (*price*), den Hubraum (*engine_s*), die Pferdestärke (*horsepow*), den Achsabstand (*wheelbas*), die Breite (*width*), die Länge (*length*), das Leergewicht (*curb_wgt*), den Tankinhalt (*fuel_cap*) und die Kraftstoffverwertung (*mpg*).

Für dieses Beispiel wird der Stream *car_sales_knn.str* verwendet, der im Ordner *Demos* unter dem Unterordner *streams* verfügbar ist. Die verwendete Datendatei ist *car_sales_knn_mod.sav*. Für weitere Informationen siehe Thema Ordner “Demos” in Kapitel 1 in *IBM SPSS Modeler 14.2- Benutzerhandbuch*.

Erstellen des Streams

Abbildung 29-2
Beispiel-Stream für KNN-Modellierung

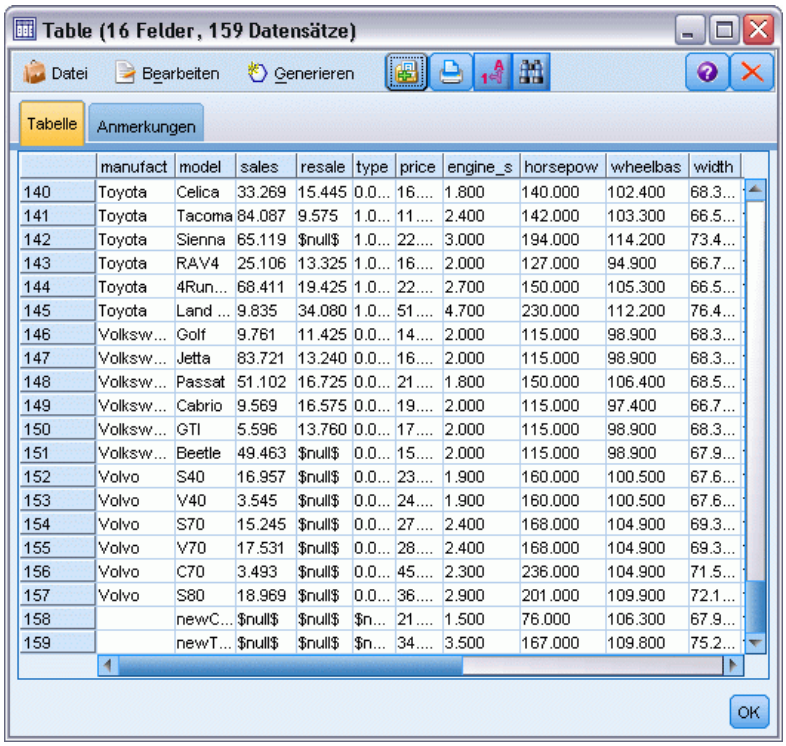


Erstellen Sie einen neuen Stream und fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf die Datei *car_sales_knn_mod.sav* im Verzeichnis *Demos* Ihrer IBM® SPSS® Modeler-Installation verweist.

Betrachten wir zunächst die vom Hersteller gesammelten Daten.

- Fügen Sie dem Statistikdatei-Quellenknoten einen Tabellenknoten hinzu.
- Öffnen Sie den Tabellenknoten und klicken Sie auf Ausführen.

Abbildung 29-3
Quellendaten für Autos und LKWs



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

Die Details für die beiden Prototypen *newCar* und *newTruck* wurden an das Ende der Datei gefügt.

Den Quellendaten lässt sich entnehmen, dass der Hersteller die Klassifizierung von “truck” (Wert 1 in der Spalte *type*) ziemlich großzügig für jede Fahrzeugart verwendet, bei der es sich um keinen PKW handelt.

Die letzte Spalte, *partition*, ist erforderlich, damit die beiden Prototypen als Prüfprofile festgelegt werden können, wenn ihre nächsten Nachbarn identifiziert werden sollen. So beeinflussen ihre Daten die Berechnungen nicht, da wir den übrigen Markt betrachten wollen. Durch Einstellen des *partition*-Wertes der beiden Prüfprofildatensätze auf 1, während alle anderen Datensätze in diesem Feld 0 enthalten, können wir dieses Feld später beim Einstellen der Fokusdatensätze verwenden – das sind die Datensätze, für die wir die nächsten Nachbarn berechnen möchten.

Lassen Sie das Ausgabefenster geöffnet, da wir es später noch brauchen.

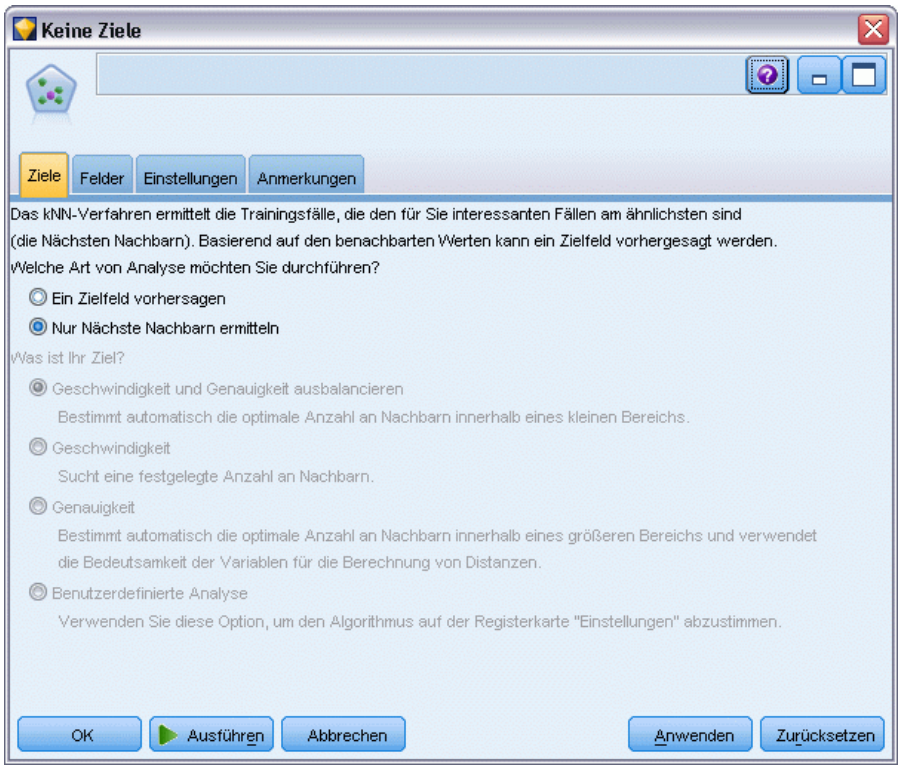
Abbildung 29-4
Typknoteneinstellungen

- Fügen Sie einen Typknoten zum Stream hinzu.
- Fügen Sie dem Statistikdatei-Quellenknoten einen Typknoten hinzu.
- Öffnen Sie den Typknoten.

  Wir wollen nur für die Felder *price* bis *mpg* einen Vergleich durchführen, d. h. wir belassen die Rolle für all diese Felder bei Eingabe.

- Setzen Sie die Rolle für alle anderen Felder (*manufact* bis *type* plus *lnsales*) auf Keine.
- Setzen Sie das Messniveau für das letzte Feld (*partition*) auf Flag. Stellen Sie sicher, dass seine Rolle auf Eingabe eingestellt ist.
- Klicken Sie auf Werte lesen, um die Datenwerte in den Stream einzulesen.
- Klicken Sie auf OK.

Abbildung 29-5
Auswahl zur Identifizierung der nächsten Nachbarn

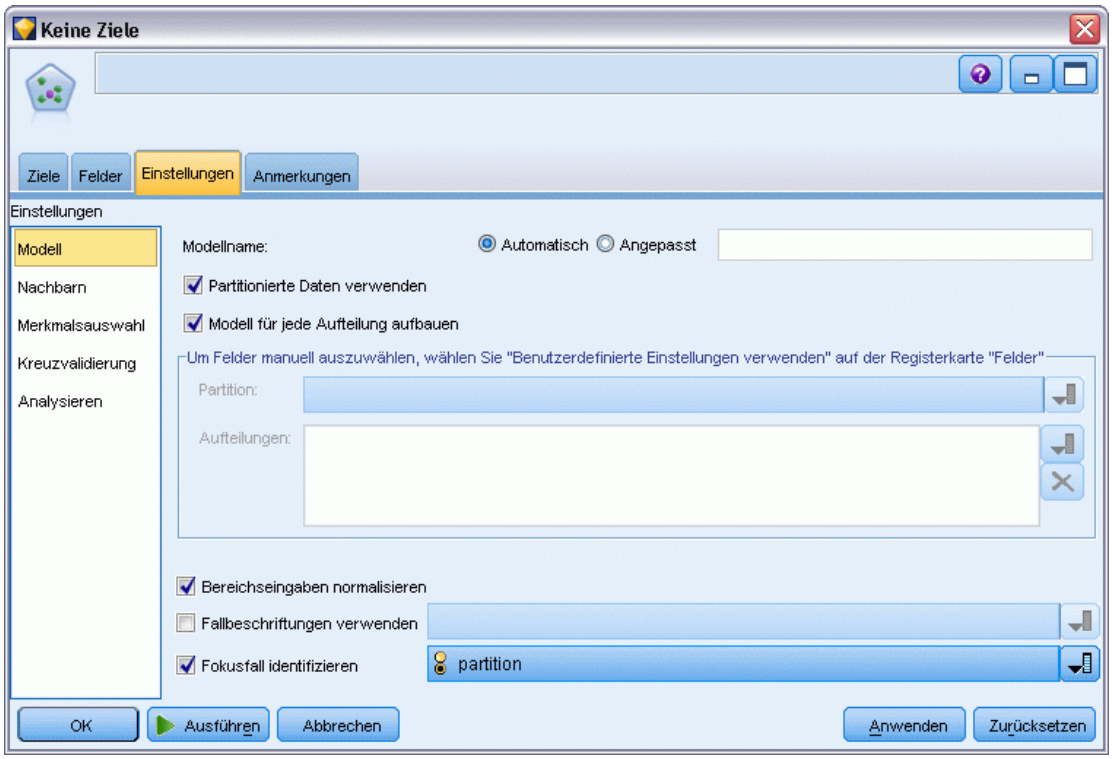


- Verbinden Sie einen KNN-Knoten mit dem Typknoten.
- Öffnen Sie den KNN-Knoten.

  Diesmal werden wir kein Zielfeld vorhersagen, da wir nur die nächsten Nachbarn für unsere beiden Prototypen finden möchten.

- Wählen Sie in der Registerkarte Ziele die Option Nur nächste Nachbarn identifizieren.
- Klicken Sie auf die Registerkarte Einstellungen.

Abbildung 29-6
Identifizieren der Fokusdatensätze mithilfe des Felds "partition"



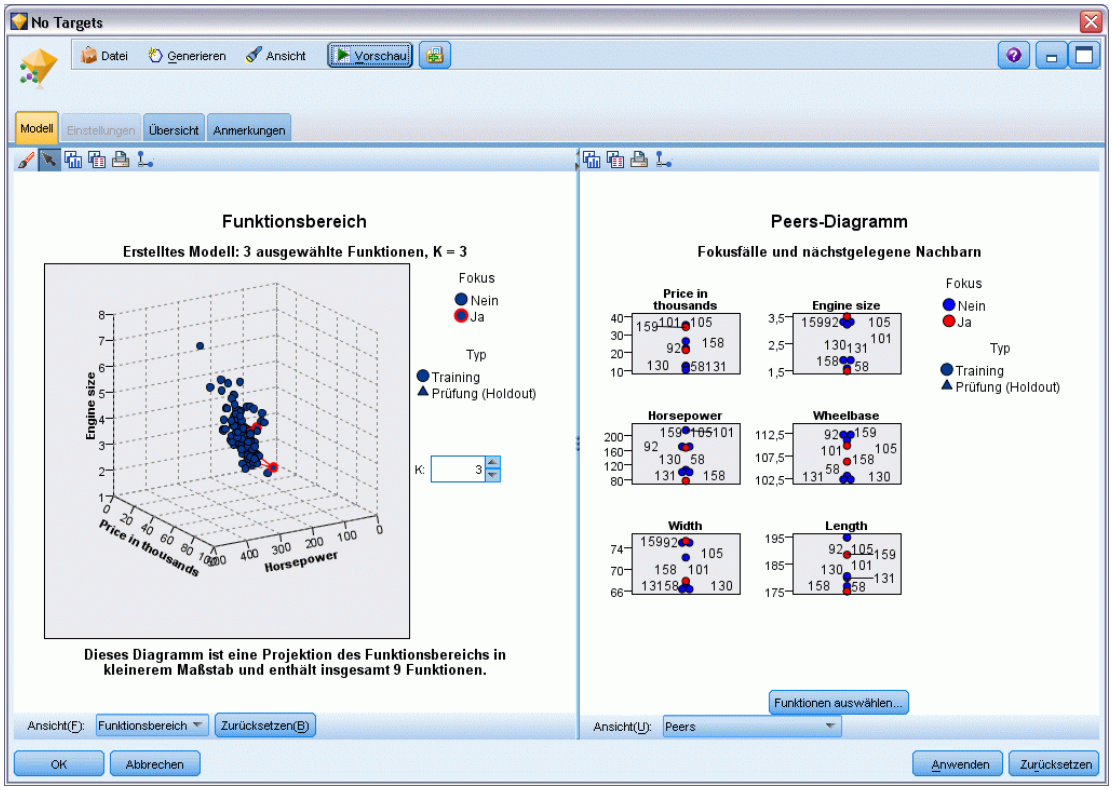
Nun können wir das Feld *partition* verwenden, um die Fokusdatensätze zu identifizieren – das sind die Datensätze, für die wir die nächsten Nachbarn ermitteln möchten. Durch Verwenden eines Flag-Felds stellen wir sicher, dass Datensätze, in denen dieses Feld auf 1 eingestellt ist, zu unseren Fokusdatensätzen werden.

Wir haben bereits gesehen, dass *newCar* und *newTruck* als einzige Datensätze den Wert 1 in diesem Feld enthalten und damit unsere Fokusdatensätze sein werden.

- Aktivieren Sie im Bereich Modell der Registerkarte Einstellungen das Kontrollkästchen Fokusdatensatz identifizieren.
- Wählen Sie partition aus der Dropdown-Liste für dieses Feld.
- Klicken Sie auf die Schaltfläche Ausführen.

Untersuchen der Ausgabe

Abbildung 29-7
Das Fenster "Modell-Viewer"



Ein Modell-Nugget wurde auf der Stream-Zeichenfläche und in der Modellpalette erstellt. Öffnen Sie eines der Nuggets, um die Anzeige des Modell-Viewers zu sehen, die aus einem Fenster mit zwei Bereichen besteht:

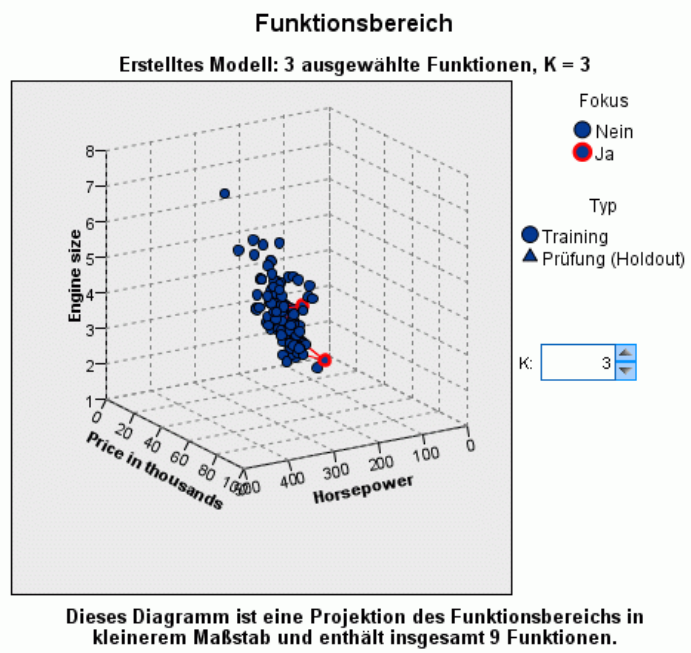
- Im ersten Bereich wird eine Übersicht des Modells, die sogenannte Hauptansicht, angezeigt. Die Hauptansicht für das Nächste-Nachbarn-Modell wird als **Prädiktorbereich** bezeichnet.
- Im zweiten Bereich wird eine der beiden folgenden Ansichten angezeigt:

  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 Sie einen Teil der Hauptansicht ansteuern.

Prädiktorbereich

Abbildung 29-8
Prädiktorbereichsdiagramm



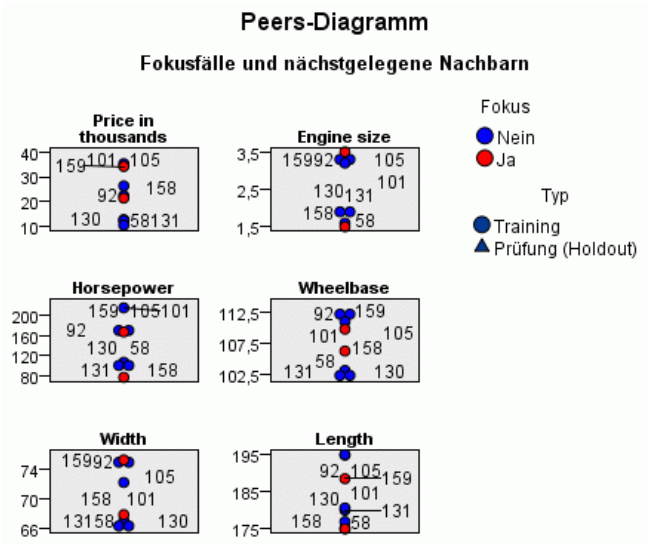
Dieses Prädiktorbereichsdiagramm ist ein interaktives 3D-Diagramm, das Datenpunkte für drei Funktionen zeichnet (d. h. für die ersten drei Eingabefelder der Quelldaten), die Preis, Hubraum und Pferdestärke darstellen.

Unsere beiden Fokusdatensätze werden rot markiert, mit Verbindungslinien zu ihren *k* nächstgelegenen Nachbarn.

Durch Klicken und Ziehen des Diagramms können Sie es drehen, um eine bessere Sicht auf die Verteilung von Punkten im Prädiktorbereich zu erhalten. Klicken Sie auf die Schaltfläche Zurücksetzen, um zur Standardansicht zurückzukehren.

Peers-Diagramm

Abbildung 29-9
Peers-Diagramm



Die Standard-Hilfsansicht ist das Peers-Diagramm, das die beiden im Prädiktorbereich ausgewählten Fokusdatensätze und ihre *k* nächsten Nachbarn für jedes von sechs Merkmalen (die ersten sechs Eingabefelder der Quelldaten) markiert.

Die Fahrzeuge werden durch ihre Datensatznummern in den Quelldaten dargestellt. Hier brauchen wir für die Identifzierung die Ausgabe aus dem Tabellenknoten.

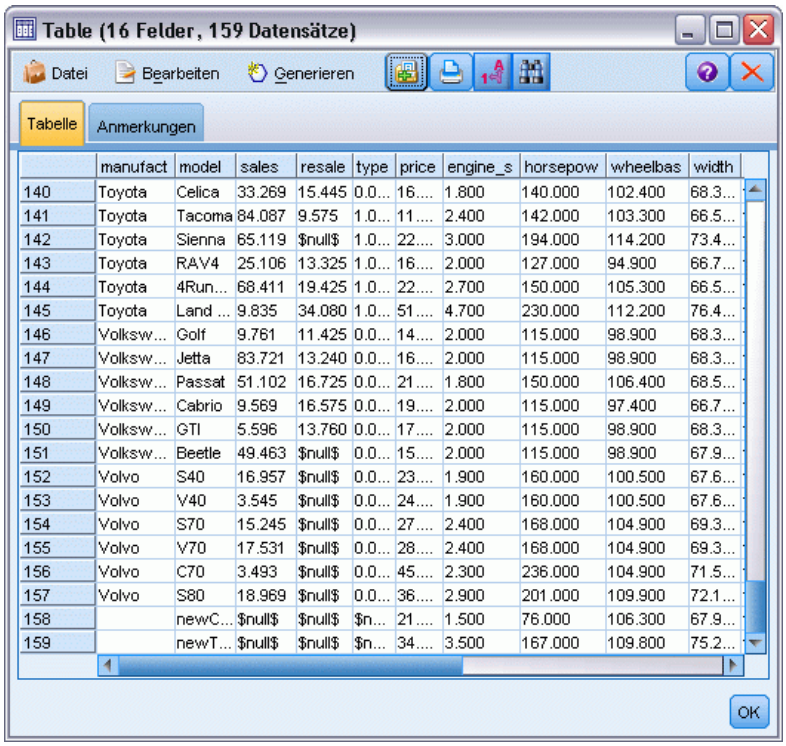
Wenn die Ausgabe des Tabellenknotens noch verfügbar ist:

- Klicken Sie auf die Registerkarte Ausgabe des Manager-Bereichs oben rechts im IBM® SPSS® Modeler-Hauptfenster.
- Doppelklicken Sie auf den Eintrag Tabelle (16 Felder, 159 Datensätze).

Wenn die Ausgabe des Tabellenknotens nicht mehr verfügbar ist:

- Öffnen Sie im SPSS Modeler-Hauptfenster den Tabellenknoten.
- Klicken Sie auf Ausführen.

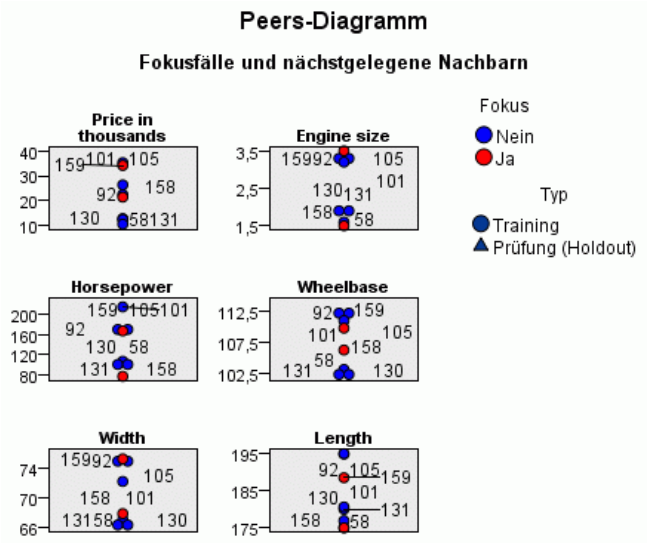
Abbildung 29-10
Identifizieren von Datensätzen nach Datensatznummer



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

Nach einem Bildlauf nach unten zum Tabellenende sehen wir, dass *newCar* und *newTruck* die letzten beiden Datensätze in unseren Daten sind, d. h. die Nummer 158 bzw. 159.

Abbildung 29-11
Vergleich der Funktionen auf dem Peers-Diagramm



Dadurch können wir auf dem Peers-Diagramm beispielsweise sehen, dass *newTruck* (159) über einen größeren Hubraum als alle seine nächsten Nachbarn verfügt, während *newCar* (158) einen kleineren Hubraum als alle *seine* nächsten Nachbarn hat.

Für jedes der sechs Merkmale können Sie die Maus über die einzelnen Punkte führen, um den tatsächlichen Wert jedes Merkmals für diesen speziellen Fall zu sehen.

Aber welche Fahrzeuge sind denn nun die nächsten Nachbarn für *newCar* und *newTruck*?

Das Peers-Diagramm ist ein wenig unübersichtlich, deshalb wollen wir zu einer einfacheren Ansicht wechseln.

- Klicken Sie auf die Dropdown-Liste Ansicht am unteren Rand des Peers-Diagramms (auf den Eintrag, der gerade Peers lautet).
- Wählen Sie Nachbar und Abstandstabelle.

Nachbar und Abstandstabelle

Abbildung 29-12
Nachbar und Abstandstabelle

k Nächstgelegene Nachbarn und Abstände

Anzeige für Anfangsfokusfälle

Fokusfall	Nächstgelegene Nachbarn			Kürzeste Abstände		
	1	2	3	1	2	3
158	131	130	58	0,979	0,990	1,011
159	105	92	101	0,580	0,634	0,644

Das sieht schon besser aus. Nun können wir die drei Modelle sehen, denen unsere beiden Prototypen auf dem Markt am ähnlichsten sind.

Für *newCar* (Fokusdatensatz 158) sind dies Saturn SC (131), Saturn SL (130) und Honda Civic (58).

Keine große Überraschung – alle drei sind Mittelklassewagen, also sollte sich *newCar* gut einfügen, insbesondere mit seiner ausgezeichneten Kraftstoffverwertung.

Für *newTruck* (Fokusdatensatz 159) sind die nächsten Nachbarn Nissan Quest (105), Mercury Villager (92) und die Mercedes M-Klasse (101).

Wie bereits früher gesehen handelt es sich dabei nicht um LKWs im herkömmlichen Sinn, sondern einfach um Fahrzeuge, die als Nicht-Automobil klassifiziert wurden. Bei Betrachtung der Tabellenknotenausgabe für die nächsten Nachbarn zeigt sich, dass *newTruck* relativ teuer und auch der schwerste seines Typs ist. Jedoch ist die Kraftstoffverwertung wieder besser als die seiner stärksten Konkurrenten, dies sollte also zu seinen Gunsten zählen.

Zusammenfassung

Nun haben Sie gesehen, wie Sie mithilfe der Nächste-Nachbarn-Analyse ein breites Spektrum an Funktionen in Fällen aus einem bestimmten Datenset vergleichen können. Zudem wurden für zwei sehr unterschiedliche Prüfprofildatensätze die Fälle berechnet, die diesen Prüfprofilen am ähnlichsten sind.

Anhang

A

Hinweise

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

Der folgende Absatz gilt nicht für Großbritannien oder andere Länder, in denen derartige Bestimmungen nicht mit dem dort geltenden Recht vereinbar sind. SPSS INC., AN IBM COMPANY, ÜBERNIMMT FÜR DIE VORLIEGENDE DOKUMENTATION KEINERLEI GEWÄHRLEISTUNG IRGENDWELCHER ART, WEDER AUSDRÜCKLICH NOCH STILLSCHWEIGEND, EINSCHLIESSLICH (JEDOCH NICHT DARAUF BEGRENZT) DER STILLSCHWEIGENDEN GEWÄHRLEISTUNGEN IN BEZUG AUF DIE NICHTVERLETZUNG VON RECHTEN DRITTER, AUF HANDELSÜBLICHKEIT ODER DIE EIGNUNG FÜR EINEN BESTIMMTEN ZWECK. Einige Staaten lassen bei bestimmten Transaktionen keine Ausschlussklauseln ausdrücklicher oder stillschweigender Gewährleistungen zu, sodass diese Erklärung möglicherweise nicht auf Sie zutrifft.

Diese Informationen können technische Ungenauigkeiten oder typografische Fehler enthalten. An den hierin enthaltenen Informationen werden in regelmäßigen Abständen Änderungen vorgenommen, die in spätere Ausgaben der Publikation eingearbeitet werden. SPSS Inc. kann jederzeit ohne Vorankündigung Verbesserungen und/oder Veränderungen an den in dieser Publikation beschriebenen Produkten und/oder Programmen vornehmen.

Alle in diesen Ausführungen enthaltenen Verweise auf Websites, die nicht zu SPSS bzw. IBM gehören, dienen lediglich der Information. Die Nennung bedeutet nicht, dass SPPS bzw. IBM den Inhalt dieser Websites unterstützen. Das Material auf diesen Websites ist kein Bestandteil des Materials für dieses SPSS Inc.-Produkt. Sie verwenden diese Websites auf eigenes Risiko.

Wenn Sie Informationen an IBM bzw. SPSS senden, räumen Sie IBM und SPSS das nicht ausschließliche Recht ein, die Informationen in jeglicher Form zu verwenden bzw. weiterzugeben, die dem Unternehmen geeignet erscheint, ohne dass ihm daraus Verbindlichkeiten Ihnen gegenüber entstehen.

Informationen zu Nicht-SPSS-Produkten stammen von den Herstellern dieser Produkte, ihren veröffentlichen Verlautbarungen oder aus anderen öffentlich verfügbaren Quellen. SPSS hat diese Produkte nicht getestet und kann daher die Richtigkeit der Angaben zu Leistung und Kompatibilität oder anderer Behauptungen in Bezug auf Nicht-SPSS-Produkte nicht bestätigen. Fragen zu den Fähigkeiten von Nicht-SPSS-Produkten sind an die Hersteller dieser Produkte zu richten.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Diese Informationen enthalten Beispiele für Daten und Berichte, die in alltäglichen Betriebsabläufen verwendet werden. Um sie möglichst umfassend darzulegen, enthalten die Beispiele Namen von Einzelpersonen, Unternehmen, Marken und Produkten. Alle diese Namen sind frei erfunden und jegliche Ähnlichkeit mit Namen und Adressen, die von einem tatsächlichen Handelsunternehmen verwendet werden, ist rein zufällig.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Marken

IBM, das IBM-Logo und ibm.com sind Marken von IBM Corporation, die in vielen Ländern weltweit eingetragen sind. Eine aktuelle Liste der IBM-Marken finden Sie im Internet unter *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS ist eine Marke von SPSS Inc., an IBM Company, die in vielen Ländern weltweit eingetragen sind.

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind entweder registrierte Marken oder Marken von Adobe Systems Incorporated in den USA und/oder anderen Ländern.

IT Infrastructure Library ist eine eingetragene Marke der Central Computer and Telecommunications Agency, die nun zum Office of Government Commerce gehört.

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 Marken oder eingetragene Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und/oder anderen Ländern.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken von Microsoft Corporation in den USA und/oder anderen Ländern.

ITIL ist eine eingetragene Marke und eine eingetragene Gemeinschaftsmarke des Office of Government Commerce und ist beim U.S. Patent and Trademark Office eingetragen.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Cell Broadband Engine ist eine Marke von Sony Computer Entertainment, Inc. in den USA und/oder anderen Ländern und wird im Rahmen einer Lizenz dieses Unternehmens verwendet.

Java und alle Java-basierten Marken und Logos sind Marken von Sun Microsystems, Inc. in den USA und/oder anderen Ländern.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Weitere Produkt- oder Servicenamen können Marken von IBM, SPSS oder anderen Unternehmen sein.

Bibliografie

Asuncion, A., als auch D. Newman. 2007. "UCI Machine Learning Repository." Available at http://mlearn.ics.uci.edu/MLRepository.html.

Index