

*IBM SPSS Modeler 16
Anwendungshandbuch*

IBM

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen“ auf Seite 347 gelesen werden.

Produktinformation

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

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Modeler 16, Applications Guide,
herausgegeben von International Business Machines Corporation, USA

© Copyright International Business Machines Corporation 2013

Informationen, die nur für bestimmte Länder Gültigkeit haben und für Deutschland, Österreich und die Schweiz nicht zutreffen, wurden in dieser Veröffentlichung im Originaltext übernommen.

Möglicherweise sind nicht alle in dieser Übersetzung aufgeführten Produkte in Deutschland angekündigt und verfügbar; vor Entscheidungen empfiehlt sich der Kontakt mit der zuständigen IBM Geschäftsstelle.

Änderung des Textes bleibt vorbehalten.

Herausgegeben von:
TSC Germany
Kst. 2877
Oktober 2013

Inhaltsverzeichnis

Kapitel 1. Informationen zu IBM SPSS

Modeler	1
IBM SPSS Modeler-Produkte	1
IBM SPSS Modeler	1
IBM SPSS Modeler Server	1
IBM SPSS Modeler Administration Console	2
IBM SPSS Modeler Batch	2
IBM SPSS Modeler Solution Publisher	2
IBM SPSS Modeler Server-Adapter für IBM SPSS Collaboration and Deployment Services	2
IBM SPSS Modeler-Editionen	2
IBM SPSS Modeler-Dokumentation	3
SPSS Modeler Professional-Dokumentation	3
SPSS Modeler Premium-Dokumentation	4
Anwendungsbeispiele	5
Ordner "Demos"	5

Kapitel 2. IBM SPSS Modeler - Übersicht 7

Erste Schritte	7
Starten von IBM SPSS Modeler	7
Starten über die Befehlszeile	7
Verbinden mit IBM SPSS Modeler Server	8
Ändern des temporären Verzeichnisses	10
Starten mehrerer IBM SPSS Modeler-Sitzungen	10
Schnittstelle von IBM SPSS Modeler auf einen Blick	11
Streamerstellungsbereich von IBM SPSS Modeler	11
Knotenpalette.	12
IBM SPSS Modeler-Manager.	13
IBM SPSS Modeler-Projekte	15
IBM SPSS Modeler-Symbolleiste	15
Anpassen der Symbolleiste	16
Anpassen des IBM SPSS Modeler-Fensters	17
Ändern der Symbolgröße für einen Stream	18
Verwenden der Maus in IBM SPSS Modeler	18
Verwenden von Tastenkombinationen	19
Drucken	20
Automatisieren von IBM SPSS Modeler	20

Kapitel 3. Einführung in die Modellierung 21

Erstellen des Streams	22
Durchsuchen des Modells	27
Bewerten des Modells	32
Scoren von Datensätzen	35
Zusammenfassung	35

Kapitel 4. Automatische Modellierung für ein Flagziel 37

Modellieren der Kundenreaktion (Automatisches Klassifikationsmerkmal)	37
Historische Daten	37
Erstellen des Streams	38
Generieren und Vergleichen von Modellen	42
Zusammenfassung	47

Kapitel 5. Automatische Modellierung für ein stetiges Ziel 49

Eigenschaftswerte (Auto-Numerisch)	49
Trainingsdaten	49
Erstellen des Streams	50
Vergleichen der Modelle	53
Zusammenfassung	55

Kapitel 6. Automatische Datenvorbereitung (ADP) 57

Erstellen des Streams	57
Vergleichen der Modellgenauigkeit	61

Kapitel 7. Vorbereiten von Daten für die Analyse (Data Audit) 65

Erstellen des Streams	65
Durchsuchen von Statistiken und Diagrammen	68
Umgang mit Ausreißern und fehlenden Werten	70

Kapitel 8. Medikamentöse Behandlung (Explorative Diagramme/C5.0) 75

Einlesen von Textdaten	75
Hinzufügen von Tabellen.	78
Erstellen eines Verteilungsdiagramms.	79
Erstellen eines Streudiagramms.	80
Erstellen eines Netzdiagramms	81
Ableiten neuer Felder	83
Erstellen eines Modells	86
Durchsuchen des Modells	88
Verwenden eines Analyseknötens	89

Kapitel 9. Screening von Prädiktoren (Merkmalauswahl) 91

Erstellen des Streams	91
Erstellen der Modelle	94
Vergleichen der Ergebnisse	95
Zusammenfassung	97

Kapitel 10. Reduzieren der Länge der Zeichenfolge für die Eingabedaten (Umcodierungsknoten) 99

Reduzieren der Länge der Zeichenfolge für die Ein- gabedaten (Umcodierung)	99
Umcodieren der Daten	99

Kapitel 11. Modellieren der Kundenreaktion (Entscheidungsliste) 105

Historische Daten	105
Erstellen des Streams.	106
Erstellen des Modells.	108
Berechnen von benutzerdefinierten Maßen mithilfe von Excel.	121

Ändern der Excel-Vorlage	127
Speichern der Ergebnisse	129

Kapitel 12. Klassifizieren von Kunden im Telekommunikationsbereich (multi-nomiale logistische Regression)	131
Erstellen des Streams	131
Durchsuchen des Modells	135

Kapitel 13. Kundenabwanderung bei Telekommunikationsunternehmen (bi-nomiale logistische Regression)	139
Erstellen des Streams	139
Durchsuchen des Modells	145

Kapitel 14. Vorhersage der Bandbreitennutzung (Zeitreihen)	151
Vorhersageerstellung mit dem Zeitreihenknoten	151
Erstellen des Streams	152
Untersuchen der Daten	153
Definieren der Datumswerte	156
Definieren der Ziele	158
Festlegen der Zeitintervalle	159
Erstellen des Modells	161
Untersuchen des Modells	163
Zusammenfassung	171
Erneutes Anwenden eines Zeitreihenmodells	171
Abrufen des Streams	172
Abrufen des gespeicherten Modells	173
Generieren eines Modellknotens	173
Generieren eines neuen Modells	174
Untersuchen des neuen Modells	176
Zusammenfassung	178

Kapitel 15. Vorhersage von Katalogverkäufen (Zeitreihen)	179
Erstellen des Streams	179
Untersuchen der Daten	182
Exponentielles Glätten	183
ARIMA (X11 ARIMA)	187
Zusammenfassung	192

Kapitel 16. Erstellen von Angeboten für Kunden (Selbstlernfunktion)	193
Erstellen des Streams	194
Durchsuchen des Modells	198

Kapitel 17. Vorhersage von Kreditausfällen (Bayes-Netz)	205
Erstellen des Streams	205
Durchsuchen des Modells	209

Kapitel 18. Erneutes Trainieren eines Modells auf monatlicher Basis (Bayes-Netz)	213
Erstellen des Streams	213
Bewerten des Modells	217

Kapitel 19. Werbeaktion für Einzelhandelsumsatz (Netz/C&RT)	225
Untersuchen der Daten	225
Lernen und Testen	227

Kapitel 20. Bedingungsüberwachung (Netz/C5.0)	229
Untersuchen der Daten	230
Datenvorbereitung	231
Lernen	232
Testen	233

Kapitel 21. Klassifizieren von Kunden im Telekommunikationsbereich (Diskriminanzanalyse)	235
Erstellen des Streams	235
Untersuchen des Modells	240
Ausgabeanalyse für die Verwendung von Diskriminanzanalysen hinsichtlich der Klassifizierung von Telekommunikationskunden	241
Zusammenfassung	245

Kapitel 22. Analysieren von intervall-zensierten Überlebensdaten (verallgemeinerte lineare Modelle)	247
Erstellen des Modells	247
Tests der Modelleffekte	251
Anpassen des Modells "Nur Behandlung"	252
Parameterschätzungen	253
Vorhergesagtes erneutes Auftreten und Überlebenswahrscheinlichkeiten	254
Modellieren der Wahrscheinlichkeit eines erneuten Auftretens nach Zeitraum	258
Tests der Modelleffekte	263
Anpassen des verkürzten Modells	263
Parameterschätzungen	264
Vorhergesagtes erneutes Auftreten und Überlebenswahrscheinlichkeiten	265
Zusammenfassung	269
Verwandte Prozeduren	270
Empfohlene Texte	270

Kapitel 23. Verwenden der Poisson-Regression für die Analyse von Schiffsschadensraten (verallgemeinerte lineare Modelle)	271
Anpassen einer Poisson-Regression mit Überdispersion	271
Statistik für Anpassungsgüte	275
Omnibus-Test	275
Tests der Modelleffekte	276
Parameterschätzungen	276
Anpassen alternativer Modelle	277
Statistik für Anpassungsgüte	279
Zusammenfassung	280
Verwandte Prozeduren	280
Empfohlene Texte	280

Kapitel 24. Anpassen einer Gammaregression an Versicherungsforderungen an Kfz-Versicherungen (verallgemeinerte lineare Modelle)	281
Erstellen des Modells	281
Parameterschätzungen	285
Zusammenfassung	285
Verwandte Prozeduren	285
Empfohlene Texte	286
Kapitel 25. Klassifizieren von Zellproben (SVM)	287
Erstellen des Streams	288
Untersuchen der Daten	292
Versuch mit einer anderen Funktion	294
Vergleichen der Ergebnisse	295
Zusammenfassung	296
Kapitel 26. Verwenden der Cox-Regression zur Modellierung der Zeit bis zur Kundenabwanderung	297
Erstellen eines geeigneten Modells	297
Zensierte Fälle	300
Codierungen für kategoriale Variablen	301
Variablenauswahl	302
Mittelwerte von Kovariaten	304
Überlebenskurve	305

Hazard-Kurve	306
Evaluation	306
Verfolgung der erwarteten Anzahl an Kunden, die gehalten werden können	311
Scoring	323
Zusammenfassung	328

Kapitel 27. Warenkorbanalyse (Regelinduktion/C5.0)	329
Datenzugriff	329
Entdecken von Affinitäten beim Warenkorbinhalt	331
Profilerstellung der Kundengruppen	334
Zusammenfassung	335

Kapitel 28. Beurteilen neuer Fahrzeugangebote (KNN)	337
Erstellen des Streams	338
Untersuchen der Ausgabe	342
Prädiktorbereich	343
Peerdiagramm	344
Nachbar und Abstandstabelle	346
Zusammenfassung	346

Bemerkungen	347
Marken	348

Index	349
--------------	------------

Kapitel 1. Informationen zu IBM SPSS Modeler

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

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

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

IBM SPSS Modeler-Produkte

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

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

IBM SPSS Modeler

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

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

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

IBM SPSS Modeler Server

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

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

IBM SPSS Modeler Administration Console

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

IBM SPSS Modeler Batch

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

IBM SPSS Modeler Solution Publisher

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

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

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

IBM SPSS Modeler-Editionen

SPSS Modeler ist in den folgenden Editionen erhältlich.

SPSS Modeler Professional

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

SPSS Modeler Premium

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

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

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

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

IBM SPSS Modeler-Dokumentation

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

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

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

SPSS Modeler Professional-Dokumentation

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

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

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

SPSS Modeler Premium-Dokumentation

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

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

Anwendungsbeispiele

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

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

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

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

Ordner "Demos"

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

Kapitel 2. IBM SPSS Modeler - Übersicht

Erste Schritte

Als Data-Mining-Anwendung bietet IBM SPSS Modeler eine strategische Methode zum Auffinden nützlicher Beziehungen in großen Datasets. 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 Modeler 16 > IBM SPSS Modeler 16

Nach einigen Sekunden wird das Hauptfenster angezeigt.

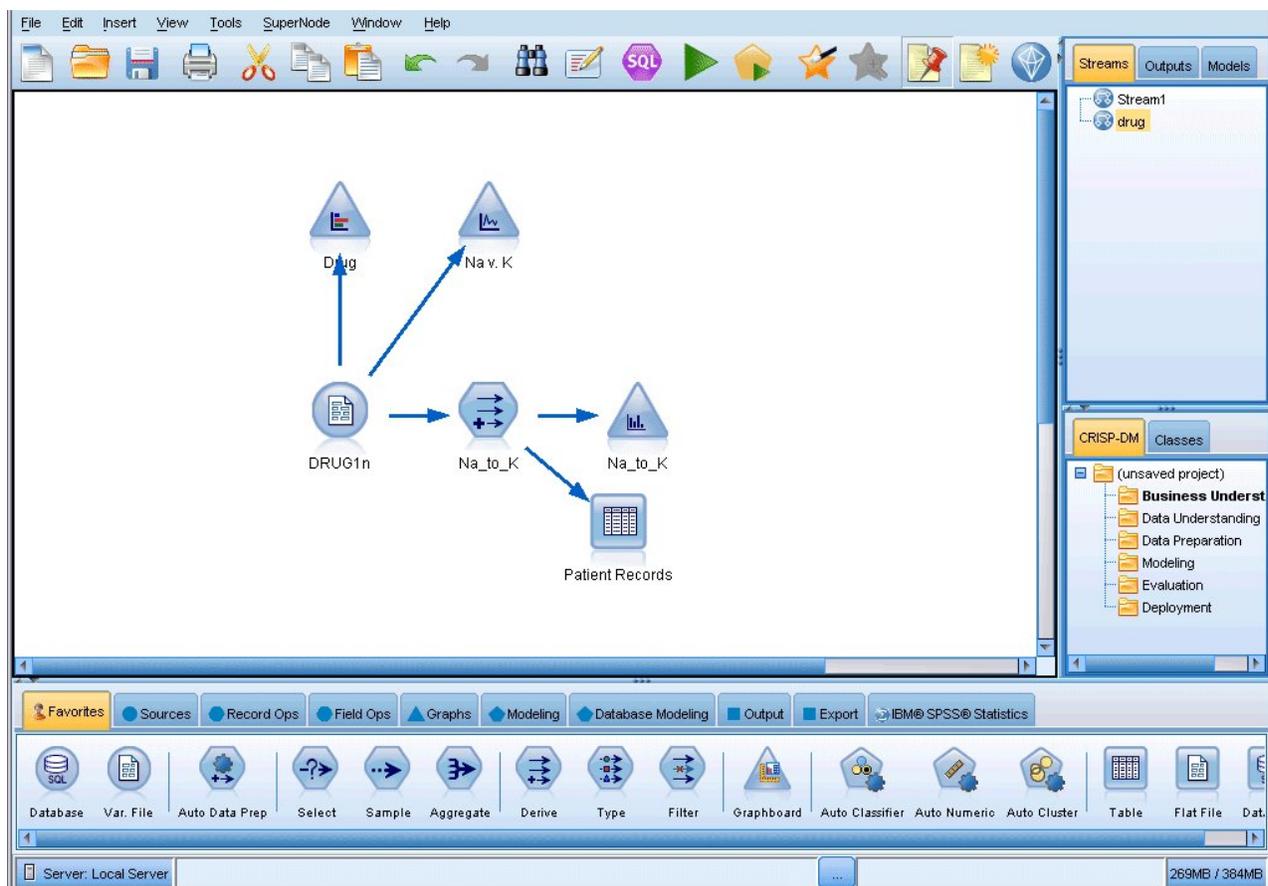


Abbildung 1. Hauptanwendungsfenster von IBM SPSS Modeler

Starten über die Befehlszeile

Sie können IBM SPSS Modeler wie folgt über die Befehlszeile Ihres Betriebssystems starten:

1. Öffnen Sie auf einem Computer, auf dem IBM SPSS Modeler installiert ist, ein DOS- oder Befehlszeilenfenster.
2. Um die IBM SPSS Modeler-Schnittstelle im interaktiven Modus zu starten, geben Sie den Befehl `modelerclient` gefolgt von den erforderlichen Argumenten ein. Beispiel:

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

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

Verbinden 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 Plug-in Coordinator of Processes von IBM SPSS Collaboration and Deployment Services mit einer IBM SPSS Modeler Server-Instanz oder einem Server-Cluster verbunden ist. Der aktuelle Verbindungsstatus wird unten links im IBM 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 Netz ausgeführten Statistics-Dienste zu durchsuchen, wird über den Coordinator of Processes bereitgestellt.

So stellen Sie eine Verbindung mit einem Server her:

1. 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 IBM SPSS Modeler-Fenster doppelklicken.
2. Legen Sie in diesem Dialogfeld die Optionen zum Verbinden mit dem lokalen Server-Computer fest oder wählen Sie eine Verbindung in der Tabelle aus.
 - Klicken Sie auf **Hinzufügen** bzw. **Bearbeiten**, um eine Verbindung hinzuzufügen bzw. zu bearbeiten. Weitere Informationen finden Sie unter im Thema „Hinzufügen und Bearbeiten der IBM SPSS Modeler Server-Verbindung“ auf Seite 9.
 - Klicken Sie auf **Suche**, um auf einen Server bzw. Server-Cluster in Coordinator of Processes zuzugreifen. Weitere Informationen finden Sie unter im Thema „Suchen nach Servern in IBM SPSS Collaboration and Deployment Services“ auf Seite 9.

Servertabelle. Diese Tabelle enthält die Menge der definierten Serververbindungen. In der Tabelle werden die Standardverbindung, der Servername sowie die 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".

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

Berechtigungs nachweise einstellen. Lassen Sie dieses Kontrollkästchen unausgewählt, um die Funktion für **Single Sign-on** zu aktivieren. Diese versucht, Sie mithilfe Ihres lokalen Benutzernamens und Kennworts beim Server anzumelden. Falls ein Single Sign-on nicht möglich ist oder Sie das Kontrollkästchen zur Inaktivierung des Single Sign-on aktivieren (z. B. zur Anmeldung an einem Administratorkonto), wird ein weiteres Fenster angezeigt, in dem Sie aufgefordert werden, Ihre Berechtigungs nachweise einzugeben.

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

Kennwort. Geben Sie das Kennwort 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.

3. Klicken Sie auf **OK**, um die Verbindung herzustellen.

So trennen Sie eine Verbindung mit einem Server:

1. 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 IBM SPSS Modeler-Fenster doppelklicken.
2. 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 über 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.

So fügen Sie Serververbindungen hinzu:

1. Klicken Sie im Menü "Extras" auf die Option **Anmelden beim Server**. Das Dialogfeld "Anmelden beim Server" wird geöffnet.
 2. Klicken Sie in diesem Dialogfeld auf **Hinzufügen**. Das Dialogfeld "Anmeldung beim Server: Server hinzufügen/bearbeiten" wird angezeigt.
 3. 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 Netz 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:

1. Klicken Sie im Menü "Extras" auf die Option **Anmelden beim Server**. Das Dialogfeld "Anmelden beim Server" wird geöffnet.
2. 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.
3. 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 Netz 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.

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 Kennwort.

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. Weitere Informationen finden Sie im Thema „Hinzufügen und Bearbeiten der IBM SPSS Modeler Server-Verbindung“ auf Seite 9.

So suchen Sie nach Servern und Clustern:

1. Klicken Sie im Menü "Extras" auf die Option **Anmelden beim Server**. Das Dialogfeld "Anmelden beim Server" wird geöffnet.
2. 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.
3. Wählen Sie den Server bzw. Server-Cluster in der Liste aus.
4. Klicken Sie auf **OK**, um das Dialogfeld zu schließen und diese Verbindung der Tabelle im Dialogfeld "Anmelden beim Server" hinzuzufügen.

Ändern des temporären Verzeichnisses

Für einige von IBM SPSS Modeler Server durchgeführte Operationen müssen möglicherweise 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:

1. Erstellen Sie ein neues Verzeichnis namens *spss* und ein Unterverzeichnis namens *servoertemp*.
2. Bearbeiten Sie die Datei *options.cfg*, die sich im Unterverzeichnis */config* Ihres IBM SPSS Modeler-Installationsverzeichnisses befindet. Bearbeiten Sie den Parameter *temp_directory* in diesem Feld, damit er wie folgt lautet: *temp_directory, "C:/spss/servoertemp"*.
3. Anschließend müssen Sie den IBM 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 Systems wird auch der Dienst neu gestartet.

Alle temporären Dateien werden nun in dieses neue Verzeichnis geschrieben.

Hinweis: Der häufigste Fehler bei diesem Vorgang besteht in der Verwendung des falschen Schrägstrichs. Zu verwenden ist der normale Schrägstrich.

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 von demselben Client-Computer aus mit zwei verschiedenen Servern ausführen möchten.

So aktivieren Sie mehrere IBM SPSS Modeler-Sitzungen:

1. Klicken Sie auf:
Start > [Alle] Programme > IBM SPSS Modeler 16
2. Klicken Sie mit der rechten Maustaste auf die Verknüpfung "IBM SPSS Modeler 16" (die mit dem Symbol) und wählen Sie die Option **Eigenschaften** aus.
3. Fügen Sie im Textfeld **Ziel** die Option *-noshare* am Ende der Zeichenfolge hinzu.
4. Wählen Sie im Windows Explorer folgende Optionen aus:

Extras > Ordneroptionen...

5. Wählen Sie auf der Registerkarte "Dateitypen" die Option für IBM SPSS Modeler-Streams aus und klicken Sie auf **Erweitert**.
6. Wählen Sie im Dialogfeld "Dateityp bearbeiten" die Option "Öffnen mit IBM SPSS Modeler" aus und klicken Sie auf **Bearbeiten**.
7. Fügen Sie im Textfeld **Anwendung für diesen Vorgang** vor dem Argument **-stream** das Argument **-noshare** ein.

Schnittstelle von IBM SPSS Modeler auf einen Blick

An jedem Punkt des Data-Mining-Prozesses können Sie über die benutzerfreundliche Benutzerschnittstelle 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 IBM SPSS Modeler besteht aus drei Schritten.

- Zunächst lesen Sie Daten in IBM SPSS Modeler ein.
- Anschließend unterziehen Sie die Daten einer Reihe von Bearbeitungen.
- Schließlich senden Sie die Daten an ein Ziel.

Diese Reihenfolge wird als **Datenstream** 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.

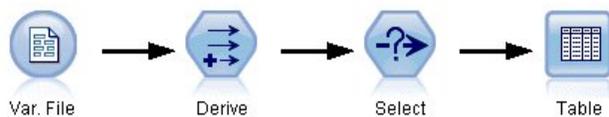


Abbildung 2. Ein einfacher Stream

Streamerstellungsbereich von IBM SPSS Modeler

Der Streamerstellungsbereich ist der größte Bereich des IBM SPSS Modeler-Fensters. Hier erstellen und bearbeiten Sie Datenstreams.

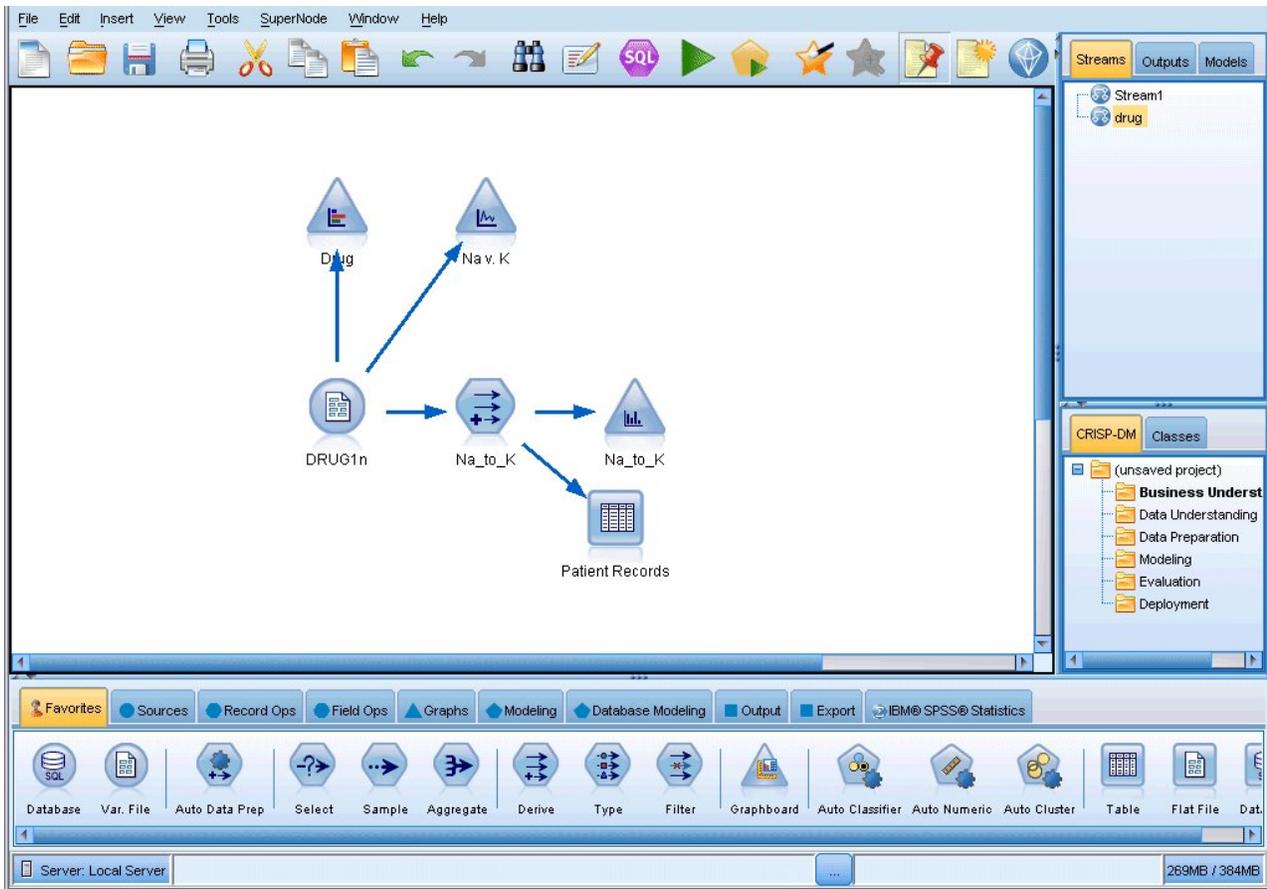


Abbildung 3. IBM SPSS Modeler-Arbeitsbereich (Standardansicht)

Die Erstellung von Streams erfolgt, indem im Haupterstellungsbereich der Benutzerschnittstelle Diagramme gezeichnet werden, in denen die für Ihren Geschäftsbetrieb relevanten Datenoperationen enthalten sind. 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 IBM SPSS Modeler mit mehreren Streams gleichzeitig arbeiten, entweder in demselben Streamerstellungsbereich oder durch Öffnen eines neuen Streamerstellungsbereichs. Während einer Sitzung werden die Streams im Stream-Manager (rechts oben im IBM SPSS Modeler-Fenster) gespeichert.

Knotenpalette

Die meisten Daten und Modellierungstools in IBM SPSS Modeler befinden sich in der **Knotenpalette** (unten im Fenster) unterhalb des Streamerstellungsbereichs.

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 Erstellungsbereich hinzuzufügen, doppelklicken Sie in den Knotenpaletten auf die entsprechenden Symbole oder ziehen Sie sie auf den Erstellungsbereich. Anschließend verbinden Sie sie, um einen **Stream** zu erstellen, der den Datenfluss darstellt.

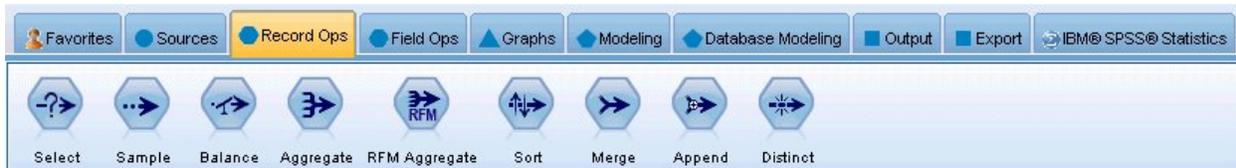


Abbildung 4. Registerkarte "Datensatzoperationen" in der Knotenpalette

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

- **Quellen.** Diese Knoten lesen Daten in IBM 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 IBM SPSS Modeler verfügbaren Modellierungsalgorithmen, wie neuronale Netze, Entscheidungsbäume, Clusteralgorithmen und Datensequenzierung.
- **Datenbankmodellierung.** Diese Knoten verwenden die Modellierungsalgorithmen, die in Microsoft SQL Server-, IBM DB2-, Oracle- und Netezza-Datenbanken verfügbar sind.
- **Ausgabe.** Diese Knoten erzeugen verschiedenste Ausgabetypen für Daten, Diagramme und Modellergebnisse, die in IBM SPSS Modeler angezeigt werden können.
- **Export.** Diese Knoten erzeugen verschiedenste Ausgabetypen, die in externen Anwendungen angezeigt werden können, z. B. in IBM SPSS Data Collection oder Excel.
- **IBM SPSS Statistics.** Diese Knoten importieren Daten aus IBM SPSS Statistics oder exportieren sie nach IBM SPSS Statistics und führen IBM SPSS Statistics-Prozeduren aus.

Je vertrauter Sie im Umgang mit IBM SPSS Modeler werden, desto besser können Sie den Paletteninhalt für Ihre eigene Verwendung anpassen.

Der Berichtsbereich, der sich unter der Knotenpalette befindet, bietet Feedback zum Fortschritt verschiedener Operationen, z. B. wann die Daten in den Datenstream 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 Managerbereich. 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 5. Registerkarte "Streams"



Abbildung 6. Registerkarte "Ausgaben"

Die Registerkarte "Ausgaben" enthält eine Reihe von Dateien, wie beispielsweise Diagramme und Tabellen, die von den Streamoperationen in IBM SPSS Modeler erstellt wurden. Sie können die auf dieser Registerkarte aufgeführten Tabellen, Diagramme und Berichte anzeigen, speichern, umbenennen und schließen.

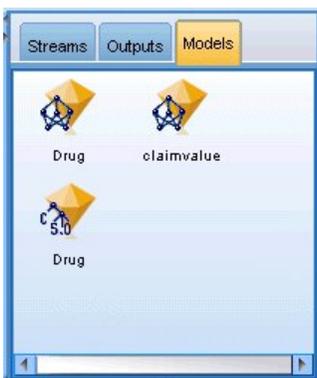


Abbildung 7. Registerkarte "Modelle" mit Modellnuggets

Die Registerkarte "Modelle" ist die leistungsstärkste der Manager-Registerkarten. Diese Registerkarte enthält sämtliche **Modellnuggets**, die die in IBM SPSS Modeler erstellten Modelle enthalten, für die aktuelle Sitzung. Auf der Registerkarte "Modelle" können diese Modelle direkt durchsucht oder dem Stream im Erstellungsbereich hinzugefügt werden.

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".

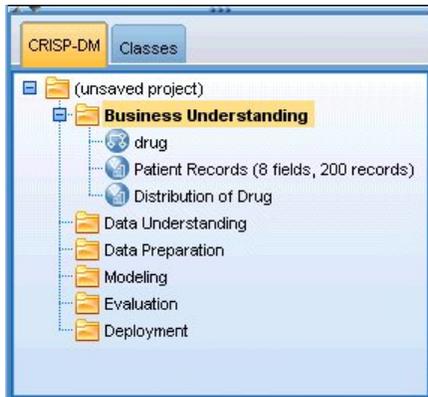


Abbildung 8. Ansicht "CRISP-DM"

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 Daten-Mining-Experten als auch Neulinge können mit dem CRISP-DM-Tool ihre Arbeit besser organisieren und an andere weitergeben.

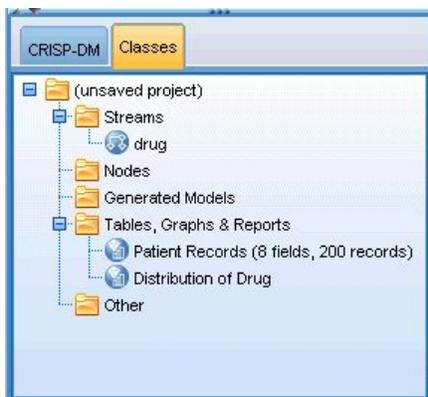


Abbildung 9. Ansicht "Klassen"

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

IBM SPSS Modeler-Symboleiste

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



Neuen Stream erstellen



Stream öffnen

	Stream speichern		Aktuellen Stream drucken
	Ausschneiden & in die Zwischenablage verschieben		In Zwischenablage kopieren
	Auswahl einfügen		Letzte Aktion rückgängig machen
	Wiederholen		Knoten suchen
	Streameigenschaften bearbeiten		Vorschau für SQL-Erzeugung
	Aktuellen Stream ausführen		Streamauswahl ausführen
	Stream anhalten (wird nur während der Streamausführung 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 Streamkommentare, Modellverknüpfungen und die Anzeige von Scoring-Verzweigungen.

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:

1. Klicken Sie im Hauptmenü auf Folgendes:

Ansicht > Symbolleiste > Anzeigen

So ändern Sie die Einstellungen für QuickInfo oder Symbolgröße:

1. 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 Trennlinien zwischen den verschiedenen Bereichen der IBM SPSS Modeler-Benutzerschnittstelle 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 Managerbereichs und des Projektbereichs verwenden. So wird die Größe des Streamerstellungsbereichs 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.

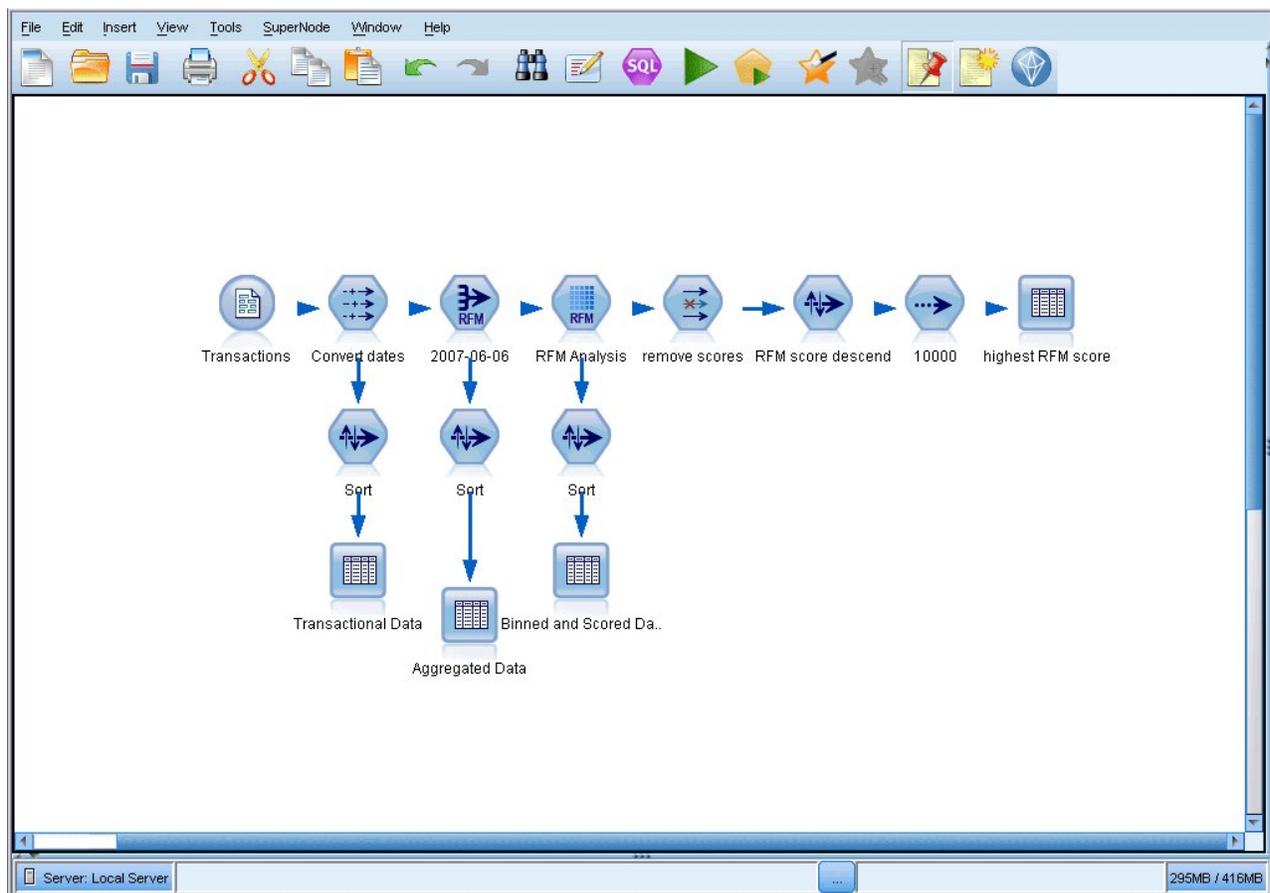


Abbildung 10. Maximierter Streamerstellungsbereich

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

Sie können auch die Anzeige der Anzeigemarkup steuern, die Streamkommentare, Modellverknüpfungen und die Anzeige von Scoring-Verzweigungen umfasst. Um diese Anzeige ein- oder auszublenden, klicken Sie auf:

Ansicht > Stream-Markup

Ändern der Symbolgröße für einen Stream

Sie können die Größe der Streamsymbole auf folgende Weisen ändern:

- Über eine Einstellung für Streameigenschaften
- Über ein Popup-Menü im Stream
- Über die Tastatur

Sie können die gesamte Streamansicht auf eine Anzahl von Größenwerten zwischen 8 und 200 % der Standardsymbolgröße skalieren.

Skalierung des gesamten Streams (Methode mit Streameigenschaften)

1. Wählen Sie im Hauptmenü Folgendes aus:
Tools > Streameigenschaften > Optionen > Layout.
2. Wählen Sie im Menü "Symbolgröße" die gewünschte Größe aus.
3. Klicken Sie auf **Anwenden**, um das Ergebnis anzuzeigen.
4. Klicken Sie auf **OK**, um die Änderung zu speichern.

Skalierung des gesamten Streams (Methode über Menü)

1. Klicken Sie mit der rechten Maustaste auf den Streamhintergrund im Erstellungsbereich.
2. Wählen Sie die Option **Symbolgröße** und anschließend die gewünschte Größe aus.

Skalierung des gesamten Streams (Methode über Tastatur)

1. Drücken Sie auf der Haupttastatur die Tastenkombination "Strg + [-]", um auf die nächstkleinere Größe umzuschalten.
2. Drücken Sie auf der Haupttastatur die Tastenkombination "Strg + Umschalt + [+]", um auf die nächstgrößere Größe umzuschalten.

Diese Funktion ist besonders nützlich, um einen allgemeinen Überblick über einen komplexen Stream zu erhalten. Sie können damit auch die für das Drucken eines Streams erforderliche Anzahl an Seiten minimieren.

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 Streamerstellungsbereich 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 Streamerstellungsbereich miteinander zu verbinden. Doppelklicken Sie auf die mitt-

lere 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 die Taste Alt drücken und gleichzeitig auf die Maus klicken und ziehen.

Verwenden von Tastenkombinationen

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

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

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

Tabelle 1. Unterstützte Tastenkombinationen

Tastenkombination	Funktion
Strg+A	Alles markieren
Strg+X	Ausschneiden
Strg+N	Neuer Stream
Strg+O	Stream öffnen
Strg+P	Drucken
Strg+C	Kopieren
Strg+V	Einfügen
Strg+Z	Rückgängig
Strg+Q	Alle Knoten unterhalb des ausgewählten Knotens auswählen
Strg+W	Auswahl aller nachgeordneten Knoten aufheben (wechselt mit Strg+Q)
Strg+E	Vom ausgewählten Knoten ausführen
Strg+S	Aktuellen Stream speichern
Alt+Pfeiltasten	Ausgewählte Knoten im Streamerstellungsbereich in die Richtung verschieben, in die die verwendete Pfeiltaste zeigt
Umschalt+F10	Popup-Menü für den ausgewählten Knoten öffnen

Tabelle 2. Unterstützte Tastenkombinationen für alte Hot Keys

Tastenkombination	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/inaktivieren
Strg+Alt+F	Cache löschen
Strg+Alt+X	Superknoten erweitern
Strg+Alt+Z	Vergrößern/Verkleinern

Table 2. Unterstützte Tastenkombinationen für alte Hot Keys (Forts.)

Tastenkombination	Funktion
Delete	Knoten oder Verbindung löschen

Drucken

Folgende Objekte können in IBM SPSS Modeler gedruckt werden:

- Streamdiagramme
- Grafiken
- Tabellen
- Berichte (über den Berichtsknoten und die Projektberichte)
- Scripts (über die Dialogfelder "Streameigenschaften", "Standalone-Script" oder "Superknotenscript")
- 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** aus.
- Um vor dem Drucken eine Vorschau anzuzeigen, wählen Sie im Menü "Datei" die Option **Druckvorschau** aus.
- 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** aus.

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 Codierungsmöglichkeiten und Automatisierungsunterstützung.

- **Control Language for Expression Manipulation (CLEM)** ist eine Sprache zur Analyse und Bearbeitung der Daten, die IBM SPSS Modeler-Streams durchlaufen. Data-Mining-Experten verwenden CLEM in hohem Maße für Streamoperationen, um so einfache Aufgaben durchzuführen, wie das Ableiten von Profit 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.
- Die **Scripterstellung** bildet ein leistungsstarkes Tool, mit dem Prozesse in der Benutzerschnittstelle automatisiert werden. Mit Scripts können dieselben Aktionen durchgeführt werden, die die Benutzer mithilfe von Maus und Tastatur durchführen. Sie können auch die Ausgabe angeben und generierte Modelle bearbeiten.

Kapitel 3. Einführung in die Modellierung

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

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

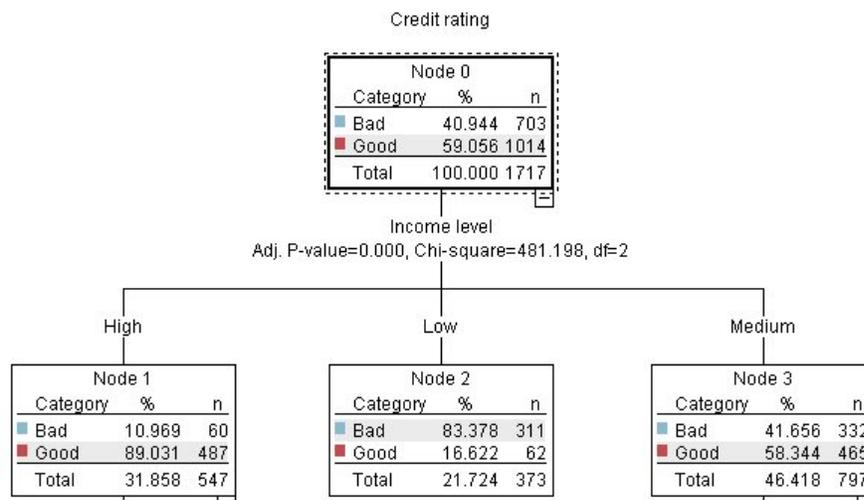


Abbildung 11. Ein einfaches Entscheidungsbaummodell

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

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

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

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

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

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

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

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

Werfen wir nun einen Blick auf den Stream.

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

Erstellen des Streams

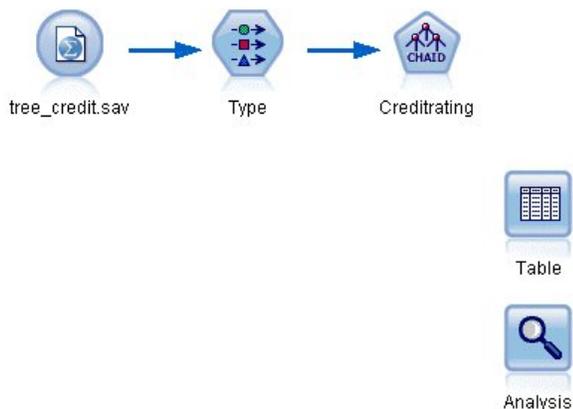


Abbildung 12. Modellierungsstream

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

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

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

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

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

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



Abbildung 13. Einlesen von Daten mit einem Quellenknoten für Statistikdateien

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

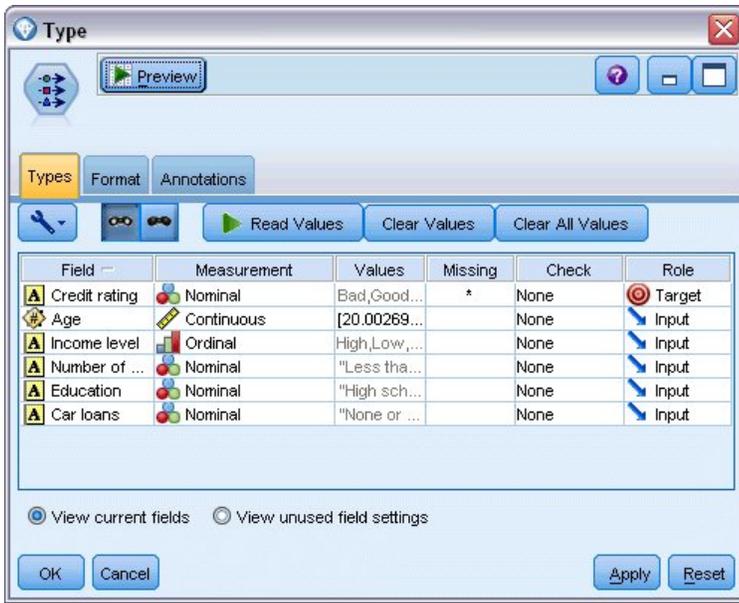


Abbildung 14. Festlegen des Ziels und der Eingabefelder mit dem Typknoten

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 Ziel-felds vorhersagt.

Der CHAID-Modellierungsknoten generiert das Modell.

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

1. Klicken Sie auf die Registerkarte "Erstellungsoptionen".

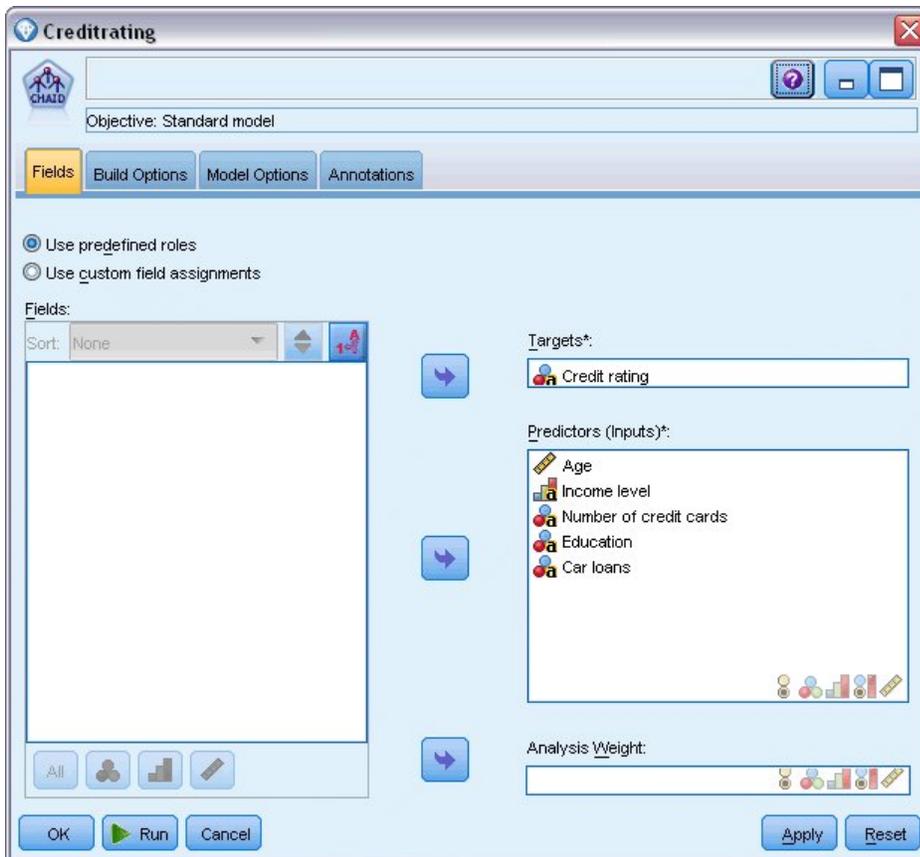


Abbildung 15. CHAID-Modellierungsknoten, Registerkarte "Felder"

Hier finden Sie einige Optionen, über die Sie die Art des aufzubauenden Modells festlegen können. Da wir ein komplett neues Modell möchten, verwenden wir die Standardoption **Neues Modell aufbauen**.

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

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

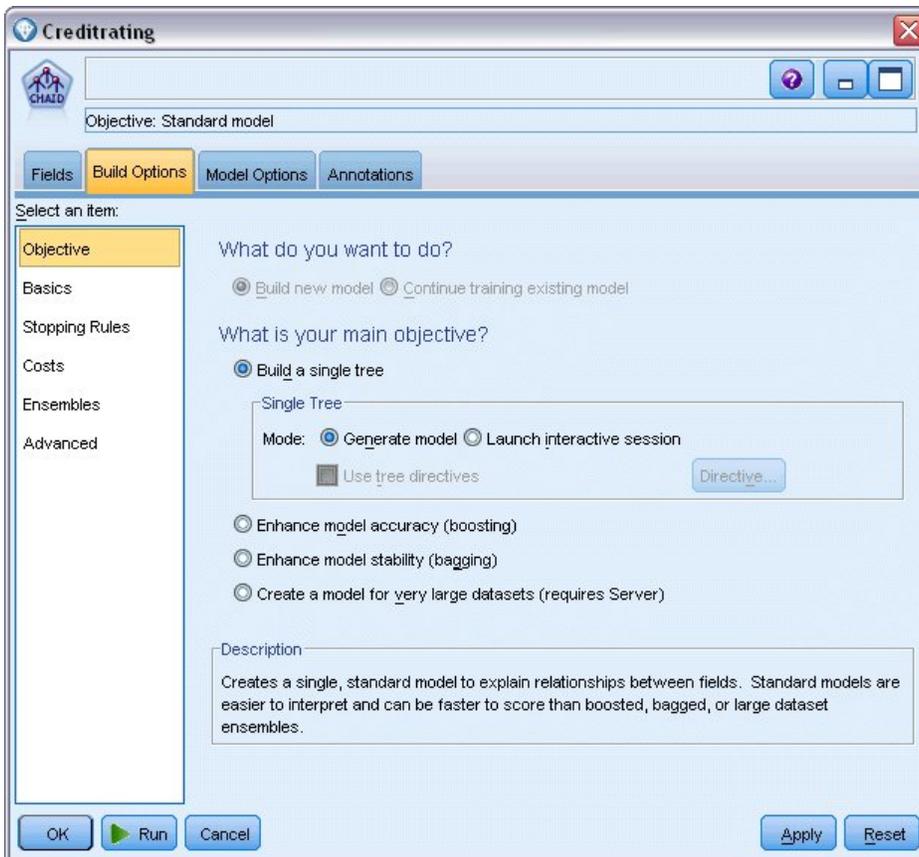


Abbildung 16. CHAID-Modellierungsknoten, Registerkarte "Erstellungsoptionen"

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

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

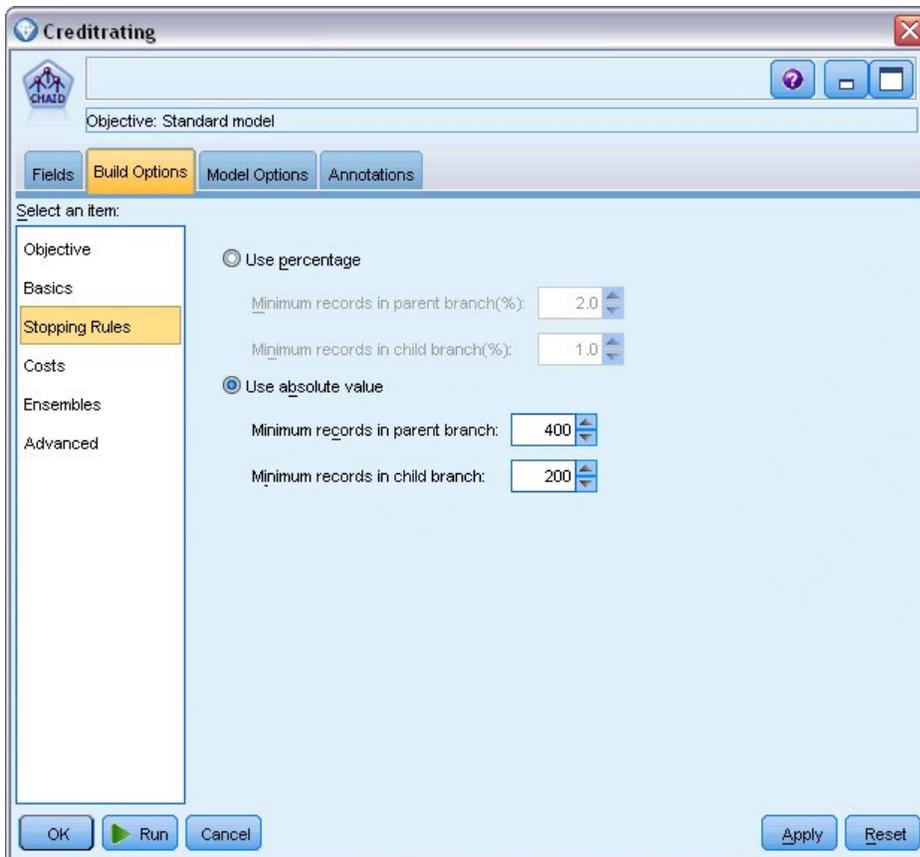


Abbildung 17. Festlegen der Stoppkriterien beim Erstellen von Entscheidungsbäumen

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

Durchsuchen des Modells

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

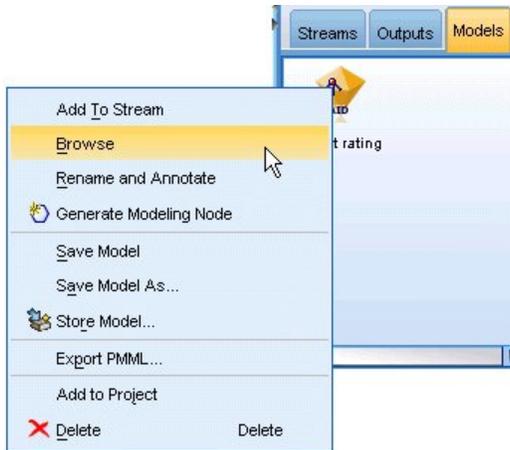


Abbildung 18. Modellpalette

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

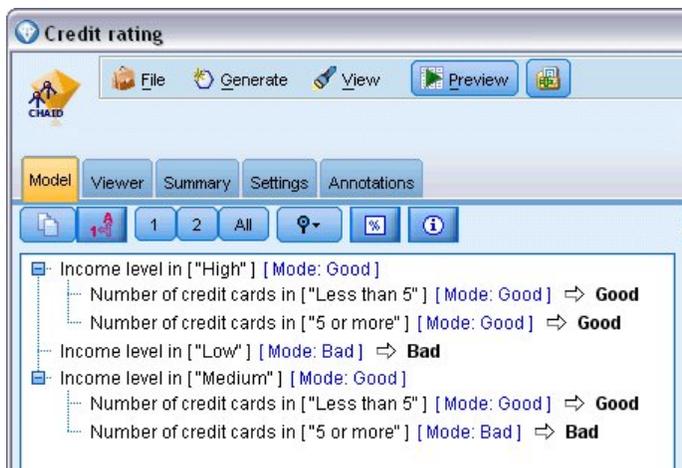


Abbildung 19. CHAID-Modellnugget, Regelset

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

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

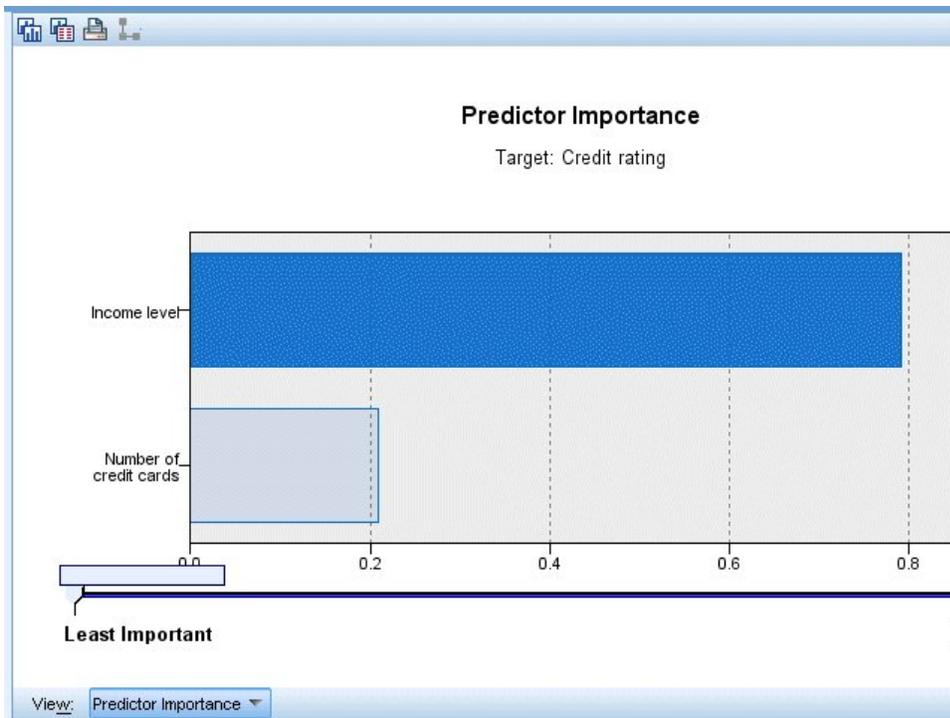


Abbildung 20. Bedeutsamkeit der Prädiktoren - Diagramm

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

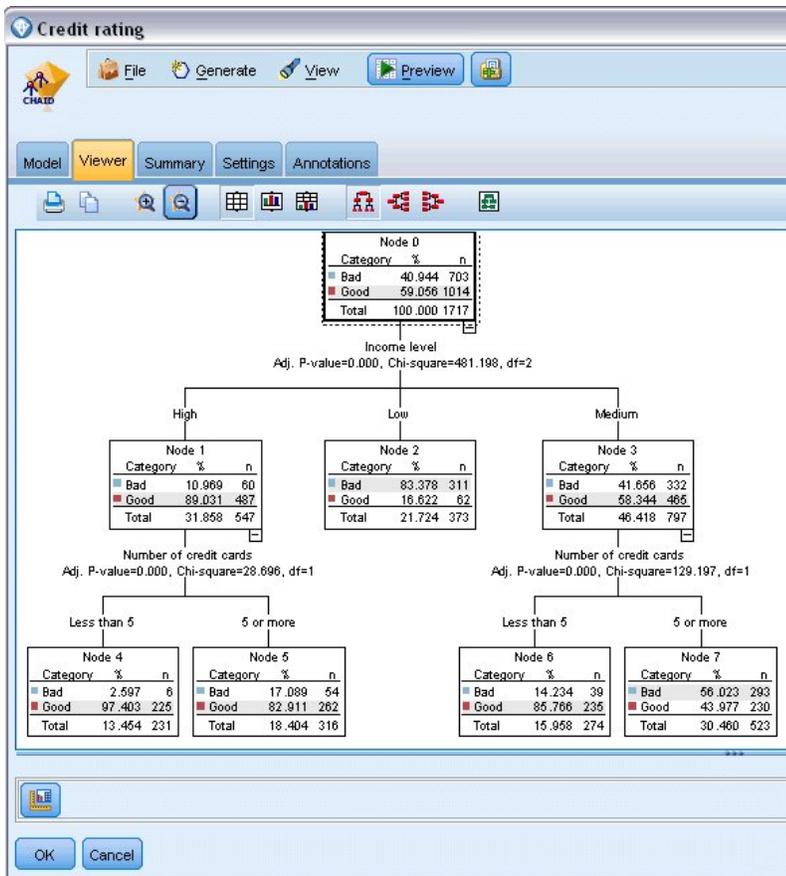


Abbildung 21. Registerkarte "Viewer" im Modellnugget, "Verkleinern" ausgewählt

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

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

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

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

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

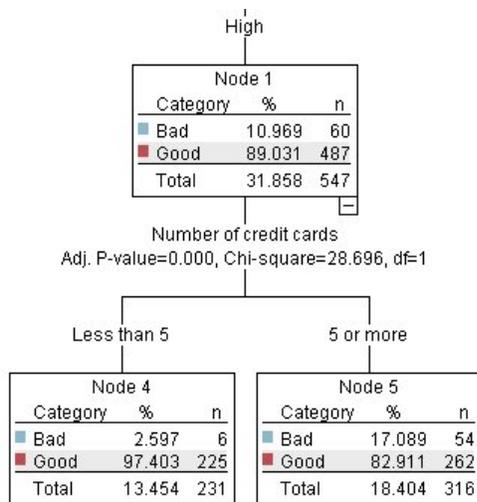


Abbildung 22. Baumansicht der Kunden mit hohem Einkommen

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 zufriedeneres Ergebnis erzielen.

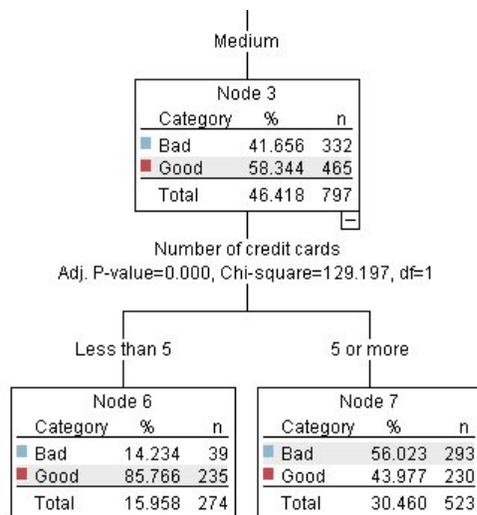


Abbildung 23. Baumansicht der Kunden mit mittlerem Einkommen

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

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

Bewerten des Modells

Wir haben das Modell durchsucht, um zu verstehen, wie das Scoring funktioniert. Aber um zu evaluieren, *mit welcher Genauigkeit* es funktioniert, müssen wir einige Datensätze scoden 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.

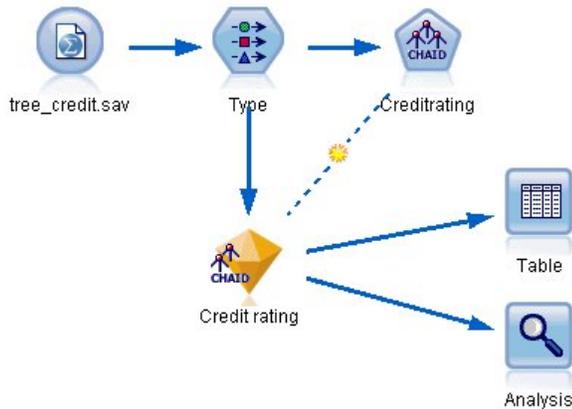


Abbildung 24. Anfügen des Modellnuggets an Ausgabeknoten zur Modellevaluierung

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

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

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

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

Abbildung 25. Tabelle mit generierten Scores und Konfidenzwerten

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

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

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

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

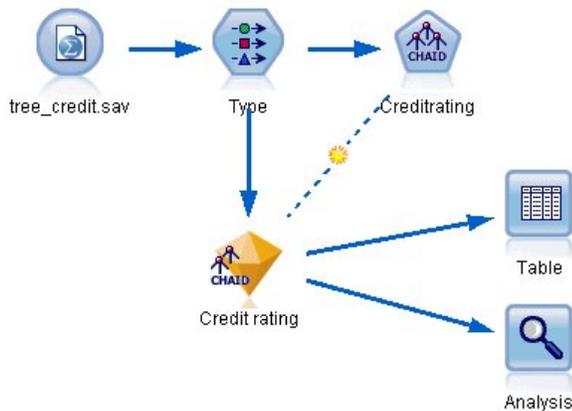


Abbildung 26. Einfügen eines Analyseknotts

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

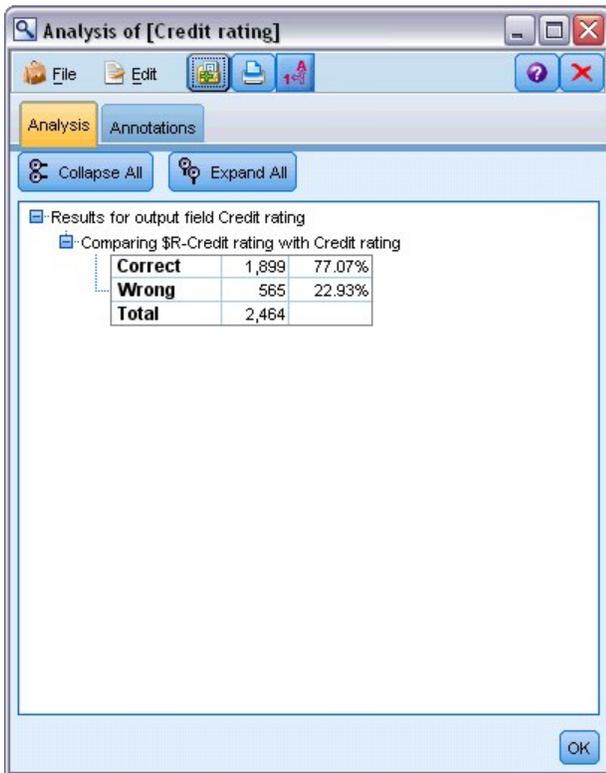


Abbildung 27. Analyseergebnisse für den Vergleich zwischen den beobachteten und vorhergesagten Ergebnissen

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

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

Mit dem Analyseknoden 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.

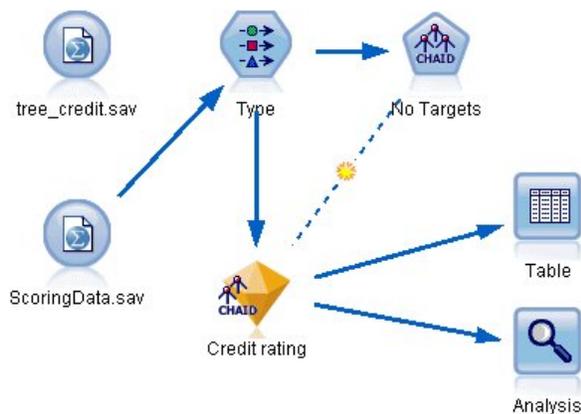


Abbildung 28. Anfügen neuer Daten zum Scoring

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

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

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

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

Zusammenfassung

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

- Der Modellierungsknoten schätzt das Modell durch Untersuchung von Datensätzen, deren Ergebnis bekannt ist, und erstellt ein Modellnugget. Dieser Vorgang wird auch als Trainieren des Modells bezeichnet.

- Das Modellnugget kann jedem Stream mit den erwarteten Feldern hinzugefügt werden, um Datensätze zu scoren. Durch Scoren der Datensätze, deren Ergebnis Sie bereits kennen (z. B. bestehende Kunden), können Sie die Leistung des Modells evaluieren.
- Sobald Sie mit der Leistungsfähigkeit des Modells zufrieden sind, können Sie neue Daten (beispielsweise potenzielle Kunden) scoren, um vorherzusagen, wie diese reagieren.
- Die zum Trainieren bzw. Schätzen des Modells verwendeten Daten können auch als analytische oder historische Daten bezeichnet werden; die Scoring-Daten können auch als operationale Daten bezeichnet werden.

Kapitel 4. Automatische Modellierung für ein Flagziel

Modellieren der Kundenreaktion (Automatisches Klassifikationsmerkmal)

Mit dem Knoten "Automatisches Klassifikationsmerkmal" können Sie automatisch mehrere verschiedene Modelle für Flagziele (beispielsweise ob ein bestimmter Kunde mit hoher Wahrscheinlichkeit einen Kredit nicht zurückzahlt oder auf ein bestimmtes Angebot eingeht) oder nominale Ziele (Setziele) erstellen und vergleichen. In diesem Beispiel wird nach einem Flagergebnis ("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 (Ensemblemodell) zusammen. Dieser Ansatz bietet gleichzeitig den Komfort einer 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 stehen die Vorteile der Automatisierung stärker im Mittelpunkt. Ein ähnliches Beispiel mit einem stetigen Ziel (numerischer Bereich) finden Sie in .

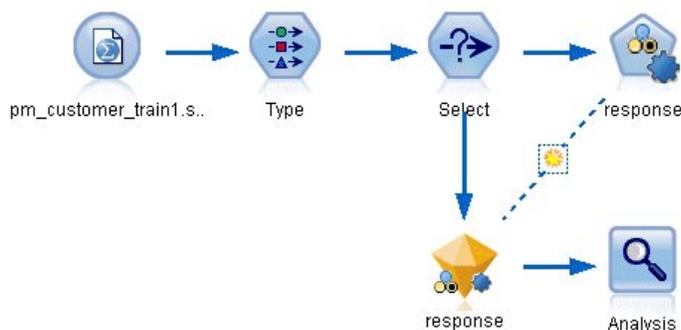


Abbildung 29. Automatisches Klassifikationsmerkmal - Beispielstream

In diesem Beispiel wird der Stream *pm_binaryclassifier.str* verwendet, der im Ordner "Demo" unter *streams* installiert ist. Als Datendatei wird die Datei *pm_customer_train1.sav* verwendet. Weitere Informationen finden Sie unter im Thema „Historische Daten“.

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 entfällt auf die Kampagne *Premium account* (Premium-Account).

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

Table (31 fields, 21,927 records)

	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

Abbildung 30. Daten zu früheren Werbeaktionen

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

1. 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 `$CLE0_DEMOS/` im Dateipfad als Verknüpfung zur Referenzierung dieses Ordners angeben. Beachten Sie, dass im Pfad, wie dargestellt, ein normaler Schrägstrich (und nicht etwa ein umgekehrter Schrägstrich) verwendet werden muss.



Abbildung 31. Einlesen der Daten

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

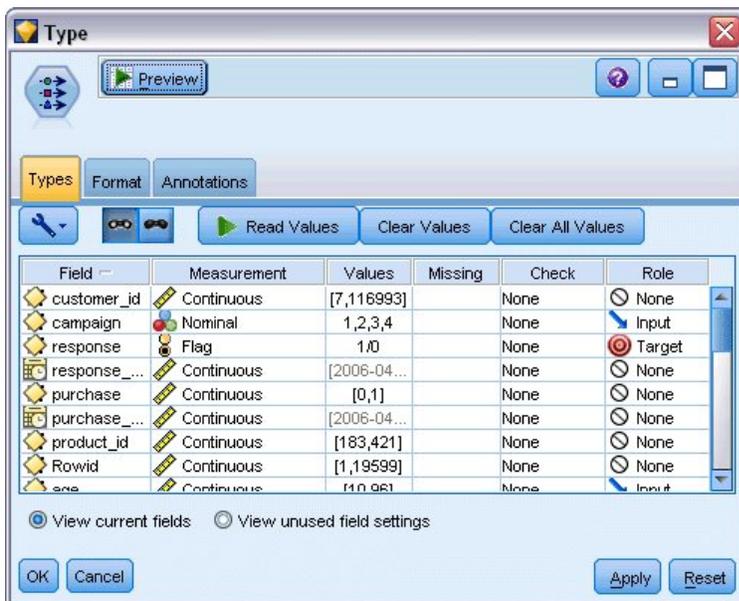


Abbildung 32. Festlegen von Messniveau und Rolle

3. Setzen Sie die Rolle für die folgenden Felder auf **Keine**: *customer_id* (Kunden-ID), *campaign* (Kampagne), *response_date* (Antwortdatum), *purchase* (Einkauf), *purchase_date* (Einkaufsdatum), *product_id* (Produkt-ID), *Rowid* (Zeilen-ID) und *X_random* (X-Zufall). Diese Felder werden beim Erstellen des Modells ignoriert.
4. Klicken Sie auf die Schaltfläche **Werte lesen** im Typknoten, um sicherzustellen, dass die Werte instanziiert werden.

Wie bereits gesehen, umfassen unsere Quelldaten Informationen über vier verschiedene Kampagnen, von denen sich jede an eine andere Art von Kundenkonto richtet. Diese Kampagnen sind in den

Daten als Ganzzahlen codiert. Wir definieren nun Beschriftungen für jede Kampagne, um deutlicher zu sehen, welcher Kontotyp jeder Ganzzahl entspricht.

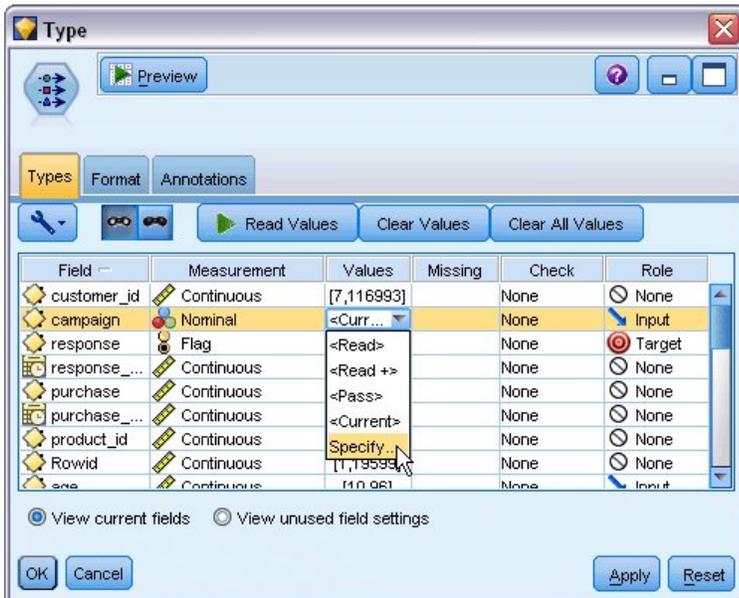


Abbildung 33. Auswahl zum Festlegen von Werten für ein Feld

5. Klicken Sie in der Zeile für das Feld **campaign** (Kampagne) auf den Eintrag in der Spalte **values** (Werte).
6. Wählen Sie **Angeben** aus der Dropdown-Liste.

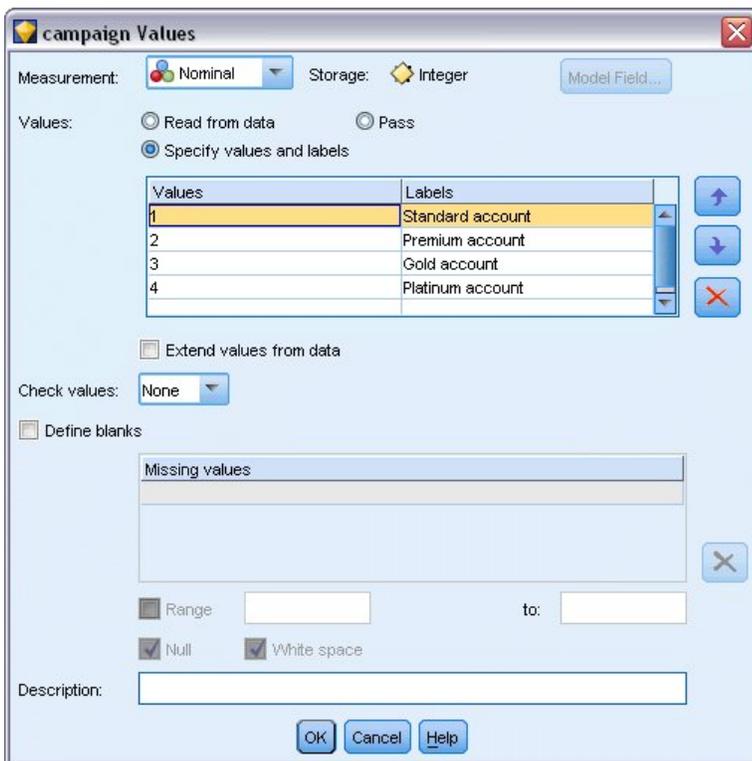


Abbildung 34. Definieren von Beschriftungen für die Feldwerte

7. Geben Sie in der Spalte **Beschriftungen** die Beschriftungen für die vier Werte des Felds **campaign** (Kampagne) wie dargestellt ein.
8. Klicken Sie auf **OK**.
Nun können Sie in Ausgabefenstern die Beschriftungen anstelle der Zahlen anzeigen.

The screenshot shows a window titled "Table (31 fields, 21,927 records) #3". The window has a menu bar with "File", "Edit", and "Generate". Below the menu bar are two tabs: "Table" and "Annotations". The "Table" tab is active, displaying a table with the following columns: customer_id, campaign, response, response_date, purchase, purchase_date, and product_id. The table contains 20 rows of data. The 'campaign' column has values such as "Premium account" and "Gold account". The 'response_date' column has some entries with timestamps like "2006-07-05 00:00:00" and "2006-07-14 00:00:00". The 'product_id' column has values like "183" and "\$null\$". An "OK" button is visible at the bottom right of the window.

	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\$

Abbildung 35. Anzeigen der Feldwertbeschriftungen

9. Verbinden Sie einen Tabellenknoten mit dem Typknoten.
10. Öffnen Sie den Tabellenknoten und klicken Sie auf **Ausführen**.
11. Klicken Sie im Ausgabefenster auf die Symbolleistenschaltfläche **Feld- und Wertbeschriftungen anzeigen**, um die Beschriftungen anzuzeigen.
12. 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 codiert als *campaign=2*), können Sie einen Auswahlknoten verwenden, um nur die betreffenden Datensätze in den Stream aufzunehmen.

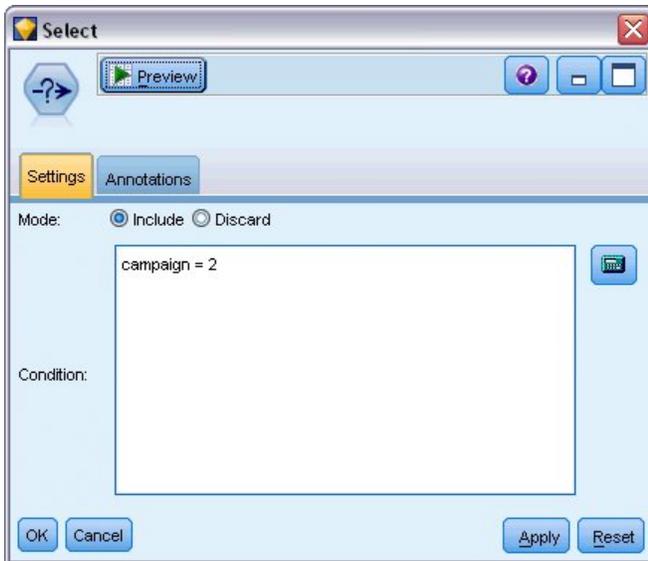


Abbildung 36. Auswählen von Datensätzen für eine einzelne Kampagne

Generieren und Vergleichen von Modellen

1. Fügen Sie einen Knoten vom Typ "Automatisches Klassifikationsmerkmal" an und wählen Sie **Gesamtgenauigkeit** als Metrik für die Rangordnung der Modelle aus.
2. Legen Sie für **Anzahl der zu verwendenden Modelle** den Wert 3 fest. Das bedeutet, dass bei der Ausführung des Knotens die drei besten Modelle erstellt werden.

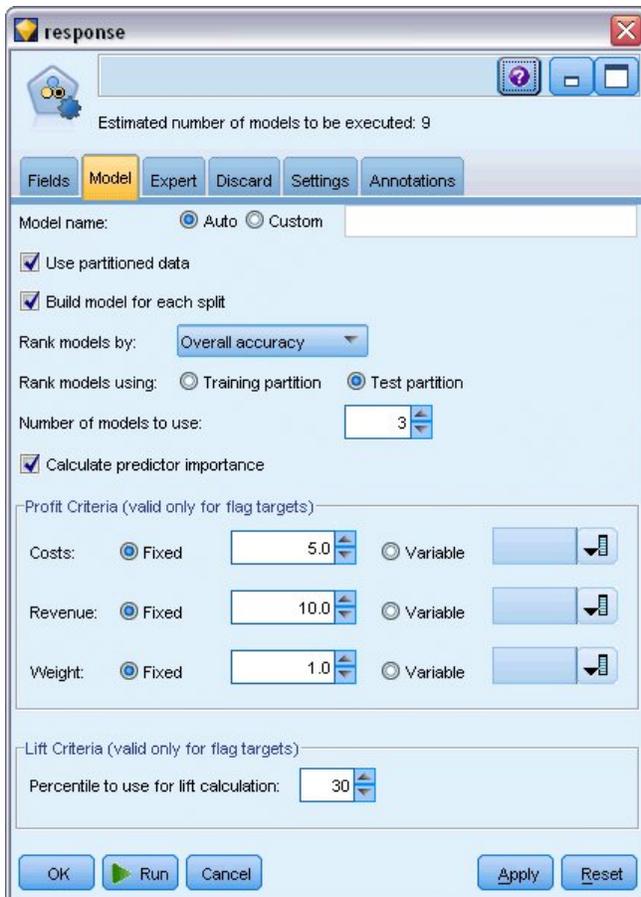


Abbildung 37. Knoten "Automatisches Klassifikationsmerkmal" - Registerkarte "Modell"

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

3. Inaktivieren Sie die Modelltypen **Diskriminanz** und **SVM**. (Bei diesen Modellen dauert das Training für die vorliegenden Daten länger. Durch den Verzicht darauf wird 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** den Wert 3 festgelegt haben, berechnet der Knoten die Genauigkeit der restlichen neun Algorithmen und erstellt ein einzelnes Modellnugget, in dem die drei genauesten enthalten sind.

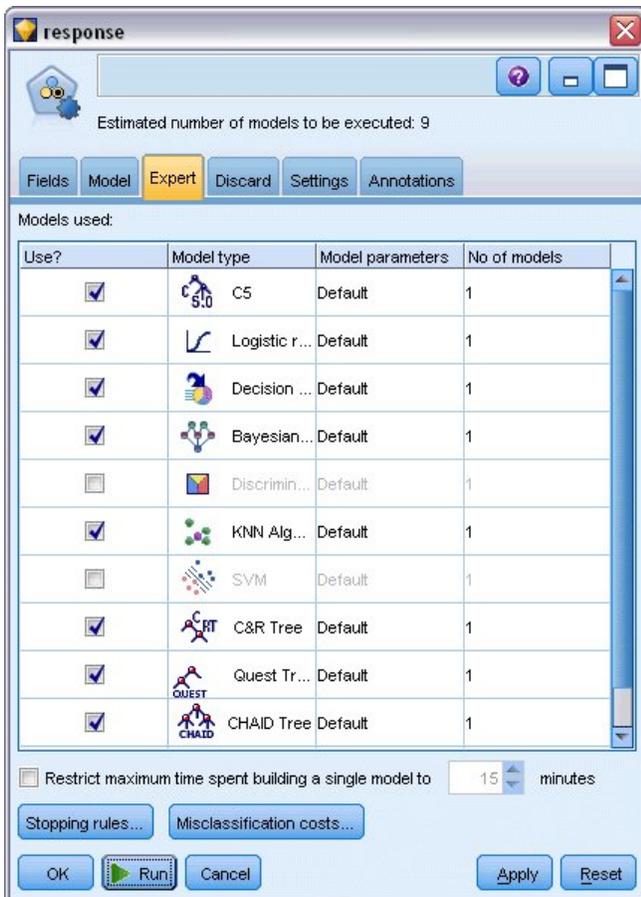


Abbildung 38. Knoten "Automatisches Klassifikationsmerkmal" - Registerkarte "Experten"

- 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*.

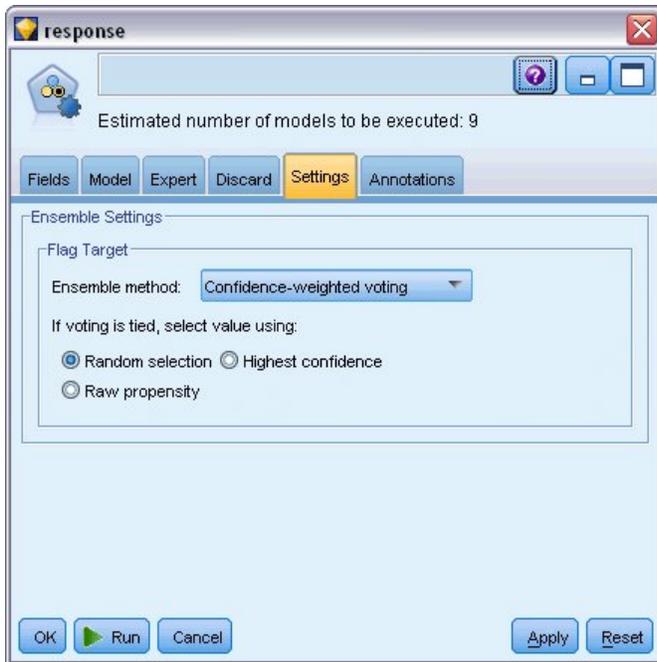


Abbildung 39. Knoten "Automatisches Klassifikationsmerkmal": Registerkarte "Einstellungen"

5. Klicken Sie auf **Ausführen**.

Nach einigen Minuten wird das generierte Modellnugget erstellt und im Erstellungsbereich sowie in der Modellpalette in der rechten oberen Fensterecke platziert. Sie können das Modellnugget durchsuchen oder auf verschiedene Weise speichern bzw. bereitstellen.

Öffnen Sie das Modellnugget. 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 Dataset erstellt werden, kann dieser Vorgang etliche Stunden in Anspruch nehmen.) Siehe Abb. 29 auf Seite 37.

Wenn Sie eines der Modelle eingehender untersuchen möchten, können Sie in der Spalte **Modell** auf ein Modellnuggetsymbol doppelklicken, um einen Drill-Down durchzuführen und die einzelnen Modellergebnisse zu durchsuchen. Anschließend können Sie Modellierungsknoten, Modellnuggets oder Evaluierungsdiagramme generieren. In der Spalte **Diagramm** können Sie ein Diagramm in voller Größe generieren, indem Sie auf einem Piktogramm doppelklicken.

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy	No. Fields Used	Area Under Curve
<input checked="" type="checkbox"/>		C5.1	< 1	4,906.667	8	2.203	92.861	10	0.777
<input checked="" type="checkbox"/>		C&RT Tree 1	3	4,602.692	9	2.778	92.365	8	0.924
<input checked="" type="checkbox"/>		CHAID Tree 1	3	4,145.668	8	2.851	91.706	4	0.927

Abbildung 40. Automatisches Klassifikationsmerkmal - Ergebnisse

Standardmäßig werden die Modelle auf der Grundlage der Gesamtgenauigkeit sortiert, da dies das Maß ist, das Sie auf der Registerkarte "Modell" des Knotens "Automatisches Klassifikationsmerkmal" 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 genauesten 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.

Fügen Sie einen Analyseknott (Ausgabepalette) nach dem Modellnugget an. Klicken Sie mit der rechten Maustaste auf den Analyseknott und wählen Sie **Ausführen** aus, um den Stream auszuführen.

Die vom Ensemblemodell 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* aufgezichnet ist) überein.

Das Ensemblemodell 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 Ensemblemodell mit höherer Wahrscheinlichkeit gute Leistungen, wenn es auf andere Datensätze als die Trainingsdaten angewendet wird.

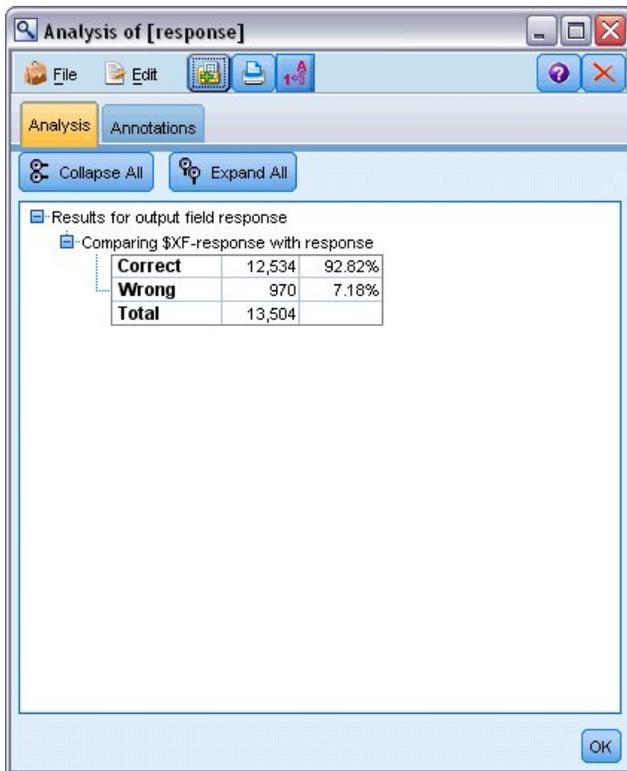


Abbildung 41. Analyse der drei Modelle

Zusammenfassung

Sie haben mithilfe des Knotens "Automatisches Klassifikationsmerkmal" eine Reihe verschiedener Modelle verglichen, die drei genauesten Modelle verwendet und dem Stream hinzugefügt und diese Modelle schließlich in einem Modellnugget "Automatischer Klassifikationsmerkmal" zusammengefasst.

- Hinsichtlich der Gesamtgenauigkeit erbrachten die Modelle "C51", "C&R-Baum" und "CHAID" die besten Leistungen bei den Trainingsdaten.
- Das Ensemblemodell 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 5. 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 ein Subset der Modelle für die weitere Analyse erstellen. Der Knoten funktioniert ebenso wie der Knoten "Automatisches Klassifikationsmerkmal", ist jedoch für stetige und nicht für Flagziele oder nominale Ziele gedacht.

Der Knoten kombiniert die besten der infrage kommenden Modelle in einem einzigen aggregierten Modellnugget. Dieser Ansatz bietet gleichzeitig den Komfort der 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.

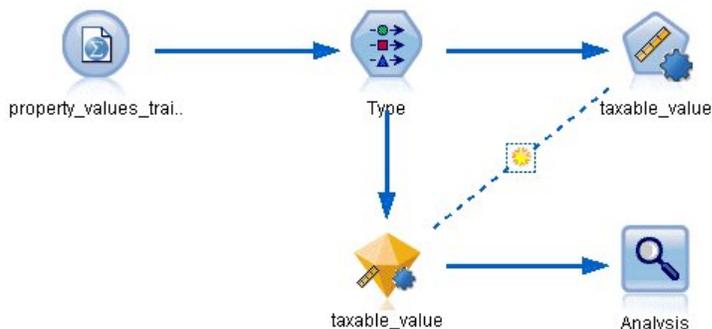


Abbildung 42. Auto-Numerisch - Beispielstream

In diesem Beispiel wird der Stream *property_values_numericpredictor.str* verwendet, der im Ordner "Demos" unter *streams* installiert ist. Als Datendatei wird die Datei *property_values_train.sav* verwendet. Weitere Informationen finden Sie unter im Thema „Ordner "Demos"“ auf Seite 5.

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	Beschriftung
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)

Feldname	Beschriftung
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 Datasets bei dem der steuerliche Wert bekannt ist, können Sie Datensätze scoren, bei denen dieser Wert noch nicht bekannt ist.

Erstellen des Streams

1. 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 dargestellt, ein normaler Schrägstrich (und nicht etwa ein umgekehrter Schrägstrich) verwendet werden muss.

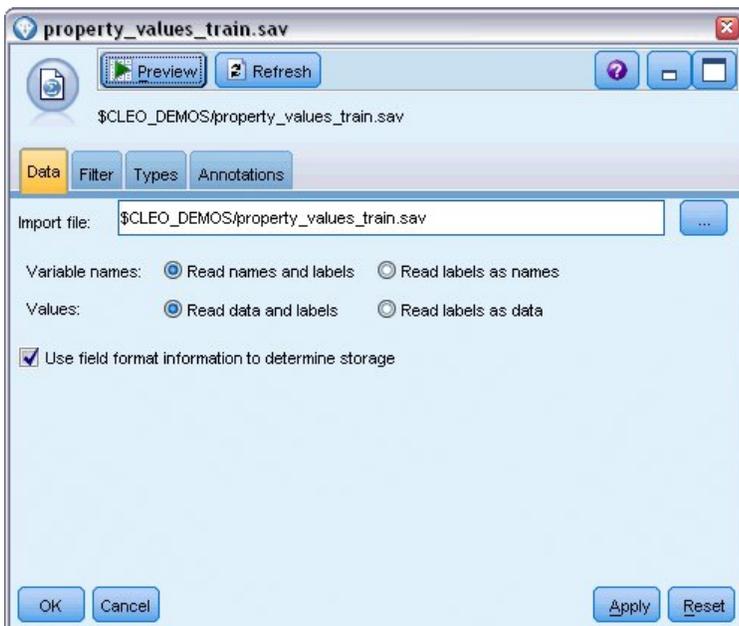


Abbildung 43. Einlesen der Daten

2. 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.

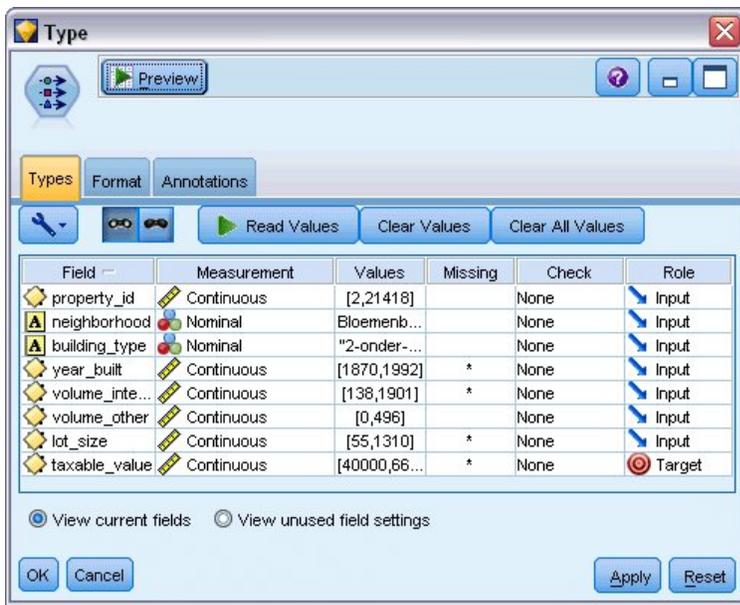


Abbildung 44. Festlegen des Zielfelds

3. Fügen Sie einen Knoten vom Typ "Auto-Numerisch" an und wählen Sie **Korrelation** als Metrik für die Rangordnung der Modelle aus.
4. Legen Sie für **Anzahl der zu verwendenden Modelle** den Wert 3 fest. Das bedeutet, dass bei der Ausführung des Knotens die drei besten Modelle erstellt werden.

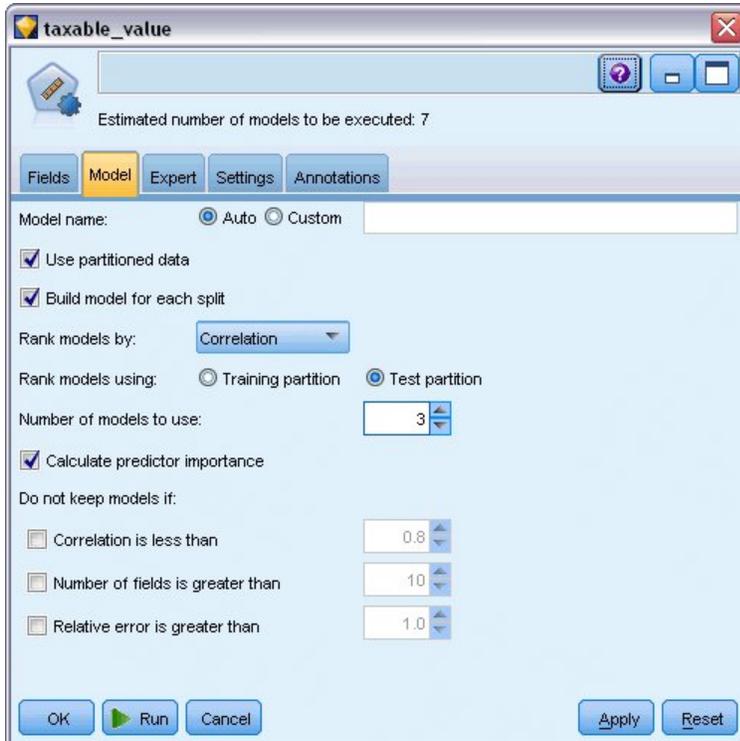


Abbildung 45. Knoten "Auto-Numerisch" - Registerkarte "Modell"

- 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** den Wert 3 festgelegt haben, berechnet der Knoten die Genauigkeit der sieben Algorithmen und erstellt ein einzelnes Modellnugget, in dem die drei genauesten enthalten sind.

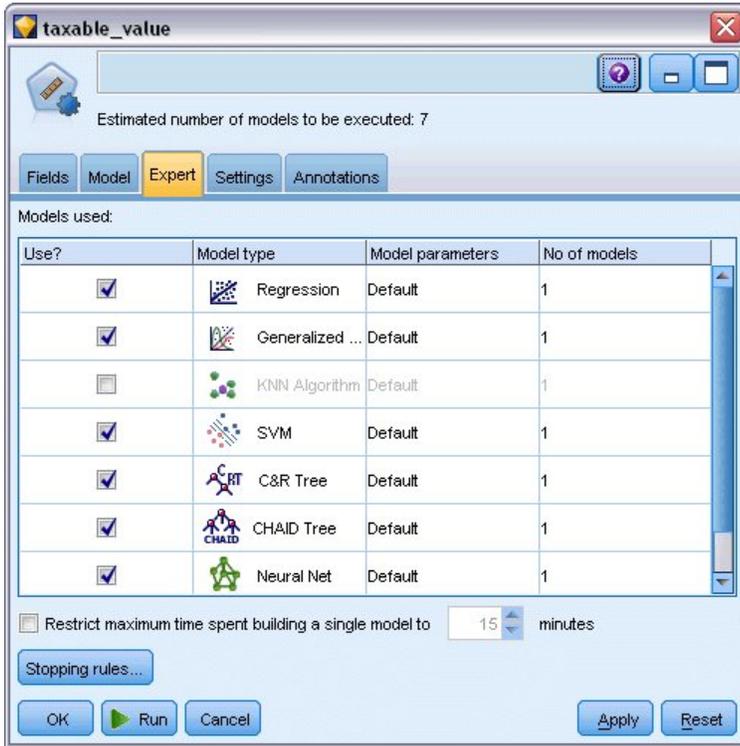


Abbildung 46. Knoten "Auto-Numerisch" - Registerkarte "Experten"

- 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.

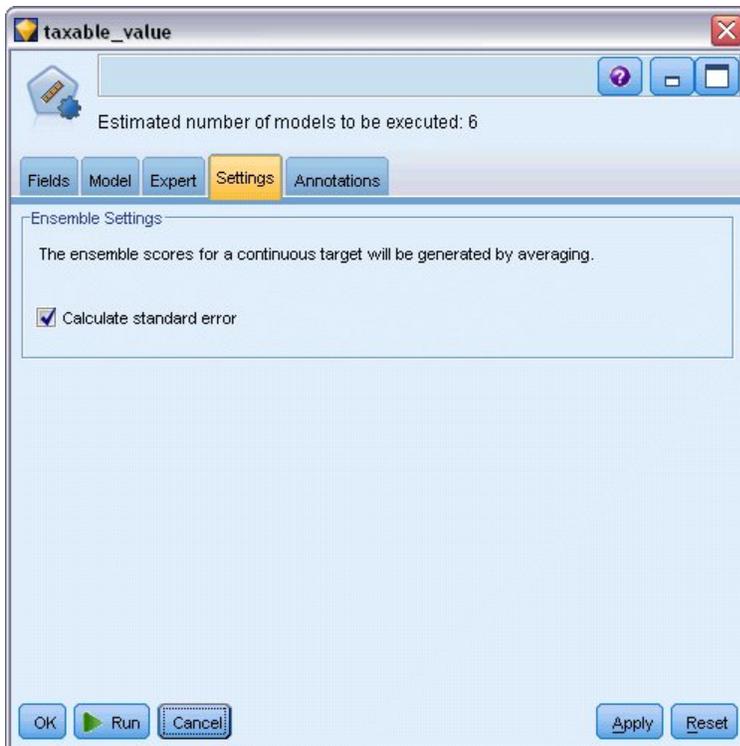


Abbildung 47. Knoten "Auto-Numerisch" - Registerkarte "Einstellungen"

Vergleichen der Modelle

1. Klicken Sie auf die Schaltfläche "Ausführen".

Das Modellnugget wird erstellt und in den Erstellungsbereich sowie in der Modellpalette in der rechten oberen Fensterecke platziert. Sie können das Nugget durchsuchen oder auf verschiedene Weise speichern bzw. bereitstellen.

Öffnen Sie das Modellnugget. 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 Dataset geschätzt werden, kann dieser Vorgang etliche Stunden in Anspruch nehmen.) Siehe Abb. 42 auf Seite 49.

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

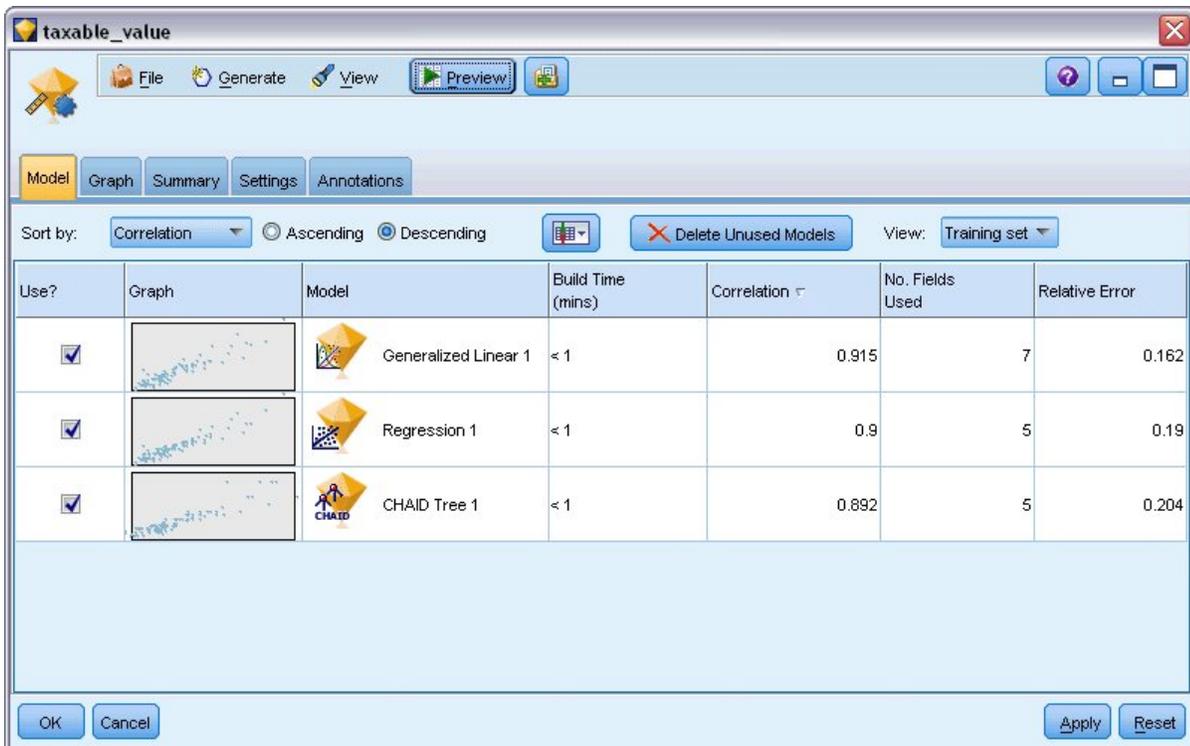


Abbildung 48. Auto-Numerisch - Ergebnisse

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.

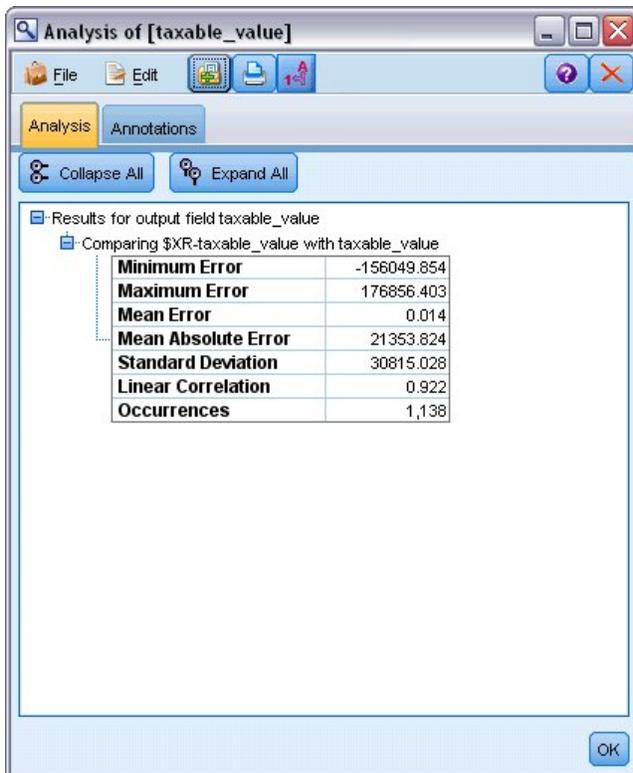
In der Spalte **Diagramm** können Sie ein Diagramm in voller Größe generieren, indem Sie auf einem Piktogramm doppelklicken.

Basierend auf diesen Ergebnissen entscheiden Sie sich, jedes der drei genauesten 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.

Fügen Sie einen Analyseknoden (Ausgabepalette) nach dem Modellnugget an. Klicken Sie mit der rechten Maustaste auf den Analyseknoden und wählen Sie **Ausführen** aus, um den Stream auszuführen.

Der vom Ensemblemodell generierte Durchschnittsscore 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 Datensätze möglicherweise eine bessere Leistung als jedes der Einzelmodelle.



Results for output field taxable_value	
Comparing \$XR-taxable_value with taxable_value	
Minimum Error	-156049.854
Maximum Error	176856.403
Mean Error	0.014
Mean Absolute Error	21353.824
Standard Deviation	30815.028
Linear Correlation	0.922
Occurrences	1,138

Abbildung 49. Auto-Numerisch - Beispielstream

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 Modellnugget "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.

Kapitel 6. 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 (ADP - Automated Data Preparation, 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 Vorkenntnisse zu den verwendeten Statistikkonzepten erforderlich sind. Wenn Sie den Knoten mit den Standardeinstellungen ausführen, werden Modelle gewöhnlich schneller erstellt und bewertet.

Bei diesem Beispiel wird der Stream *ADP_basic_demo.str* verwendet, 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. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. Die Datei *ADP_basic_demo.str* befindet sich im Verzeichnis *streams*.

Erstellen des Streams

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

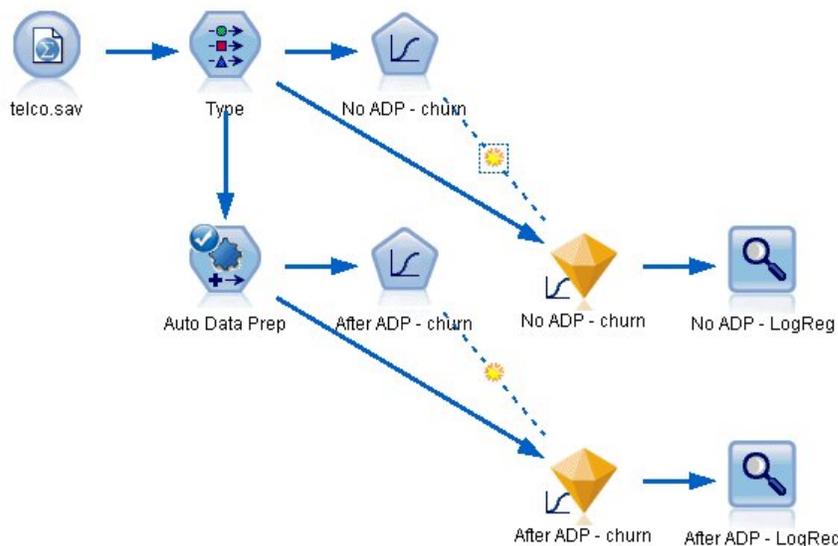


Abbildung 50. Erstellen des Streams

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



Abbildung 51. Auswahl des Ziels

3. Verbinden Sie einen logistischen Knoten mit dem Typknoten.
4. 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.

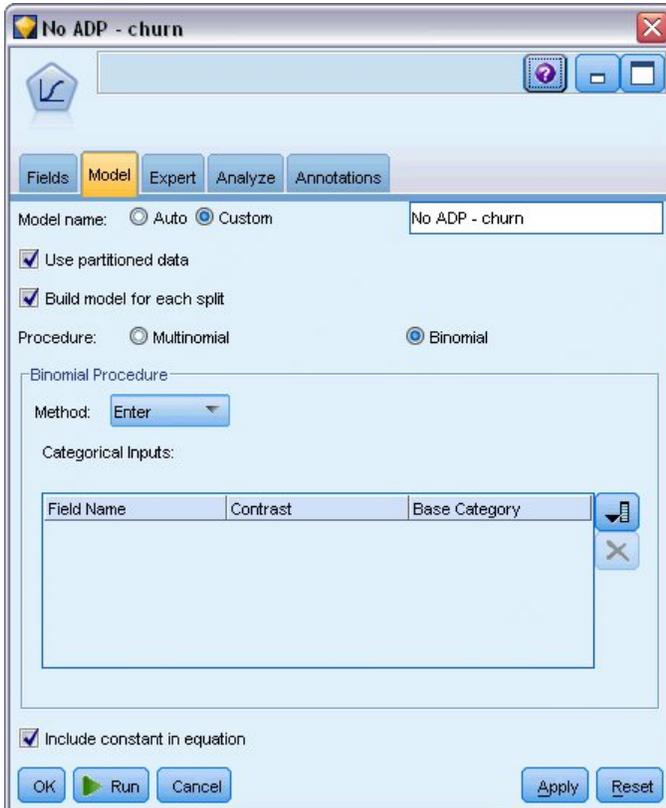


Abbildung 52. Auswählen der Modelloptionen

5. Verbinden Sie einen ADP-Knoten mit dem Typknoten. Übernehmen Sie auf der Registerkarte "Ziele" die Standardeinstellungen, um Ihre Daten durch Ausgleich von Geschwindigkeit und Genauigkeit vorzubereiten.
6. 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.

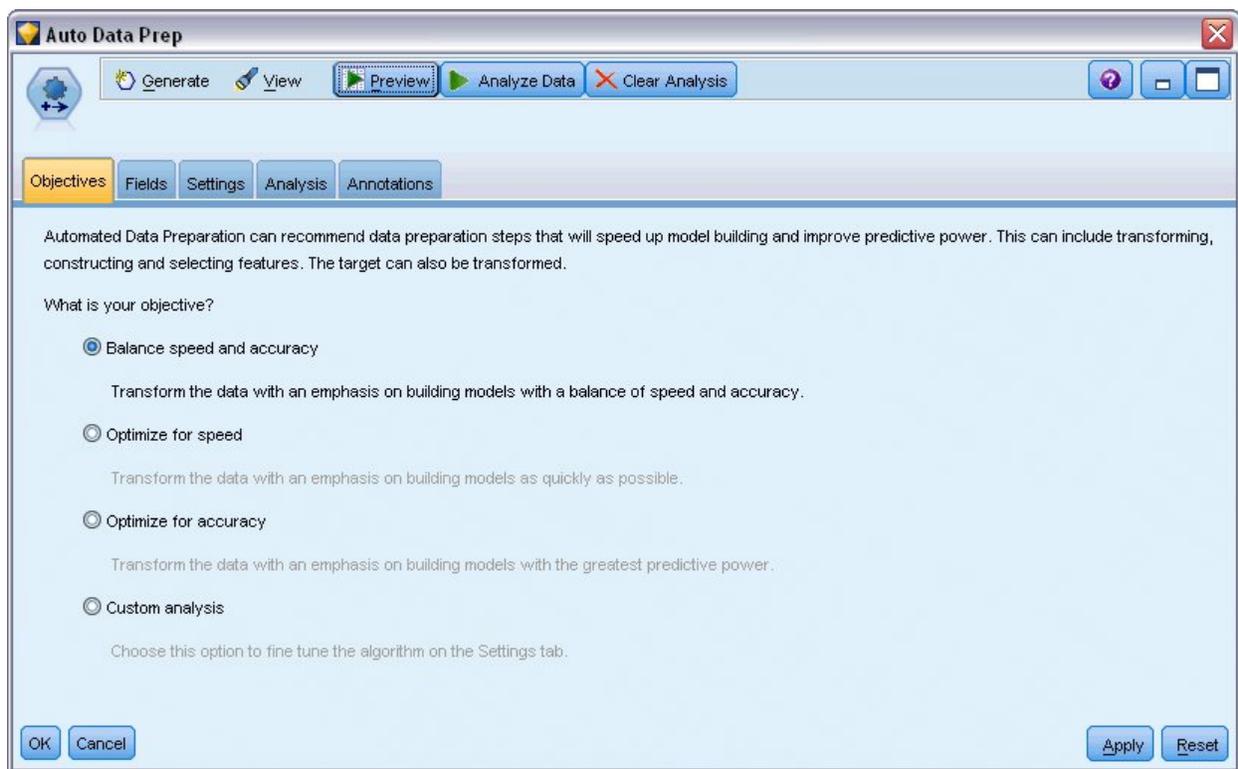


Abbildung 53. ADP-Standardziele

Die Ergebnisse der Datenverarbeitung werden auf 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 verwendet verworfen wurden.

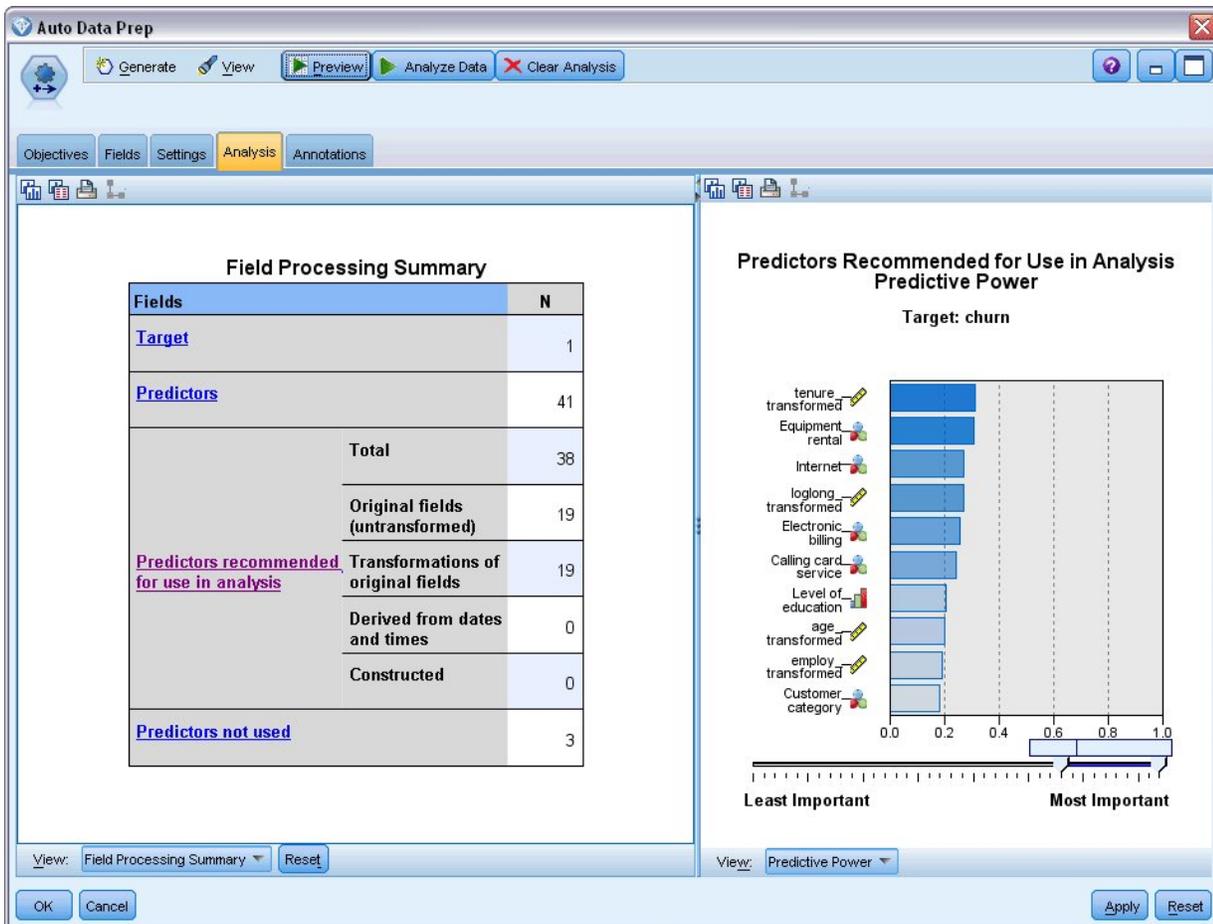


Abbildung 54. Übersicht der Datenverarbeitung

7. Verbinden Sie einen logistischen Knoten mit dem ADP-Knoten.
8. 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.

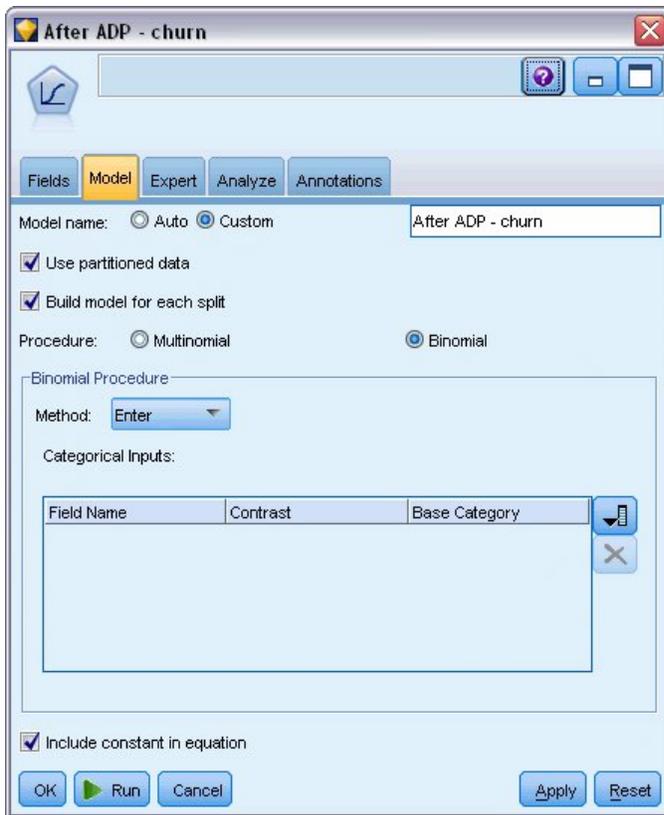


Abbildung 55. Auswählen der Modelloptionen

Vergleichen der Modellgenauigkeit

1. Führen Sie beide Logistikknoten aus, um die Modellnuggets zu erstellen; diese werden dem Stream und der Modellalette in der rechten oberen Ecke hinzugefügt.

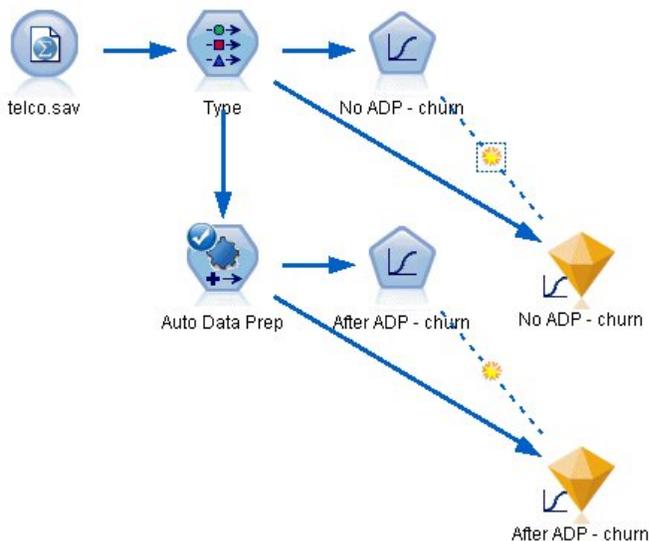


Abbildung 56. Anfügen der Modellnuggets

2. Fügen Sie Analyseknotten an die Modellnuggets an und führen Sie die Analyseknotten mit den zugehörigen Standardeinstellungen aus.

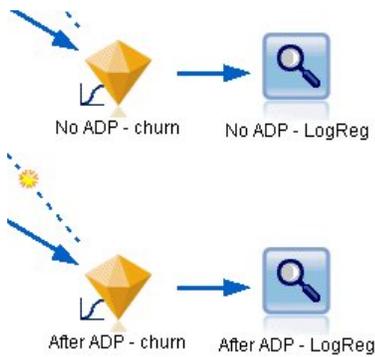


Abbildung 57. Anfügen der Analyseknotten

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

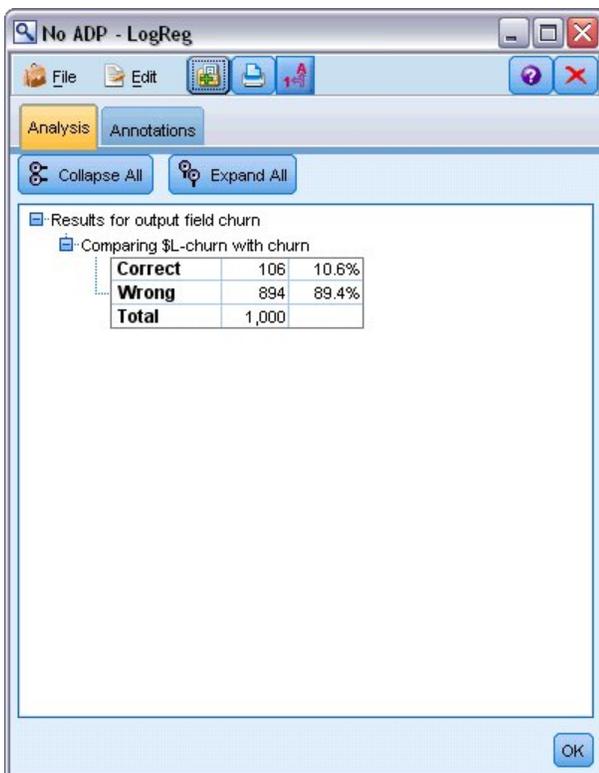


Abbildung 58. Ergebnisse des Modells ohne ADP

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.

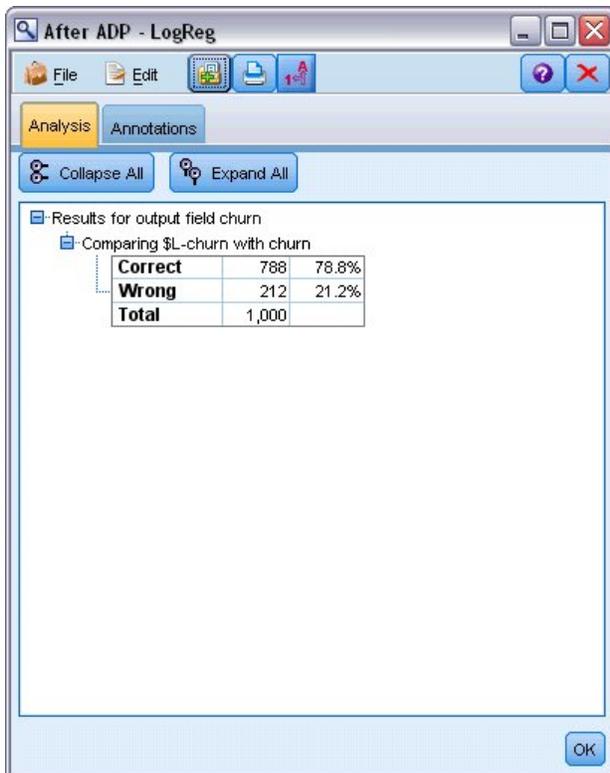


Abbildung 59. Ergebnisse des Modells mit ADP

Zusammenfassend lässt sich sagen, dass Sie durch einfaches Ausführen des ADP-Knotens zur Feineinstellung der Verarbeitung Ihrer Daten ein genaueres Modell mit wenig direkter Datenbearbeitung 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 sich der ADP-Knoten als vorteilhaft erweisen.

Erläuterungen der mathematischen Grundlagen für die in IBM SPSS Modeler verwendeten Modellierungsmethoden finden Sie im *IBM 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 ein Subset der Datensätze für Test- und Validierungszwecke zurückhalten.

Kapitel 7. 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 Datenexploration 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 auf die Datendatei *telco.sav* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. Die Datei *telco_dataaudit.str* befindet sich im Verzeichnis *streams*.

Erstellen des Streams

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

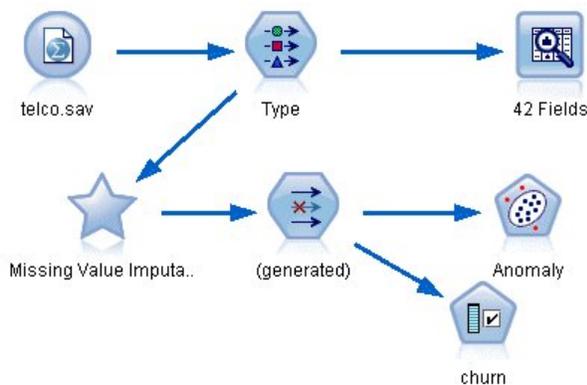


Abbildung 60. Erstellen des Streams

2. 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.

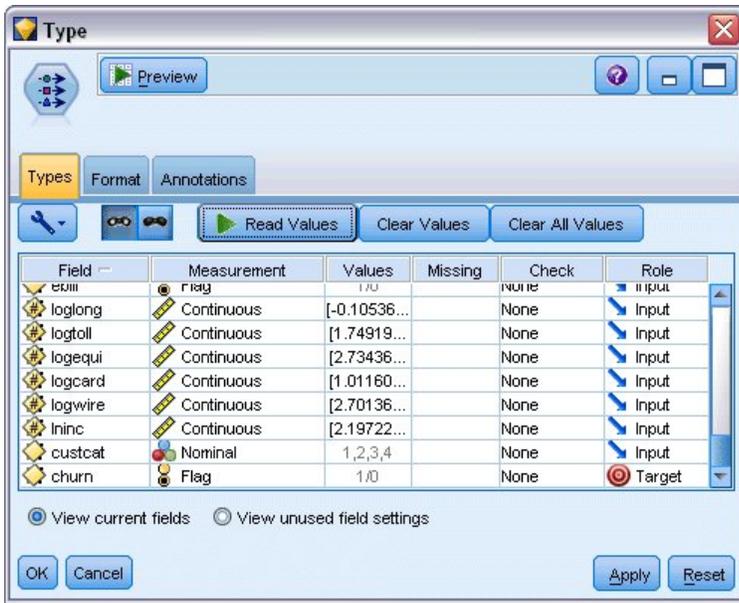


Abbildung 61. Festlegen des Ziels

3. 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 62. 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.

4. Fügen Sie dem Stream einen Data Audit-Knoten 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.

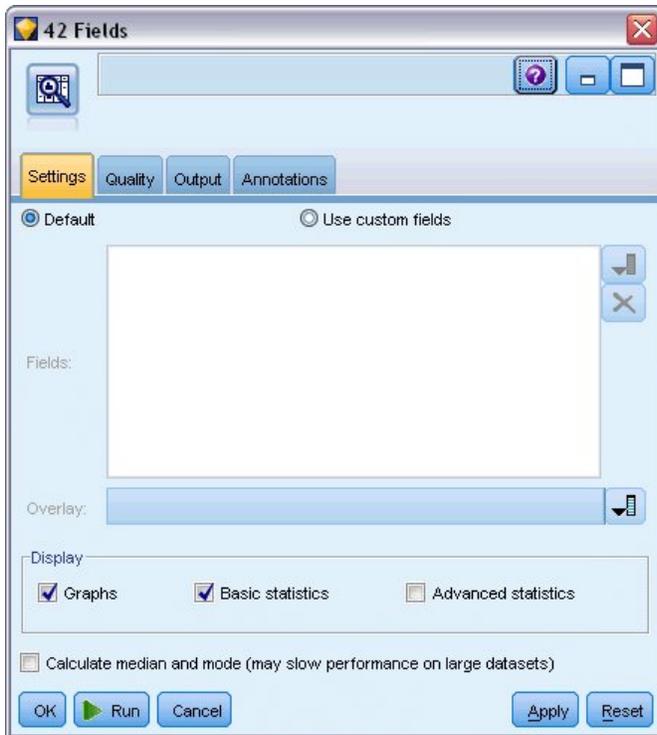


Abbildung 63. Data Audit-Knoten - Registerkarte "Einstellungen"

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

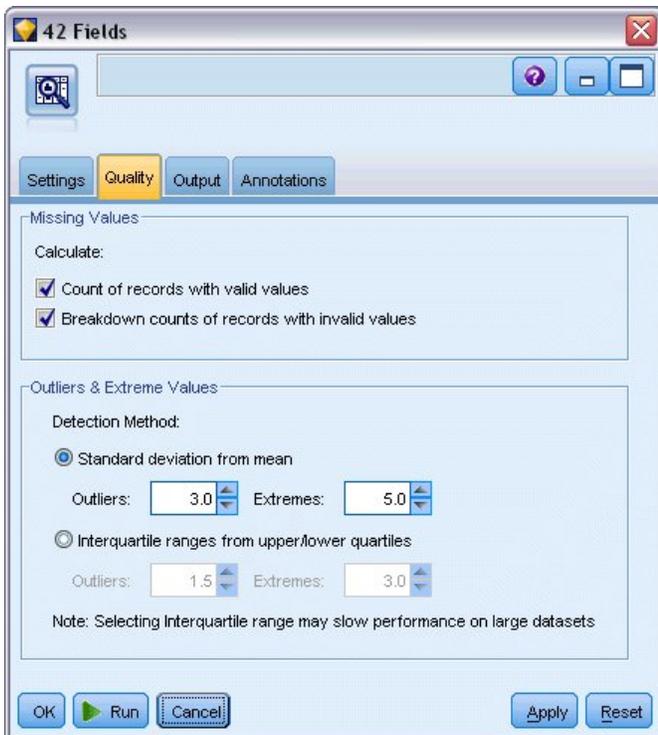


Abbildung 64. Data Audit-Knoten - Registerkarte "Qualität"

Durchsuchen von Statistiken und Diagrammen

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

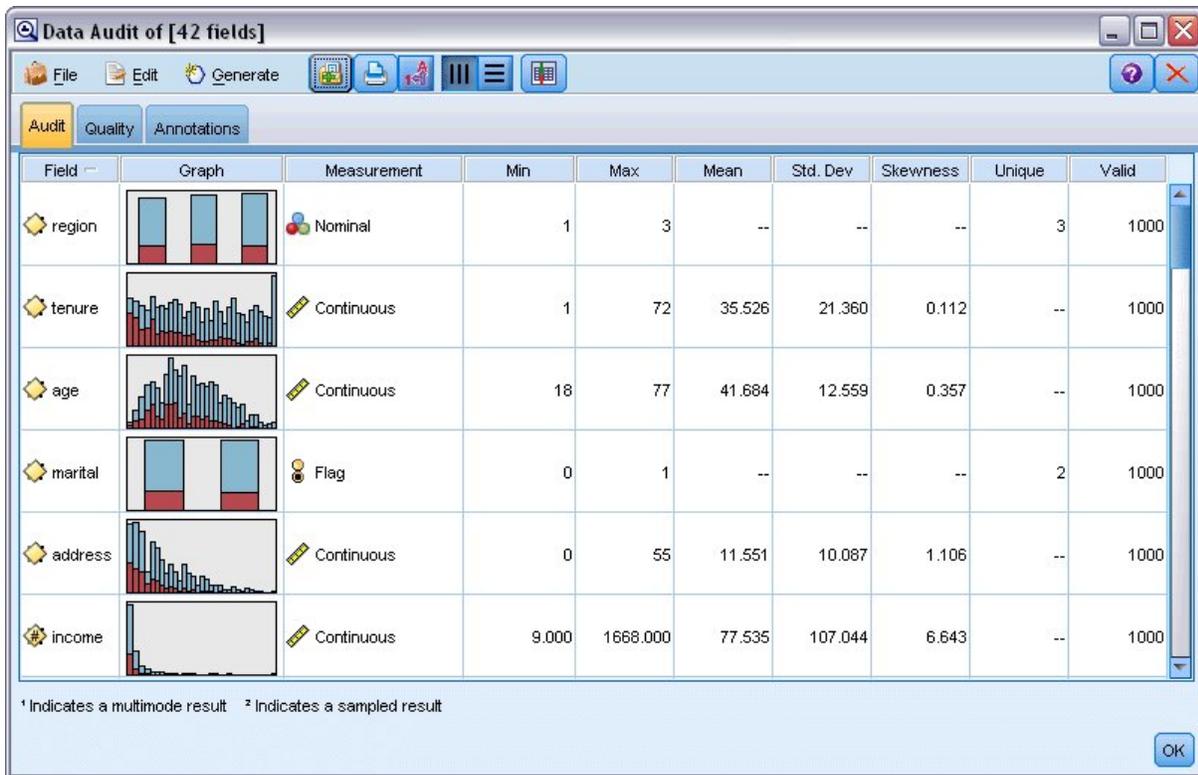


Abbildung 65. Data Audit-Browser

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

1. Außerdem können Sie über die Symbolleiste oder das Menü "Bearbeiten" die anzuzeigenden Statistiken auswählen.

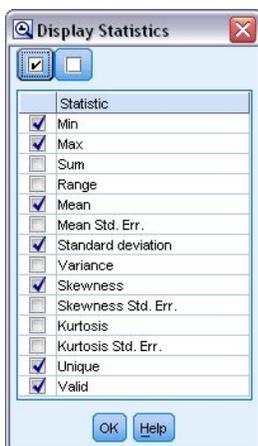


Abbildung 66. Statistik anzeigen

Doppelklicken Sie auf ein Piktogramm 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 Wertbeschriftungen umschalten. Alternativ können Sie auf die Schaltfläche "Bearbeitungsmodus" klicken, um das Diagramm weiter anzupassen.

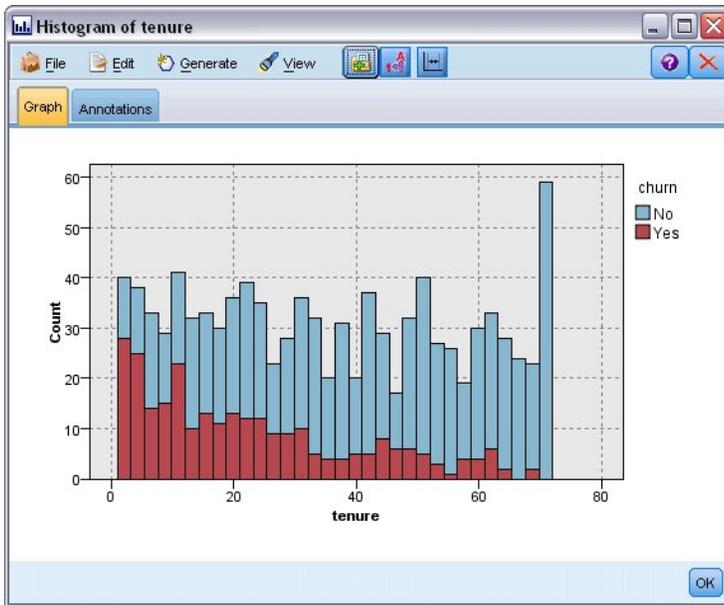


Abbildung 67. Histogramm der Beschäftigungsdauer

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

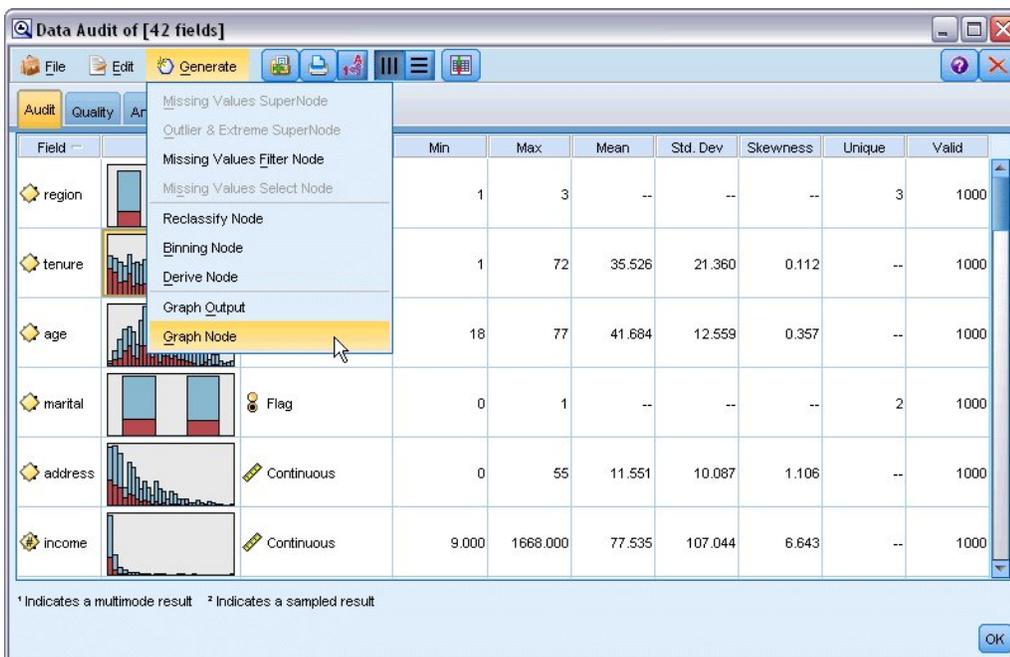


Abbildung 68. Generieren eines Diagrammknotens

Umgang mit Ausreißern und fehlenden Werten

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

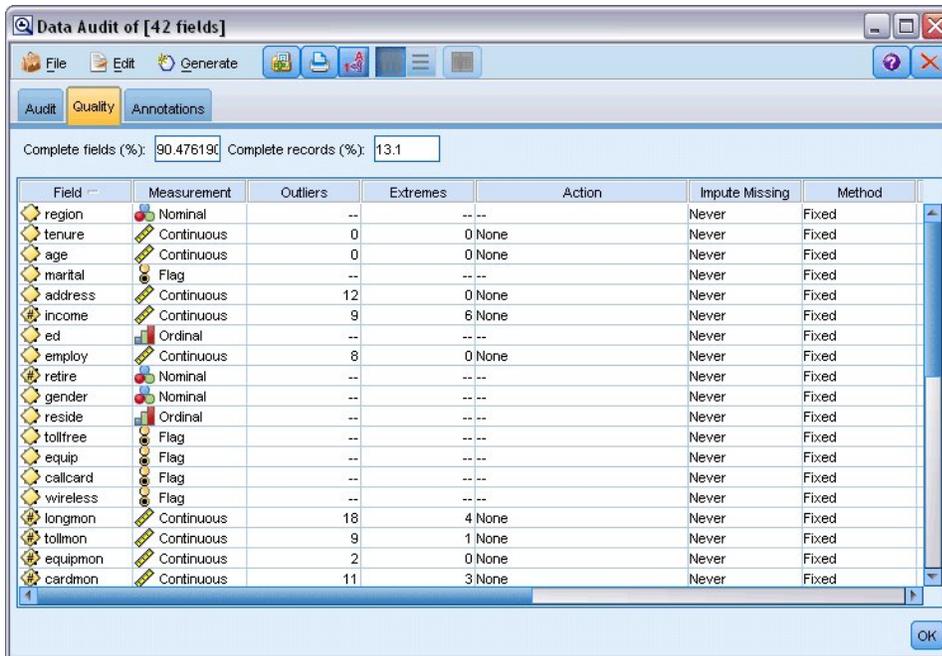


Abbildung 69. Data Audit-Browser - Registerkarte "Qualität"

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.

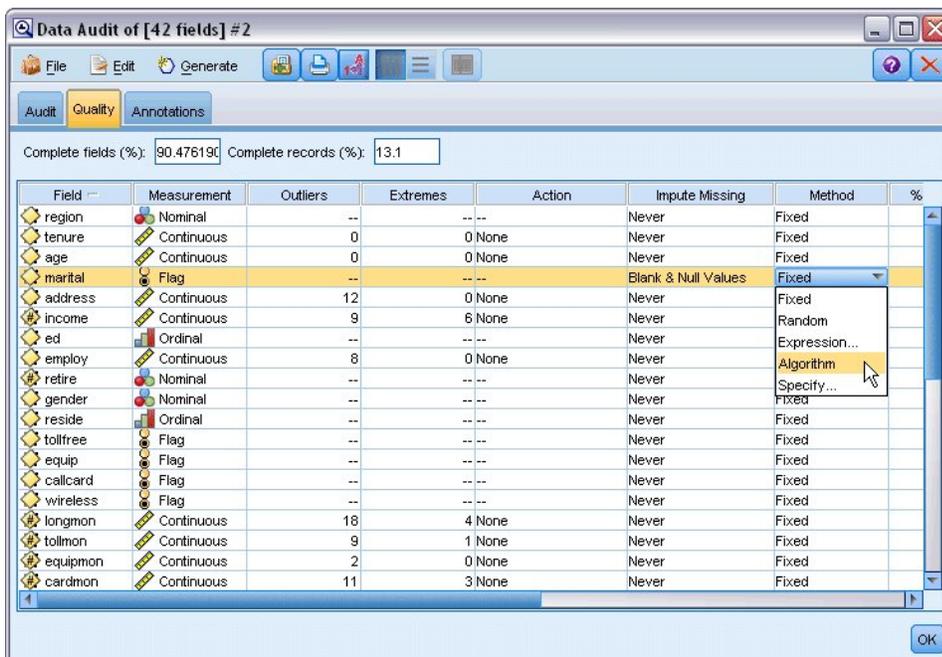


Abbildung 70. Auswahl einer Imputationsmethode

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:

Generieren > Superknoten für fehlende Werte

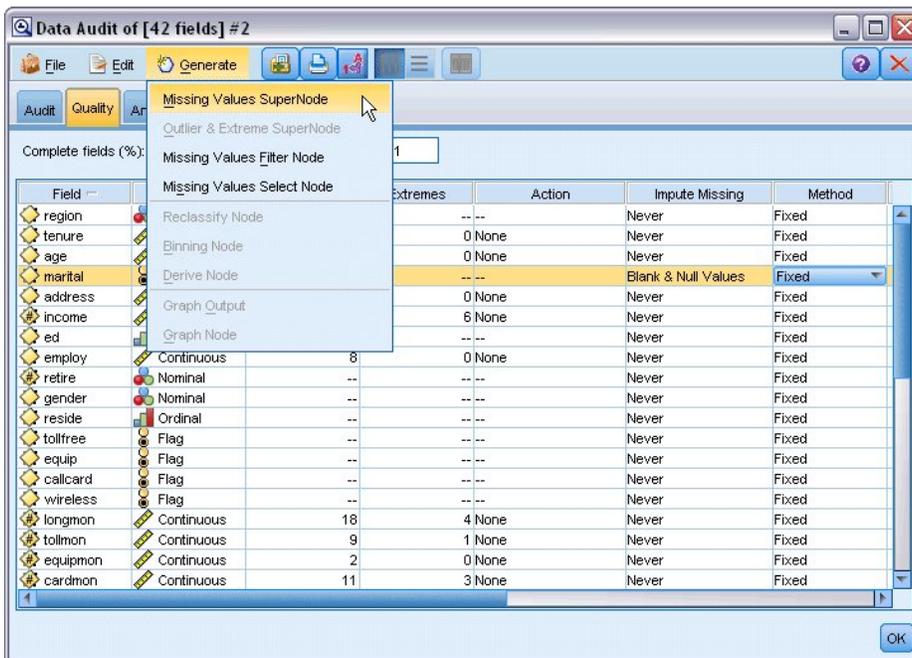


Abbildung 71. Generieren des Superknotens

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

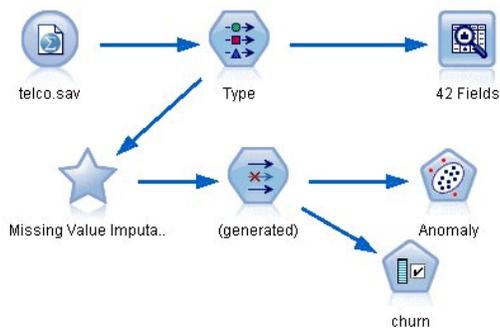


Abbildung 72. Stream mit Superknoten für fehlende Werte

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.

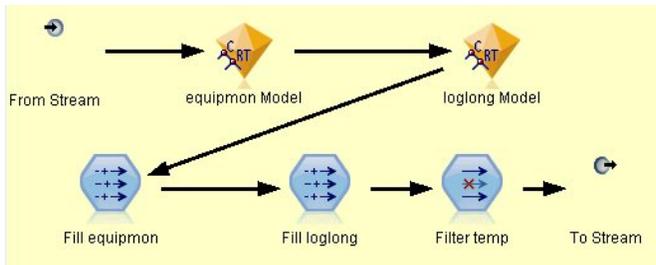


Abbildung 73. Vergrößern des Superknotens

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 74. 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 Nullwert setzen - und generieren Sie einen Superknoten zur Anwendung der Transformationen.

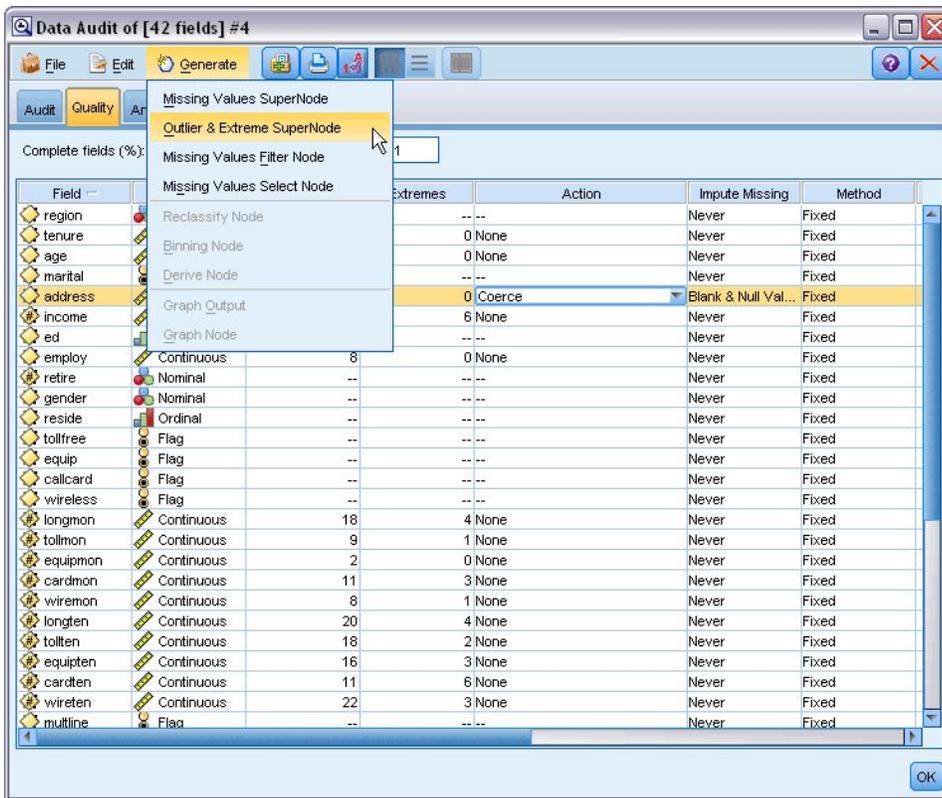


Abbildung 75. Generieren eines Filterknotens

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

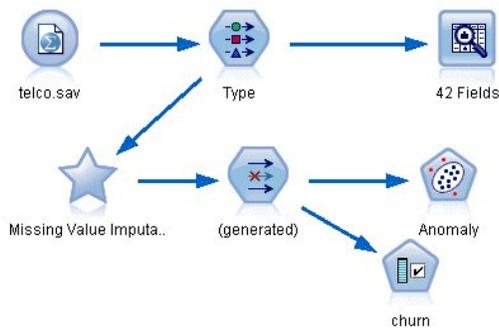


Abbildung 76. Stream mit Superknoten für fehlende Werte

Kapitel 8. 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 auf die Datendatei *DRUG1n* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. Die Datei *druglearn.str* befindet sich im Verzeichnis *streams*.

In dieser Demonstration werden die folgenden Datenfelder verwendet:

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

Einlesen von Textdaten



Abbildung 77. Hinzufügen eines Knotens vom Typ "Datei (var.)"

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.

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.

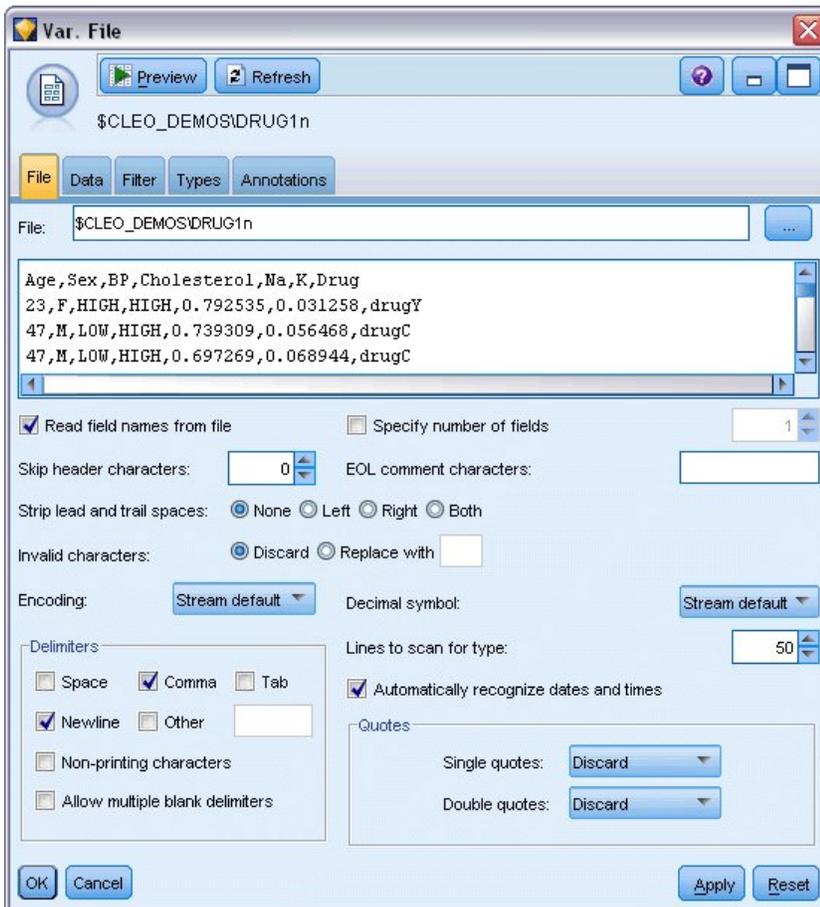


Abbildung 78. Dialogfeld "Datei (var.)"

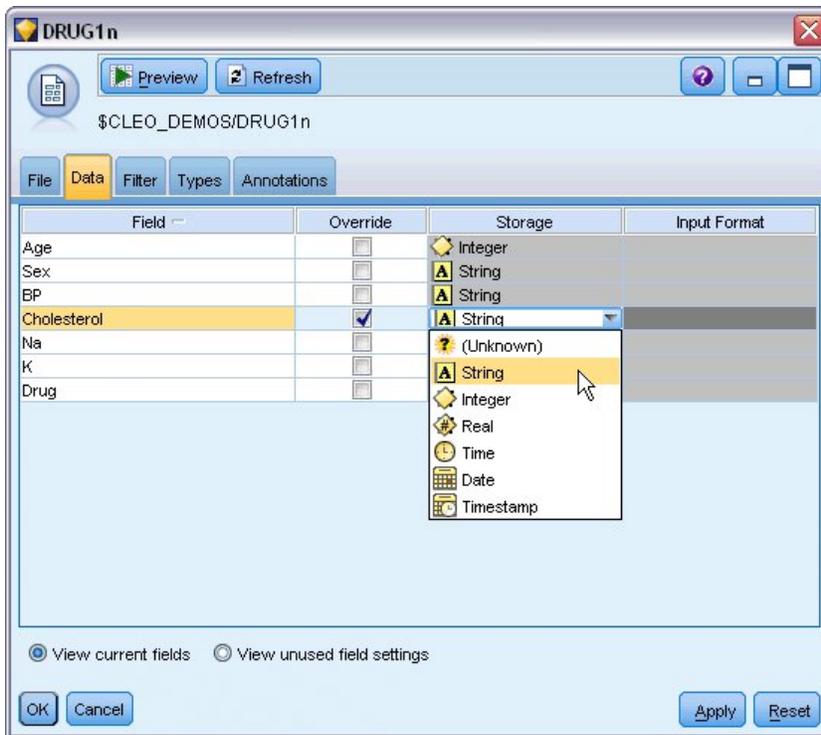


Abbildung 79. Ändern des Speichertyps für ein Feld

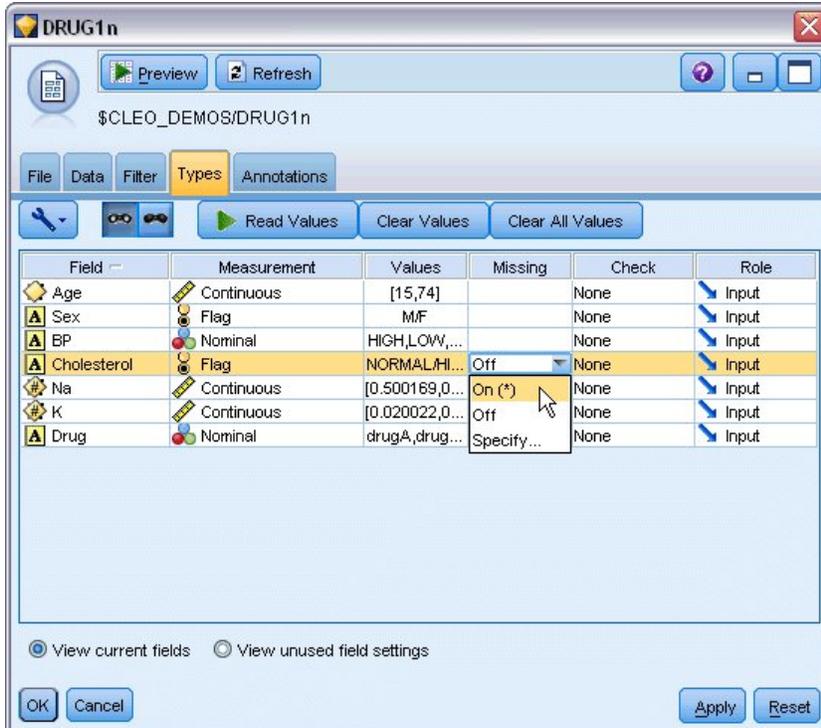


Abbildung 80. Auswählen von Werteoptionen auf der Registerkarte "Typen"

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 Feldty-

pen 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 auf den Erstellungsbereich und legen es dort ab.

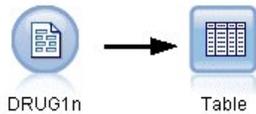


Abbildung 81. Mit der Datenquelle verbundener Tabellenknoten

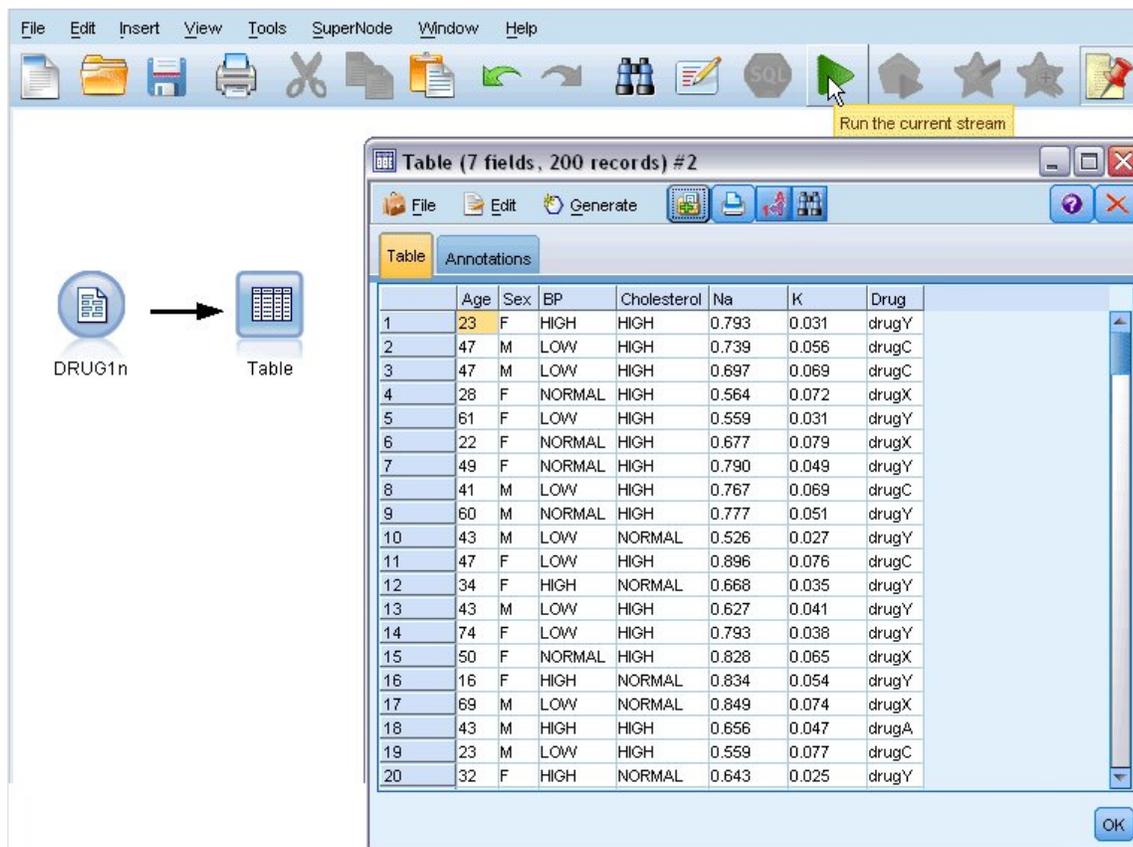


Abbildung 82. Ausführen eines Streams über die Symbolleiste

Wenn Sie auf einen Knoten in der Palette doppelklicken, wird dieser automatisch mit dem ausgewählten Knoten im Streamerstellungsbereich 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 Taste Alt 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 dem Stream einen Verteilungsknoten hinzu und verbinden Sie ihn mit dem Quellenknoten, doppelklicken Sie dann auf den Knoten, um die Anzeigeeoptionen zu bearbeiten.

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

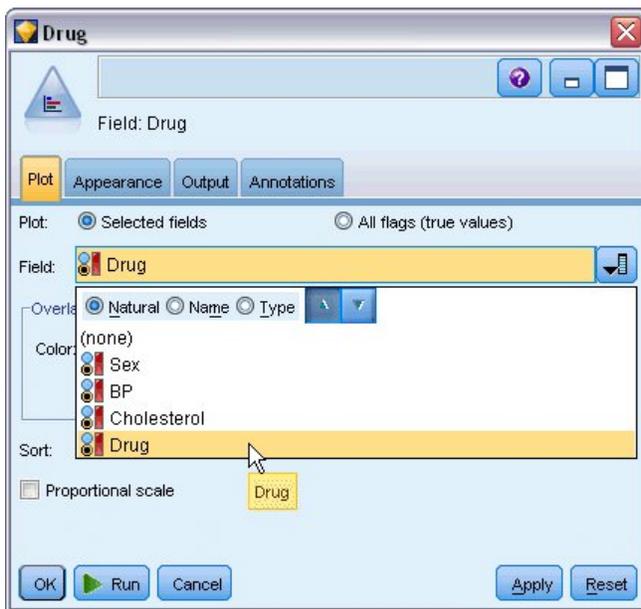


Abbildung 83. Auswählen von "drug" als Zielfeld

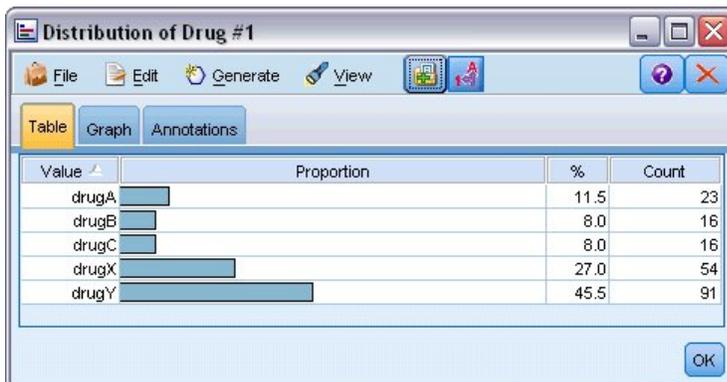


Abbildung 84. Verteilung der Ansprechquoten auf den Medikamententyp

Aus dem so entstandenen 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 ansprechen.

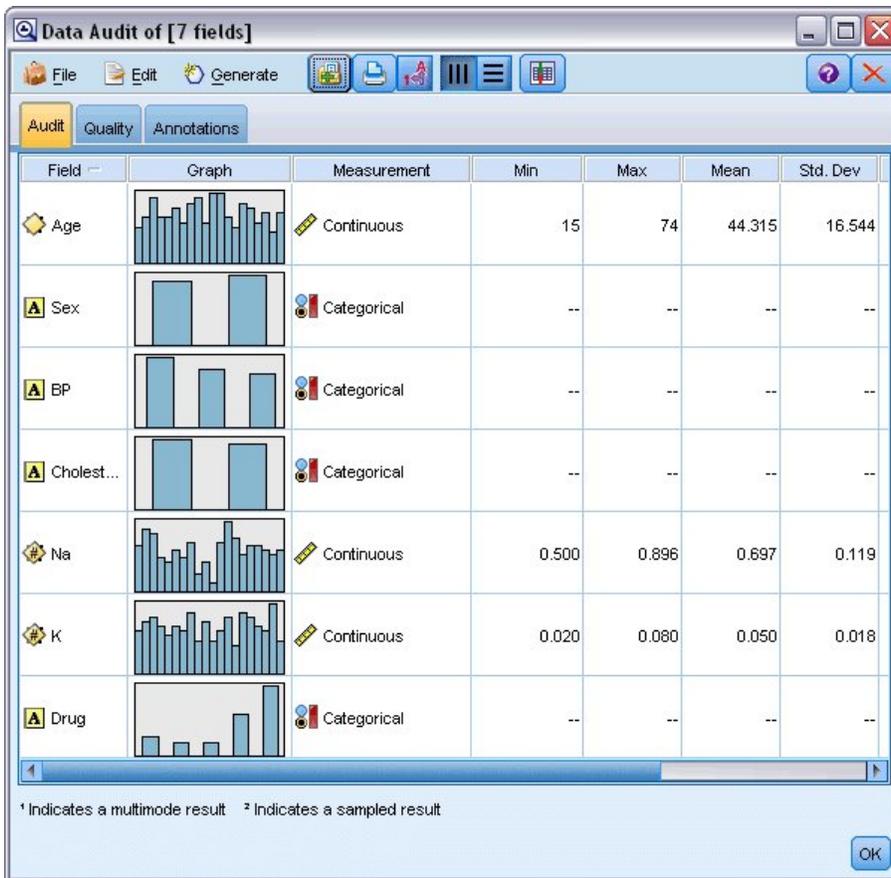


Abbildung 85. Ergebnisse eines Data Audit

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.

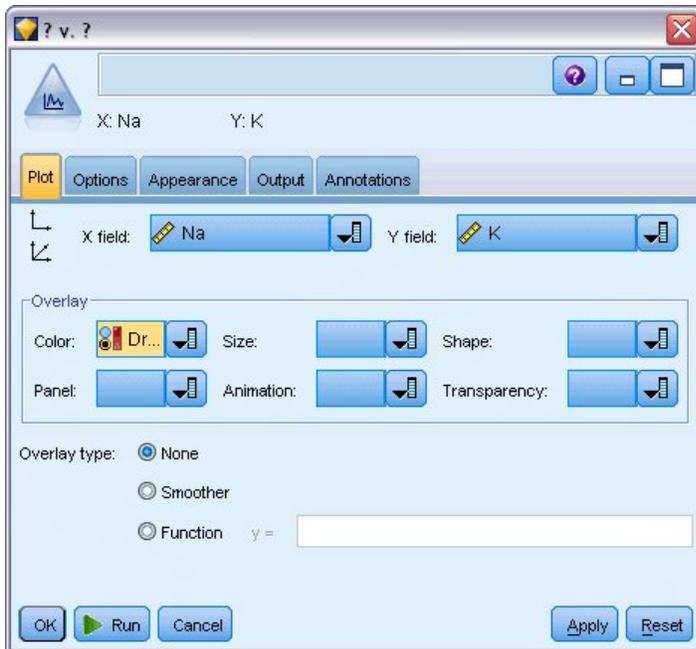


Abbildung 86. Erstellen eines Streudiagramms

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

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

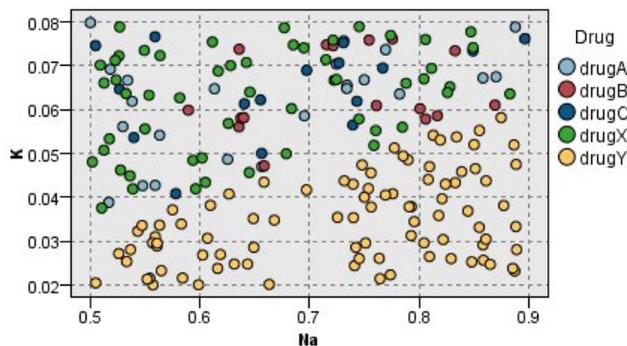


Abbildung 87. Streudiagramm der Medikamentenverteilung

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) aus. Klicken Sie dann auf **Ausführen**.

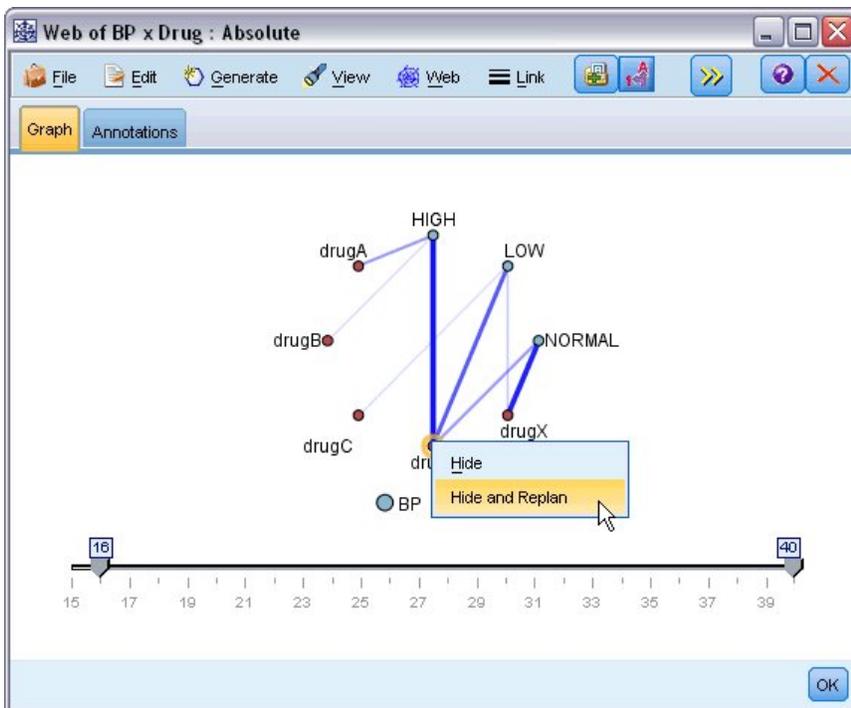


Abbildung 88. Netzdiagramm: Medikamente im Vergleich zum Blutdruck

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**.

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 sind 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.

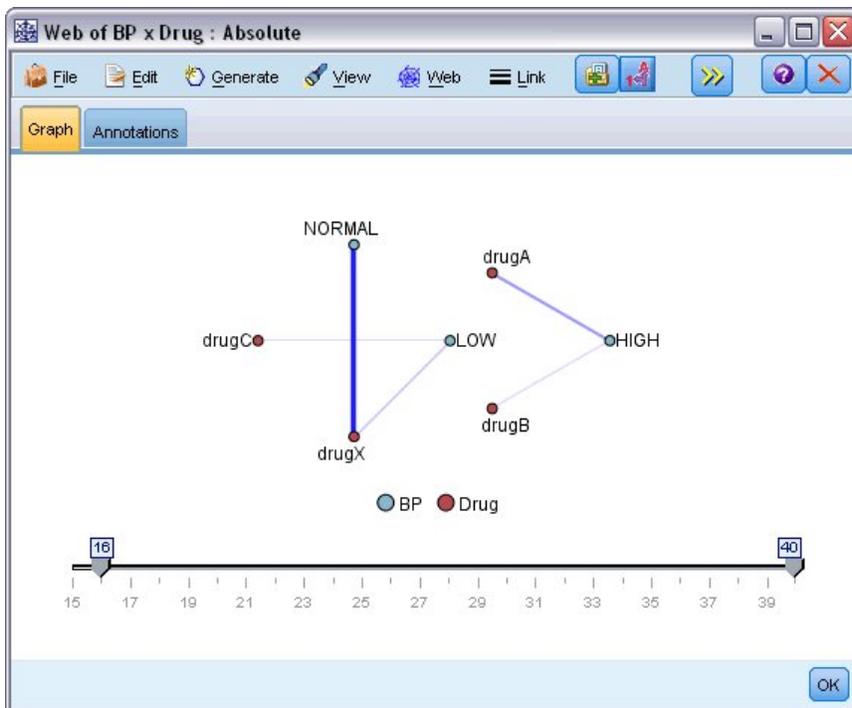


Abbildung 89. Netzdiagramm, Medikament Y und alle zugehörigen Verbindungen sind ausgeblendet

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 zur Voraussage, in welchen Fällen jedes der fünf Medikamente eingesetzt werden soll, ein Modell erstellen. Um das Stream-Layout zu vereinfachen, beginnen Sie damit, dass Sie alle Knoten mit Ausnahme des Quellenknotens DRUG1n löschen. Fügen Sie einen Ableitungsknoten (Registerkarte "Feldoperationen") an DRUG1n an und doppelklicken Sie dann auf den Ableitungsknoten, um ihn zu bearbeiten.

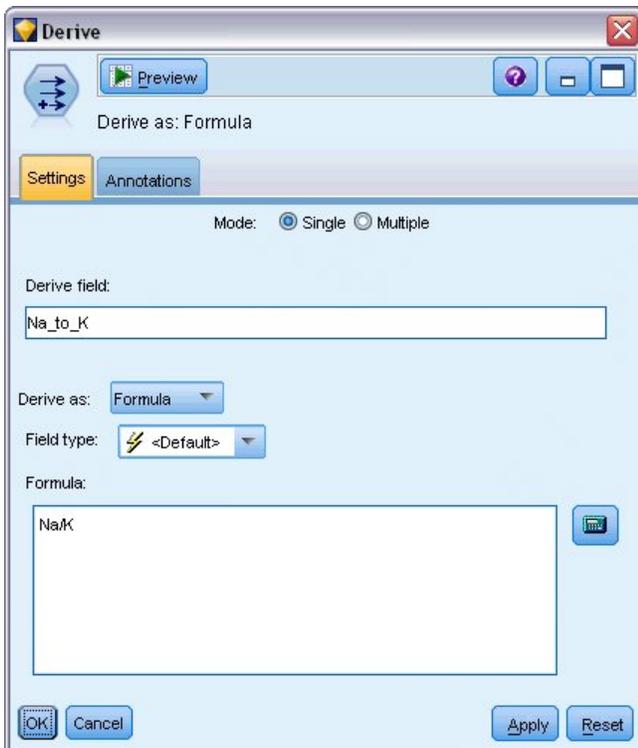


Abbildung 90. Bearbeiten des Ableitungsknotens

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. Hierdurch 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.

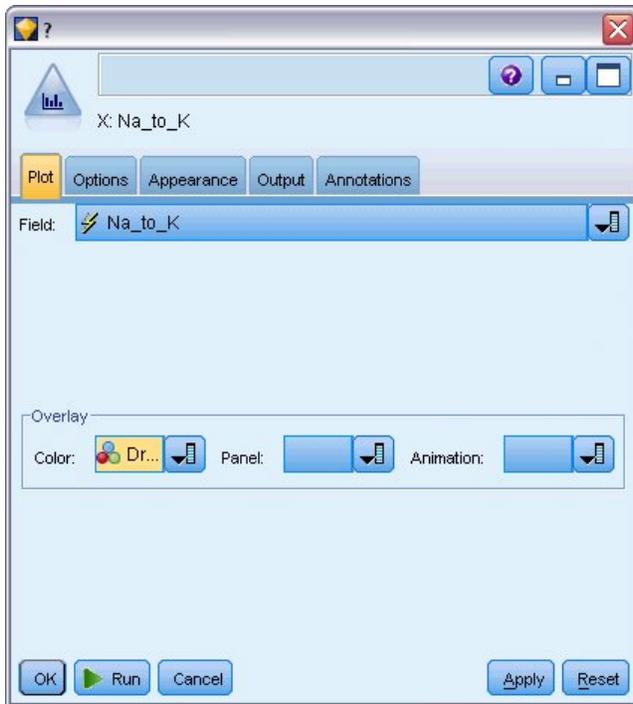


Abbildung 91. Bearbeiten des Histogrammknotens

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

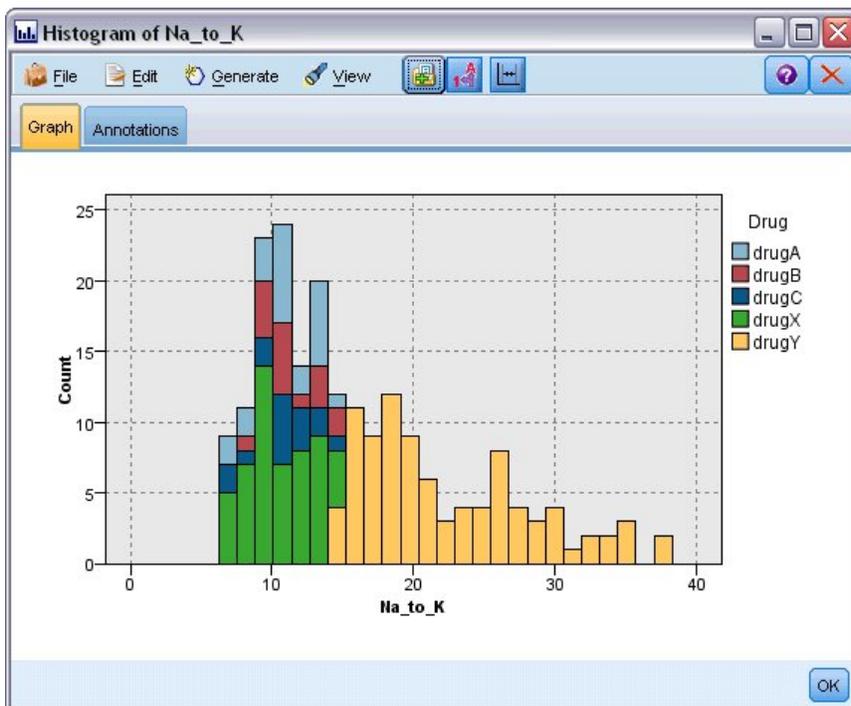


Abbildung 92. Histogrammanzeige

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.

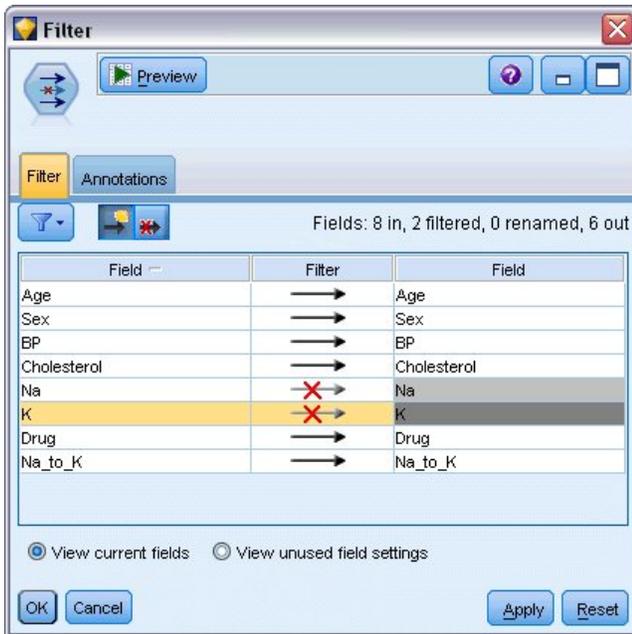


Abbildung 93. Bearbeiten des Filterknotens

Klicken Sie auf der Registerkarte "Filter" neben *Na* und *K* jeweils auf den angezeigten Pfeil. Über jedem Pfeil wird ein rotes X angezeigt, um anzuzeigen, 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. Übernehmen Sie für die anderen Felder die Rolle **Eingabe**, damit diese Felder als Prädiktor verwendet werden.

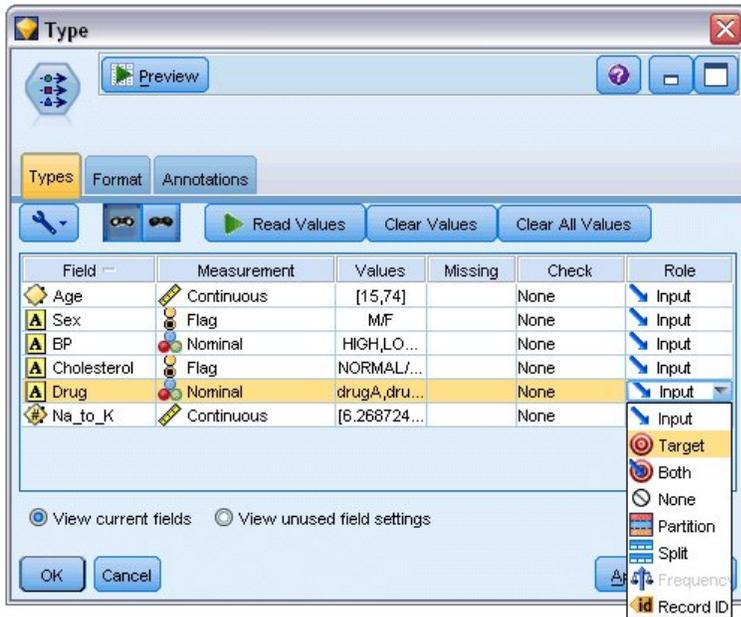


Abbildung 94. Bearbeiten des Typknotens

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.

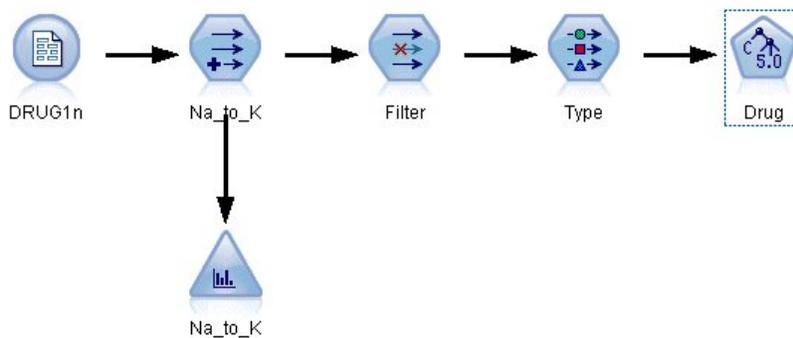


Abbildung 95. Hinzufügen eines C5.0-Knotens

Durchsuchen des Modells

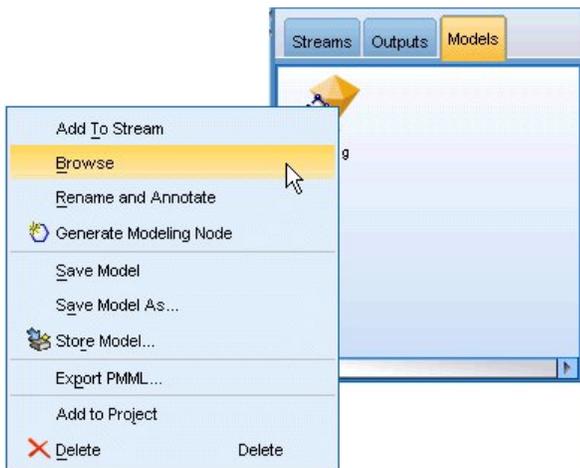


Abbildung 96. Durchsuchen des Modells

Wenn der Knoten C5.0 ausgeführt wird, wird das generierte Modellnugget dem Stream und der 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.

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**.

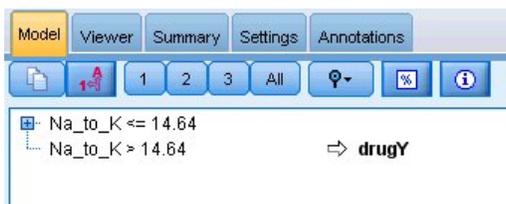


Abbildung 97. Regelbrowser

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.

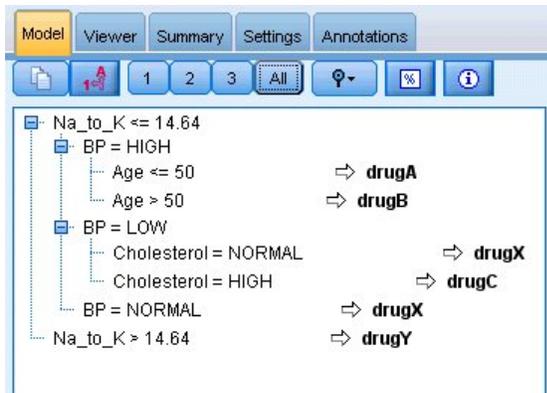


Abbildung 98. Vollständig erweiterter Regelbrowser

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.

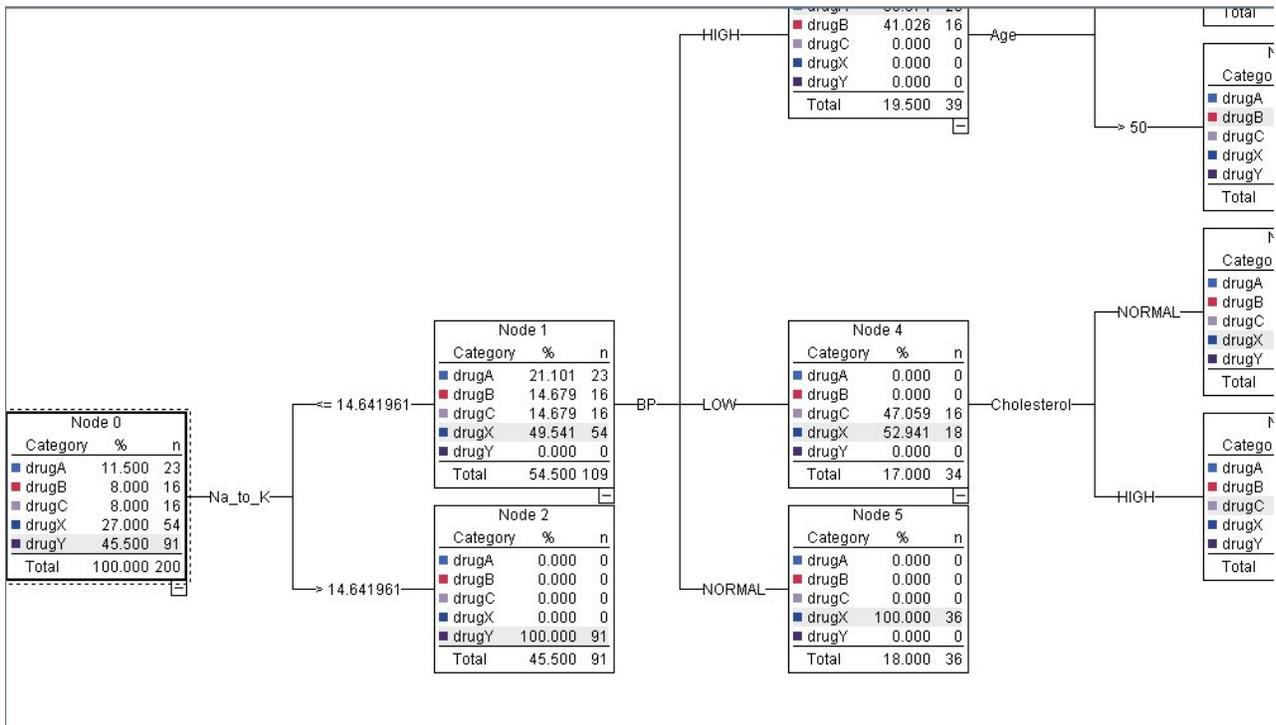


Abbildung 99. Entscheidungsbaum in grafischem Format

Verwenden eines Analyseknotts

Sie können die Genauigkeit des Modells mithilfe eines Analyseknotts bewerten. Verbinden Sie einen Analyseknott (aus der Ausgabeknotenpalette) mit dem Modellnugget, öffnen Sie den Analyseknott und klicken Sie auf **Ausführen**.

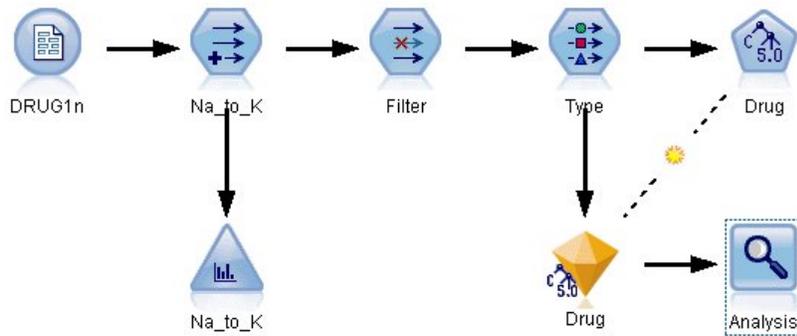


Abbildung 100. Hinzufügen eines Analyseknotts

Die Ausgabe des Analyseknotts zeigt, dass das Modell bei diesem künstlichen Dataset die Wahl des Medikaments für jeden Datensatz im Dataset korrekt vorhergesagt hat. Mit einem realen Dataset werden Sie kaum eine 100%ige Genauigkeit erreichen, Sie können jedoch mithilfe des Analyseknotts bestimmen, ob das Modell für Ihre spezielle Anwendung über eine ausreichende Genauigkeit verfügt.

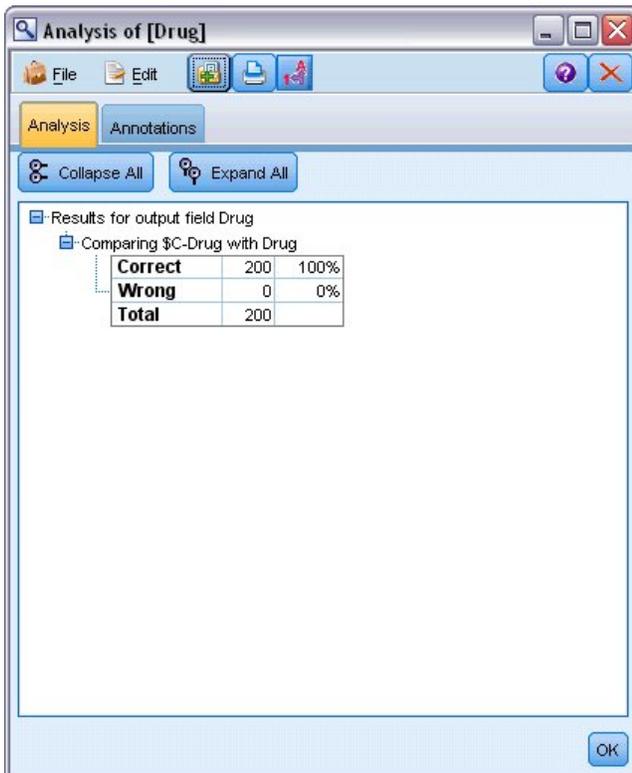


Abbildung 101. Analyseknott - Ausgabe

Kapitel 9. Screening von Prädiktoren (Merkmalauswahl)

Mit dem Merkmalauswahlknoten 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 Merkmalauswahlknoten 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 Prädiktoren 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 künftige ähnliche Angebote reagieren.

In diesem Beispiel wird ein Stream namens *featureselection.str* verwendet, der Bezug auf die Datendatei *customer_dbase.sav* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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 Merkmalauswahl. Alle Prädiktorfelder im Dataset dienen als Eingaben für den CHAID-Baum.
- Mit Merkmalauswahl. Der Merkmalauswahlknoten 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 Merkmalauswahl erzielt werden können.

Erstellen des Streams

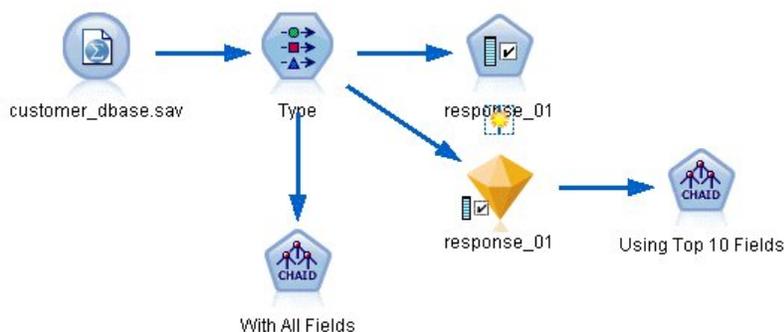


Abbildung 102. Beispielstream für die Merkmalauswahl

1. Platzieren Sie einen Quellenknoten für Statistikdateien in einem leeren Streamerstellungsbereich. 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 Beispielstreamdatei *featureselection.str* im Verzeichnis *streams* öffnen.)
2. 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**.

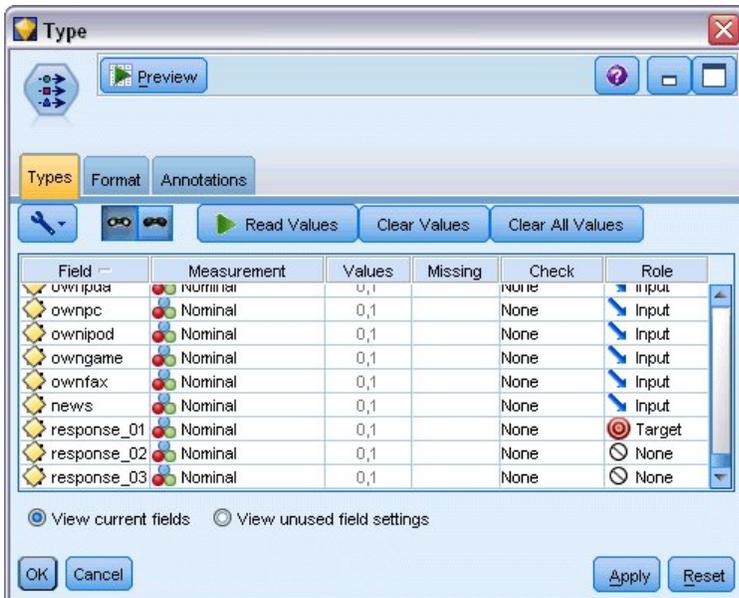


Abbildung 103. Hinzufügen eines Typknotens

3. Fügen Sie dem Stream einen Merkmalauswahlmodellierungsknoten hinzu. An diesem Knoten können Sie die Regeln und Kriterien für das Screening bzw. Ausschließen von Feldern angeben.
4. Führen Sie den Stream aus, um das Modellnugget "Merkmalauswahl" zu erstellen.
5. Klicken Sie mit der rechten Maustaste auf das Modellnugget im Stream oder in der Modellalette und wählen Sie **Bearbeiten** oder **Durchsuchen**, um die Ergebnisse zu betrachten.



Abbildung 104. Registerkarte "Modell" im Modellnugget "Merkmalauswahl"

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.

6. 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.
7. 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 Modellnugget.
8. Um Ergebnisse ohne Merkmalauswahl zu vergleichen, müssen Sie zwei CHAID-Modellierungsknoten in den Stream aufnehmen: einen, bei dem die Merkmalauswahl verwendet wird, und einen bei dem auf Merkmalauswahl verzichtet wird.
9. Verbinden Sie einen CHAID-Knoten mit dem Typknoten und den anderen mit dem generierten Modellnugget "Merkmalauswahl".

10. 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.

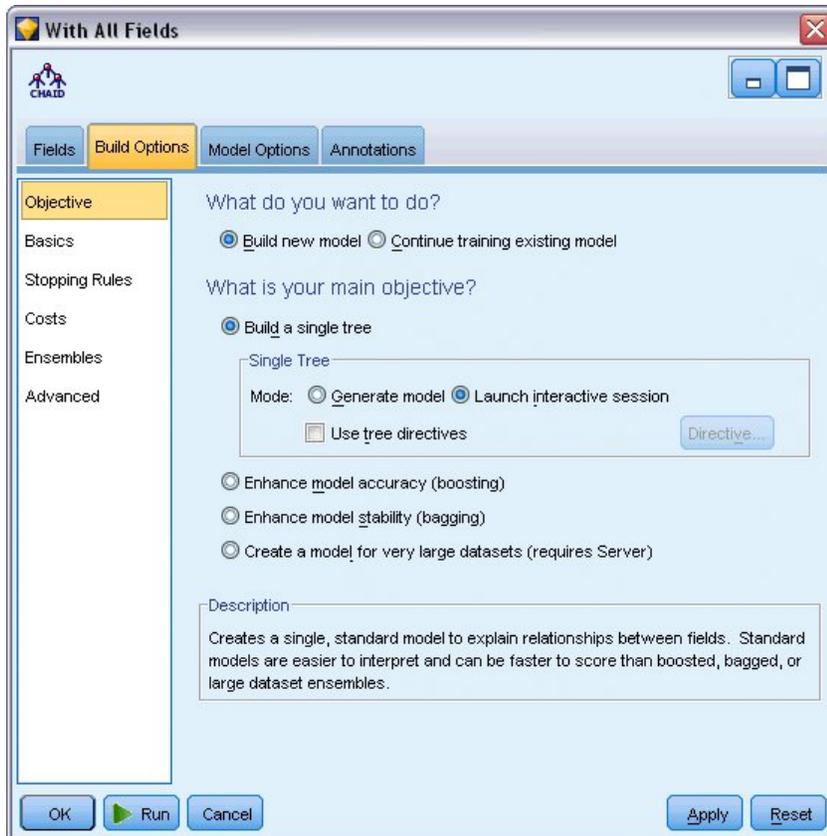


Abbildung 105. Zieleinstellungen für den CHAID-Modellierungsknoten für alle Prädiktorfelder

Erstellen der Modelle

1. Führen Sie den CHAID-Knoten aus, der alle Prädiktoren im Dataset verwendet (den Knoten, der mit dem Typknoten verbunden ist). Achten Sie darauf, wie lange die Ausführung dauert. Im Ergebnisfenster wird eine Tabelle angezeigt.
2. Wählen Sie in den Menüs die Optionsfolge **Baum > Baum erweitern** aus, um den Baum zu erweitern und anzuzeigen.

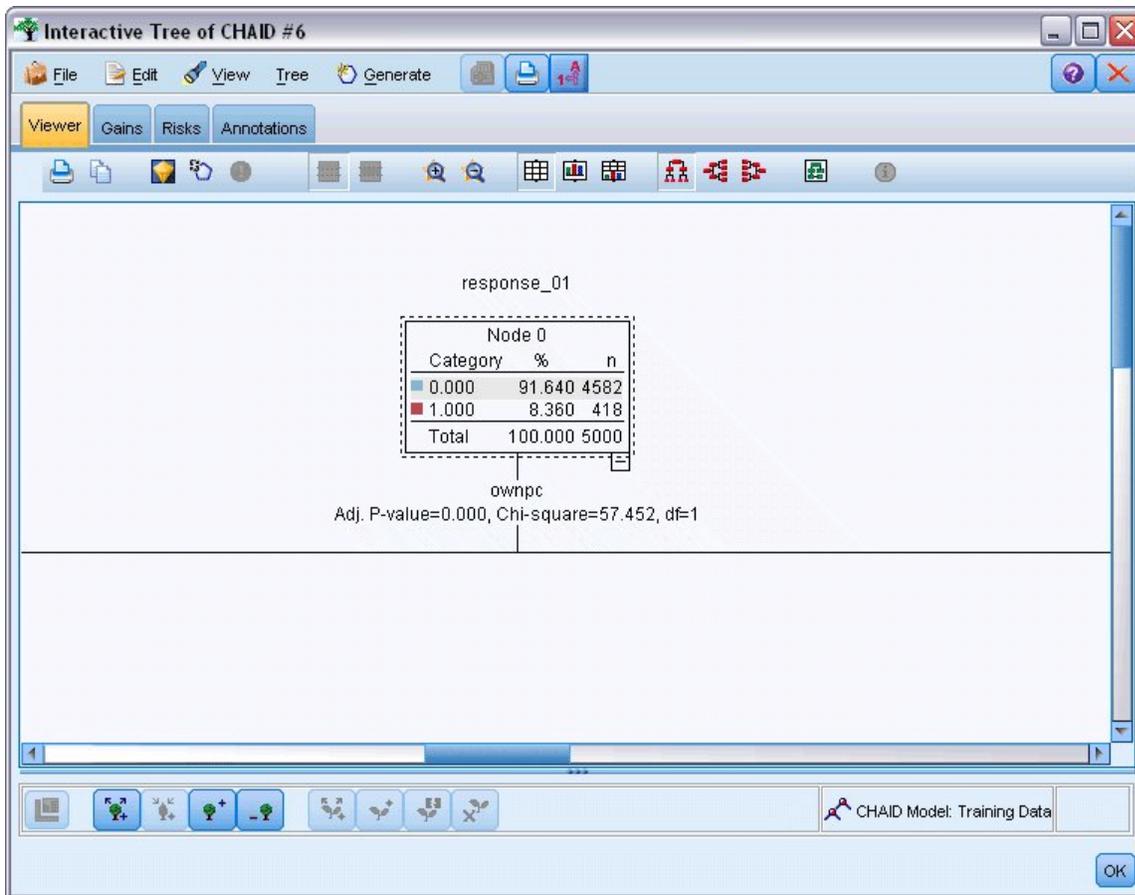


Abbildung 106. Erweitern des Baums im Tree Builder

3. 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 Dataset recht klein ist, beträgt der Unterschied in der Ausführungsdauer vermutlich nur wenige Sekunden. Bei größeren, realen Datensets kann der Unterschied jedoch beträchtlich sein - es kann sich um Minuten oder sogar Stunden handeln. Mithilfe der Merkmalauswahl 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 - im Vergleich zu allen Datensätzen im Dataset - in die Zielkategorie fallen. Ein Liftwert von 148 % beispielsweise gibt an, dass die Datensätze im Knoten mit 1,48-mal höherer Wahrscheinlichkeit in die Zielkategorie fallen als alle Datensätze im Dataset. Der Lift wird auf der Registerkarte "Gewinne" in der Spalte *Index* angegeben.

1. Klicken Sie im Tree Builder für das vollständige Prädiktorenset auf die Registerkarte "Gewinne". Ändern Sie die Zielkategorie in 1,0. Ändern Sie die Anzeige in Quartile. Klicken Sie dazu zuerst auf die Symbolleistschaltfläche "Quartile". Wählen Sie anschließend im Dropdown-Menü neben dieser Schaltfläche die Option **Quartil** aus.
2. 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).

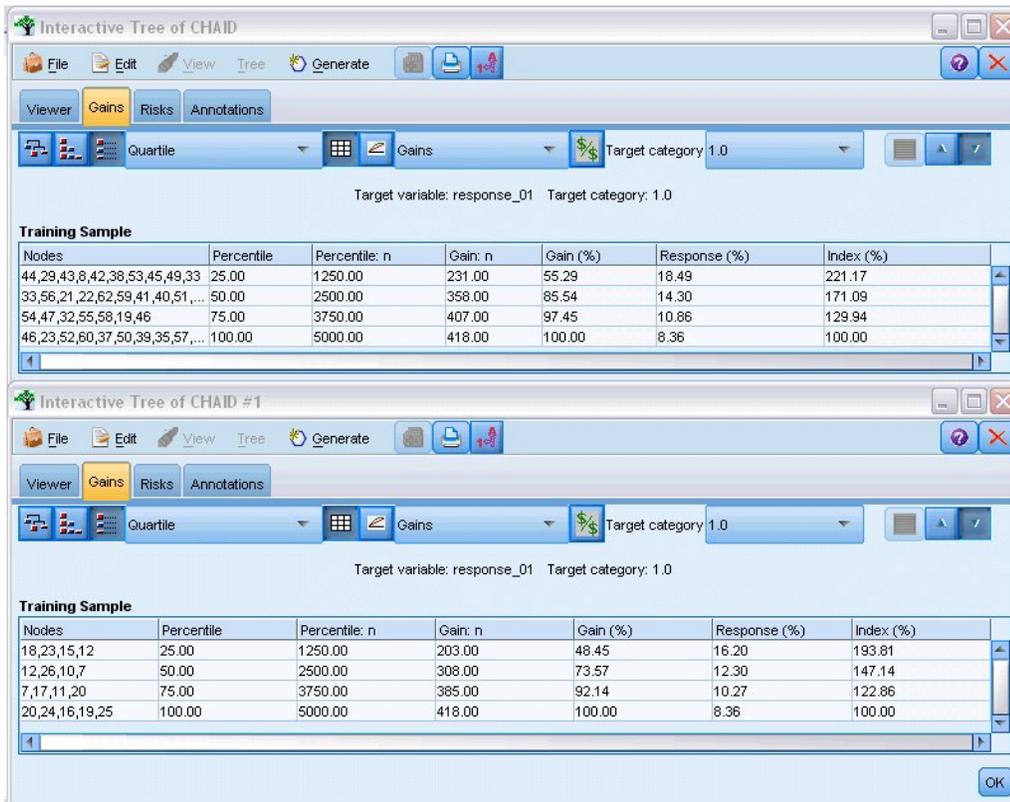


Abbildung 107. Gewinntabellen für beide CHAID-Modelle

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 festzustellen, 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 festzustellen, 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 Merkmalauswahl 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 Merkmalauswahlmodell dem Modell mit allen Prädiktoren vorzuziehen ist.

Zusammenfassung

Fassen wir die Vorteile der Merkmalauswahl 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 Merkmalauswahlschritts die Modellerstellung mit dem kleineren Prädiktorenset merklich schneller. Mit einem größeren realen Dataset 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 Datensets verallgemeinern (allerdings müsste dies sicherheitshalber getestet werden).

Sie könnten einen Baumerstellungsalgorithmus für die Merkmalauswahl 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 Merkmalauswahlknoten 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 10. Reduzieren der Länge der Zeichenfolge für die Eingabedaten (Umcodierungsknoten)

Reduzieren der Länge der Zeichenfolge für die Eingabedaten (Umcodierung)

Bei Modellen vom Typ "Binomiale logistische Regression" und Modellen vom Typ "Automatisches Klassifikationsmerkmal", die ein Modell vom Typ "Binomiale logistische Regression" enthalten, sind die Zeichenfolgenfelder auf maximal acht Zeichen begrenzt. Zeichenfolgen mit mehr als acht Zeichen können mithilfe eines Umcodierungsknotens neu codiert werden.

In diesem Beispiel wird ein Stream namens *reclassify_strings.str* verwendet, der Bezug auf die Datendatei *drug_long_name* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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 Zeichenfolgen entstehen können. Außerdem wird erläutert, wie die Zeichenfolgendetails mithilfe des Umcodierungsknotens 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 "Automatisches Klassifikationsmerkmal" ein Modell vom Typ "Binomiale logistische Regression" erstellt wird.

Umcodieren der Daten

1. Stellen Sie mithilfe eines Quellenknotens für variable Dateien eine Verbindung mit dem Dataset *drug_long_name* im Ordner *Demos* her.

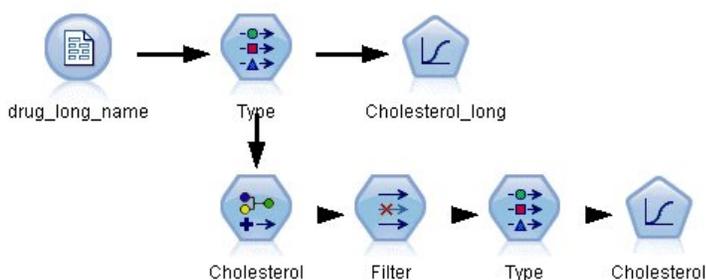


Abbildung 108. Beispielstream zur Umcodierung von Zeichenfolgen für die binomiale logistische Regression

2. Fügen Sie dem Quellenknoten einen Typknoten hinzu und wählen Sie **Cholesterol_long** als Ziel aus.
3. Fügen Sie dem Typknoten einen Knoten vom Typ "Logistische Regression" hinzu.
4. Klicken Sie im Knoten "Logistische Regression" auf die Registerkarte "Modell" und wählen Sie die Prozedur **Binomial** aus.

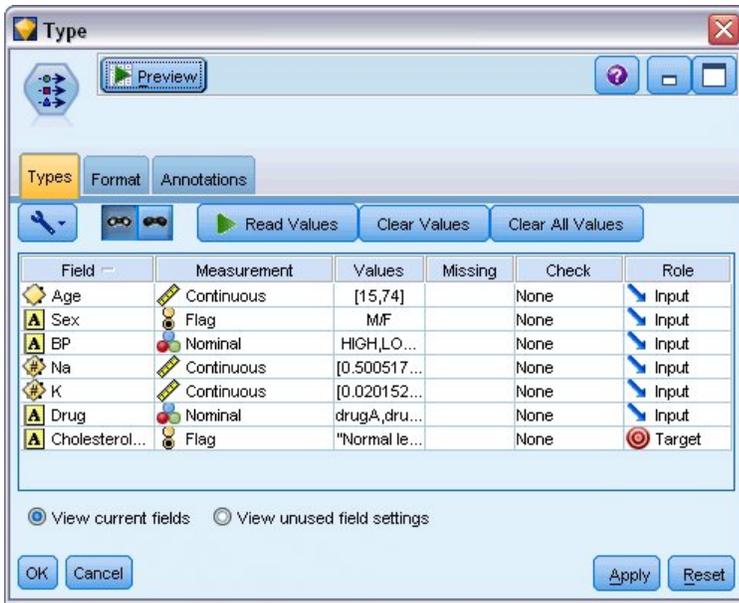


Abbildung 109. Lange Zeichenfolgendetails im Feld "Cholesterol_long"

- Wenn Sie den Knoten "Logistische Regression" in *reclassify_strings.str* ausführen, wird eine Fehlermeldung angezeigt, die Sie darauf hinweist, dass die Werte der Zeichenfolge **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.

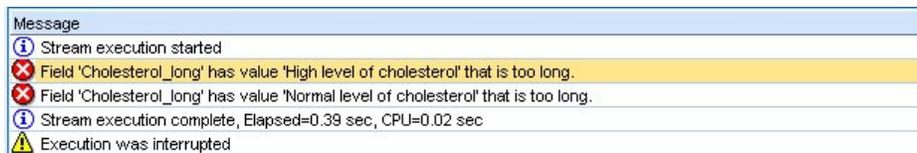


Abbildung 110. Fehlermeldung bei Ausführung des Knotens "Binomiale logistische Regression"

- Fügen Sie dem Typknoten einen Umcodierungsknoten hinzu.
- Wählen Sie im Feld "Umcodieren" den Eintrag **Cholesterol_long** aus.
- Geben Sie **Cholesterol** als neuen Feldnamen ein.
- Klicken Sie auf die Schaltfläche **Ermitteln**, um die Werte von **Cholesterol_long** der Spalte "Ursprünglicher Wert" hinzuzufügen.
- Geben Sie in der Spalte "Neuer Wert" den Wert **High** neben dem ursprünglichen Wert **High level of cholesterol** und den Wert **Normal** neben dem ursprünglichen Wert **Normal level of cholesterol** ein.



Abbildung 111. Umcodieren der langen Zeichenfolgen

11. Fügen Sie dem Umcodierungsknoten einen Filterknoten hinzu.
12. Klicken Sie in der Spalte "Filter", um **Cholesterol_long** zu entfernen.

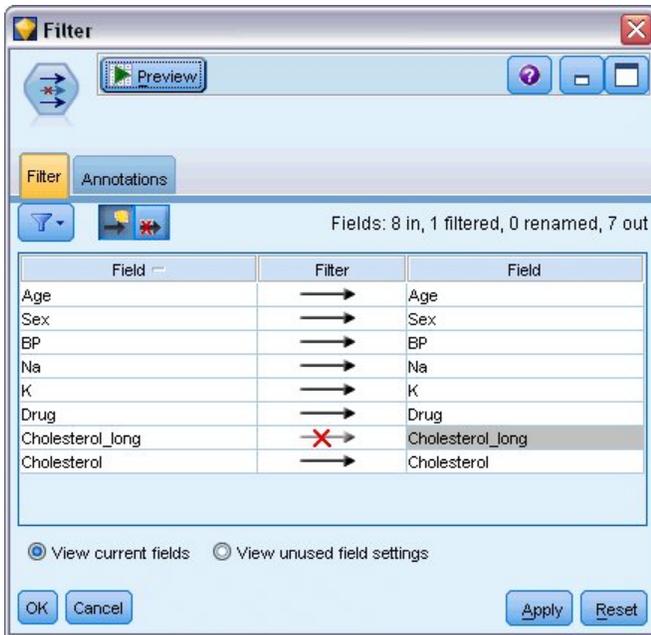


Abbildung 112. Filtern des Felds "Cholesterol_long" aus den Daten

13. Fügen Sie dem Filterknoten einen Typknoten hinzu und wählen Sie **Cholesterol** als Ziel aus.

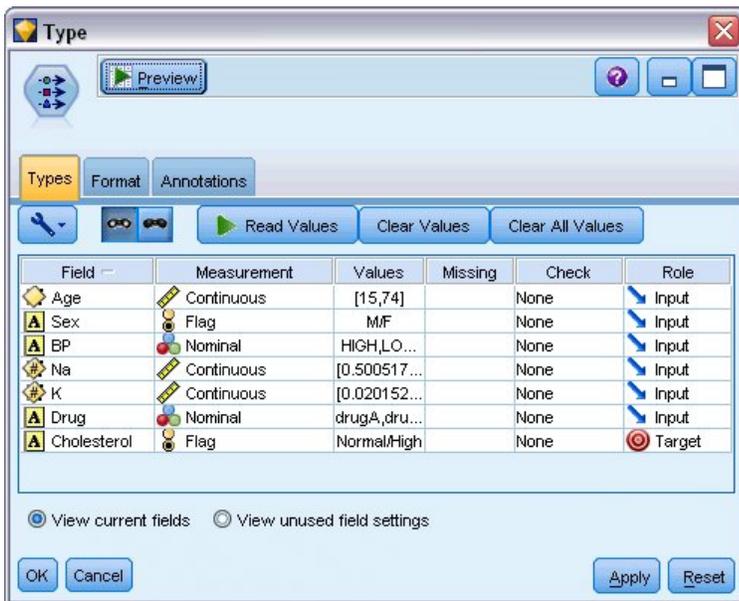


Abbildung 113. Kurze Zeichenfolgendetails im Feld "Cholesterol"

14. Fügen Sie dem Typknoten einen Logistikknoten hinzu.
15. Klicken Sie im Logistikknoten auf die Registerkarte "Modell" und wählen Sie die Prozedur **Binomial** aus.
16. Sie können nun den binomialen Logistikknoten ausführen und ein Modell generieren, ohne dass eine Fehlernachricht angezeigt wird.

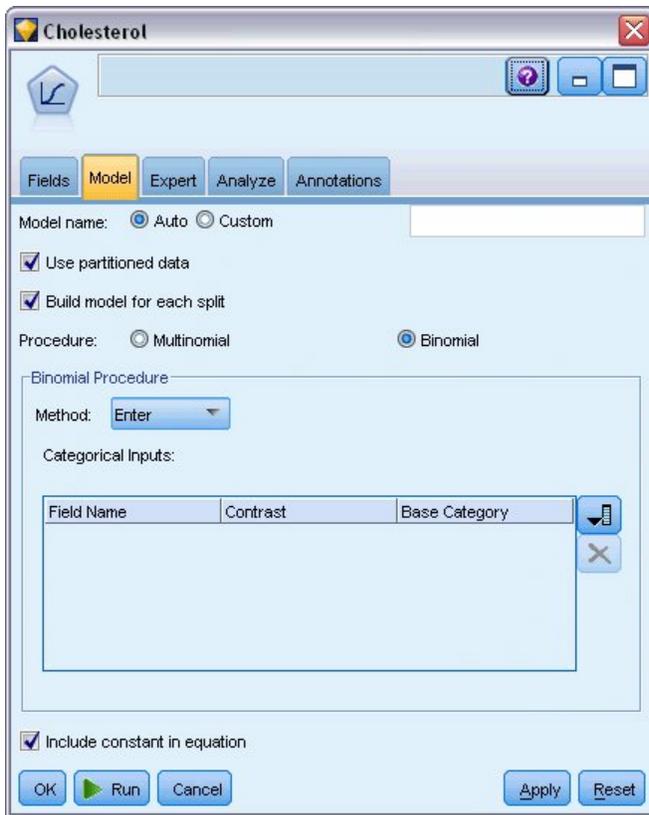


Abbildung 114. Auswählen von Binomial als Prozedur

Dieses Beispiel zeigt nur einen Teil eines Streams. Wenn Sie weitere Informationen über die Streamtypen benötigen, bei denen eine Umcodierung langer Zeichenfolgen erforderlich sein kann, stehen Ihnen folgende Beispiele zur Verfügung:

- Knoten "Automatisches Klassifikationsmerkmal". Weitere Informationen finden Sie unter im Thema „Modellieren der Kundenreaktion (Automatisches Klassifikationsmerkmal)“ auf Seite 37.
- Knoten "Binomiale logistische Regression". Weitere Informationen finden Sie unter im Thema Kapitel 13, „Kundenabwanderung bei Telekommunikationsunternehmen (binomiale logistische Regression)“, auf Seite 139.

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.

Kapitel 11. Modellieren der Kundenreaktion (Entscheidungsliste)

Der Entscheidungslistenalgorithmus 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 "Automatisches Klassifikationsmerkmal" erstellt werden.

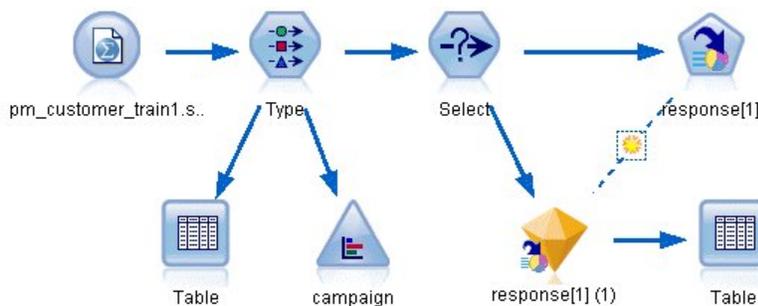


Abbildung 115. Beispielstream für Entscheidungsliste

In diesem Beispiel wird ein Stream namens *pm_decisionlist.str* verwendet, der Bezug auf die Datendatei *pm_customer_train1.sav* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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).

	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\$

Abbildung 116. Daten zu früheren Werbeaktionen

Die Werte des Felds *campaign* (Kampagne) sind in den Daten tatsächlich als ganze Zahlen codiert. Die Beschriftungen sind im Typknoten definiert (Beispiel: 2 = *Premium account*). Sie können die Anzeige der Wertbeschriftungen 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

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf die Datei *pm_customer_train1.sav* im Ordner *Demos* Ihrer IBM SPSS Modeler-Installation verweist. (Sie können `%CLE0_DEMOS%` im Dateipfad als Verknüpfung zur Referenzierung dieses Ordners angeben.)

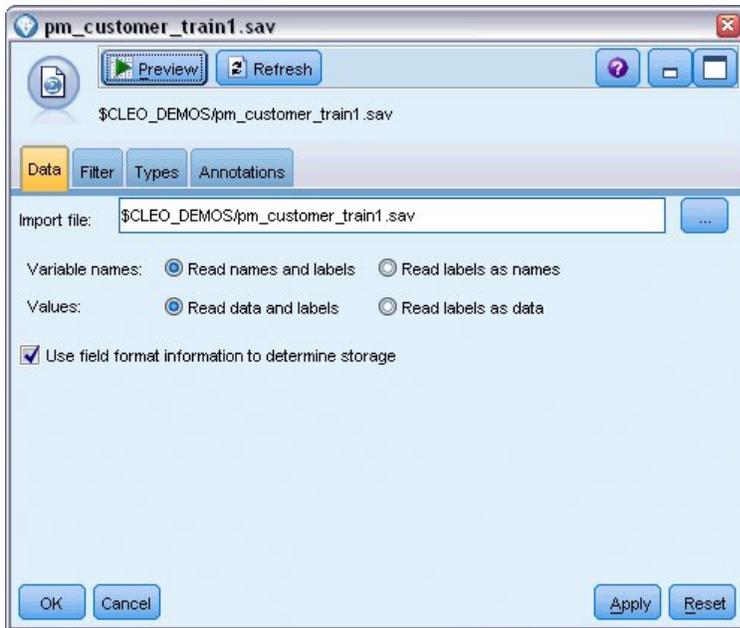


Abbildung 117. Einlesen der Daten

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

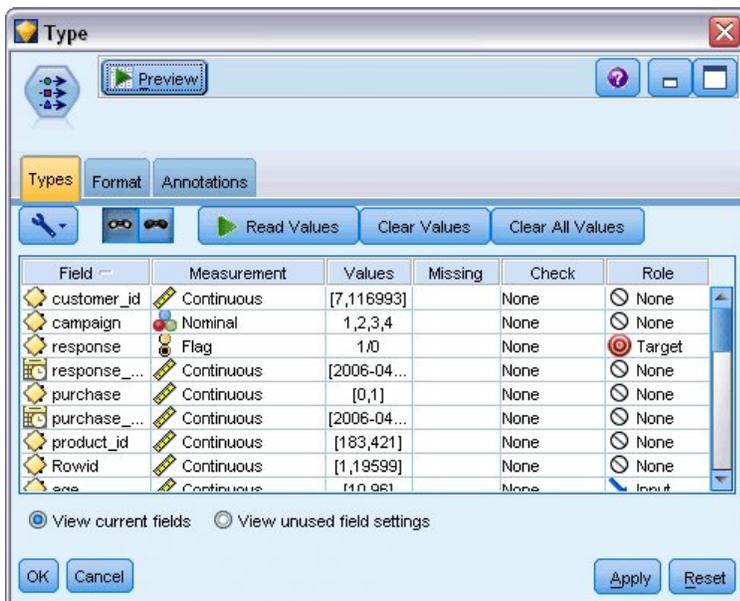


Abbildung 118. Festlegen von Messniveau und Rolle

3. Setzen Sie die Rolle für die folgenden Felder auf **Keine**: *customer_id* (Kunden-ID), *campaign* (Kampagne), *response_date* (Antwortdatum), *purchase* (Einkauf), *purchase_date* (Einkaufsdatum), *product_id* (Produkt-ID), *Rowid* (Zeilen-ID) und *X_random* (X-Zufall). Diese Felder dienen jeweils in den Daten bestimmten Zwecken, werden jedoch nicht zum Erstellen des tatsächlichen Modells verwendet.
4. 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 codiert als *campaign=2*), können Sie einen Auswahlknoten verwenden, um nur die betreffenden Datensätze in den Stream aufzunehmen.

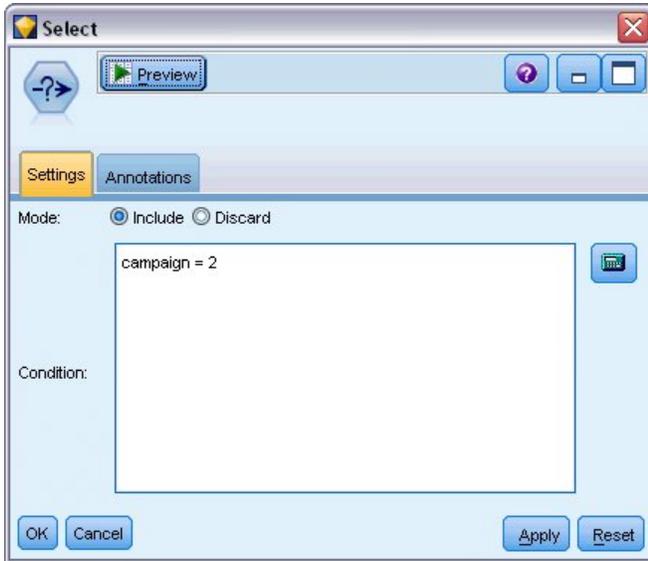


Abbildung 119. Auswählen von Datensätzen für eine einzelne Kampagne

Erstellen des Modells

1. Fügen Sie dem Stream einen Entscheidungslistenknoten 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.

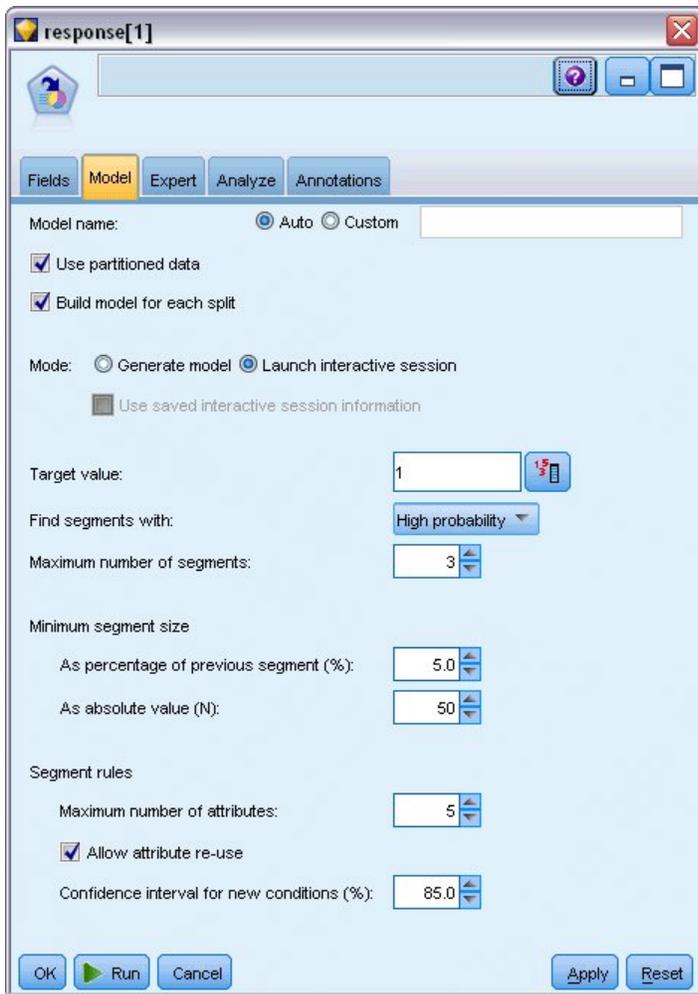


Abbildung 120. Entscheidungslistenknoten - Registerkarte "Modell"

2. Wählen Sie **Interaktive Sitzung starten** aus.
3. Um das Modell für dieses Beispiel einfach zu halten, geben Sie als maximale Anzahl an Segmenten den Wert 3 an.
4. Ändern Sie das Konfidenzintervall für neue Bedingungen auf 85 %.
5. Setzen Sie auf der Registerkarte "Experten" den Wert von **Modus** auf **Experten**.

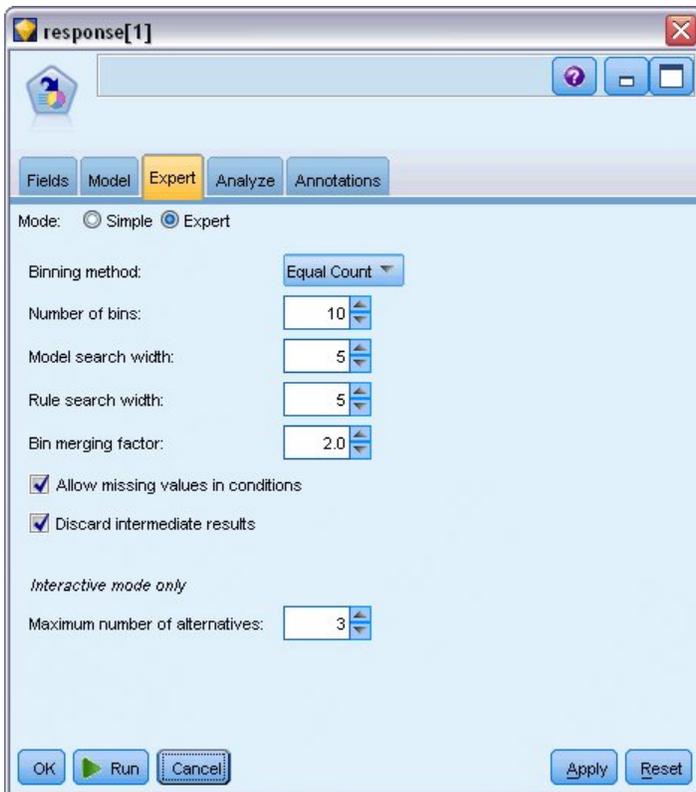


Abbildung 121. Entscheidungslistenknoten - Registerkarte "Experten"

6. 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.
7. Klicken Sie auf **Ausführen**, um den Viewer "Interaktive Liste" anzuzeigen.

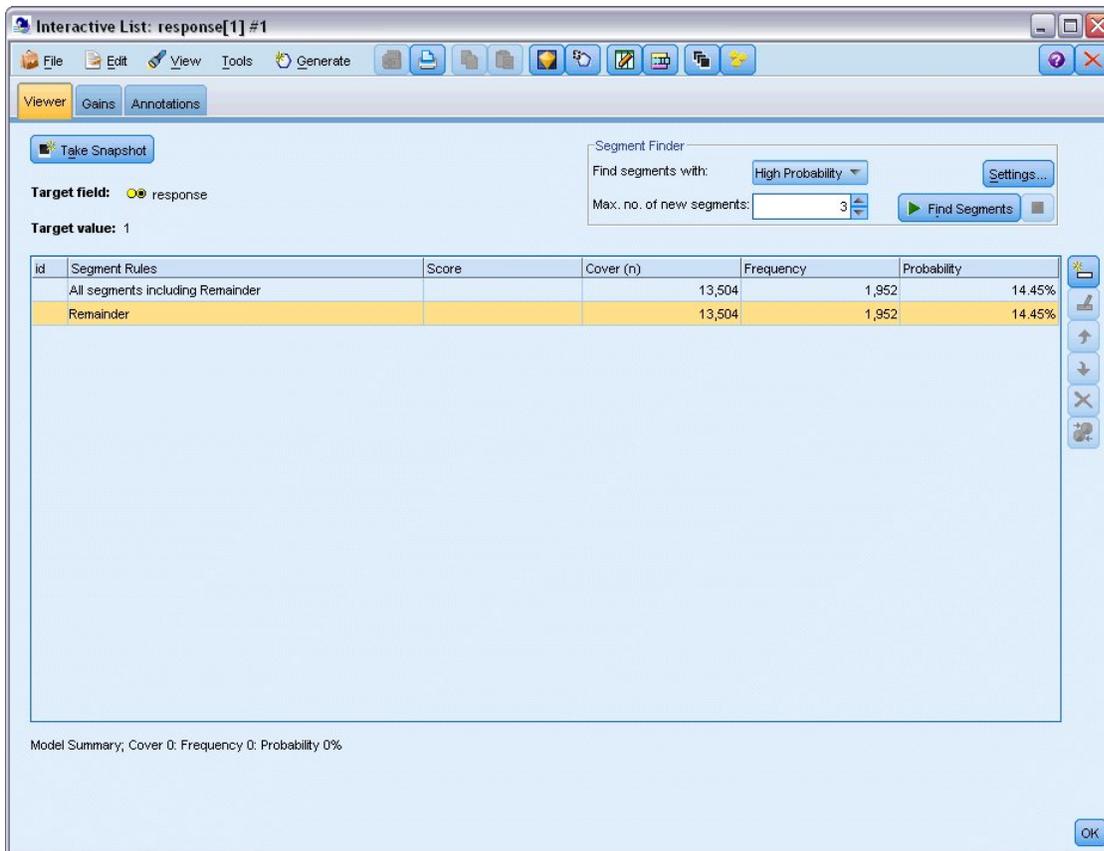


Abbildung 122. Viewer "Interaktive Liste"

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.

8. Wählen Sie im Viewer "Interaktive Liste" folgende Optionsfolge aus den Menüs aus:

Tools > Segmente suchen

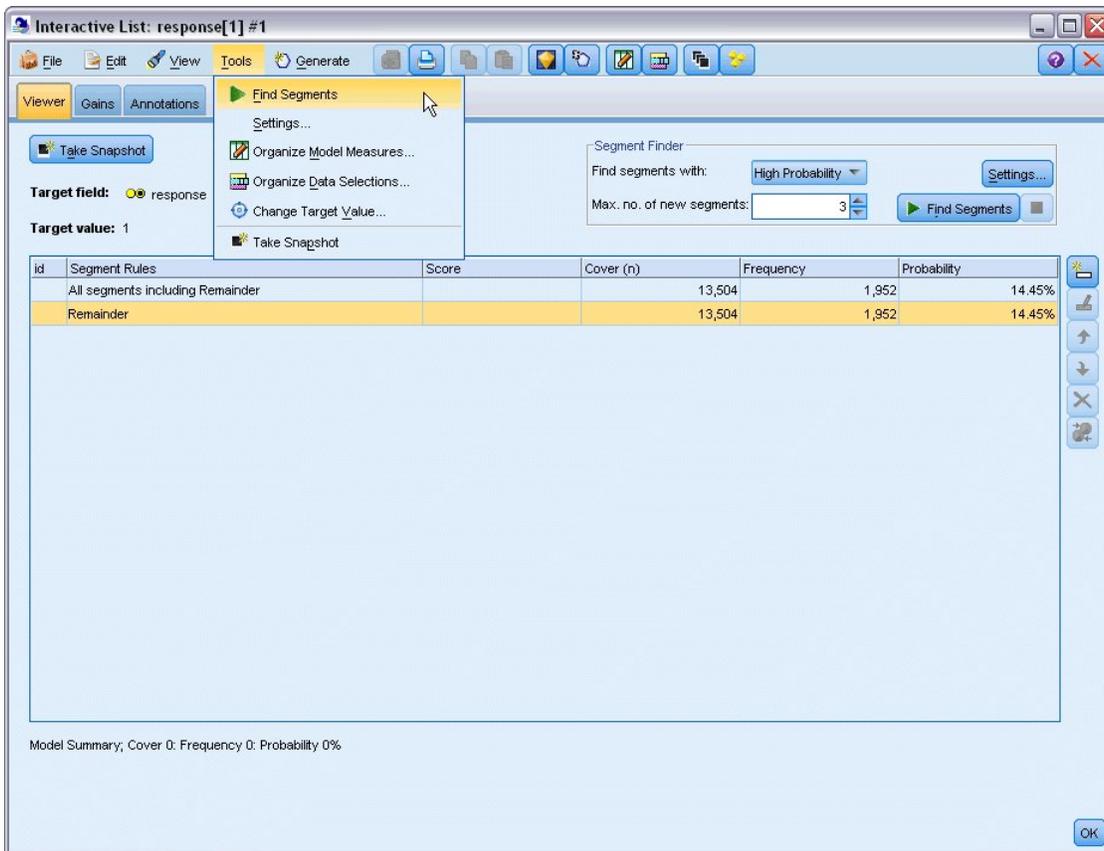


Abbildung 123. Viewer "Interaktive Liste"

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.

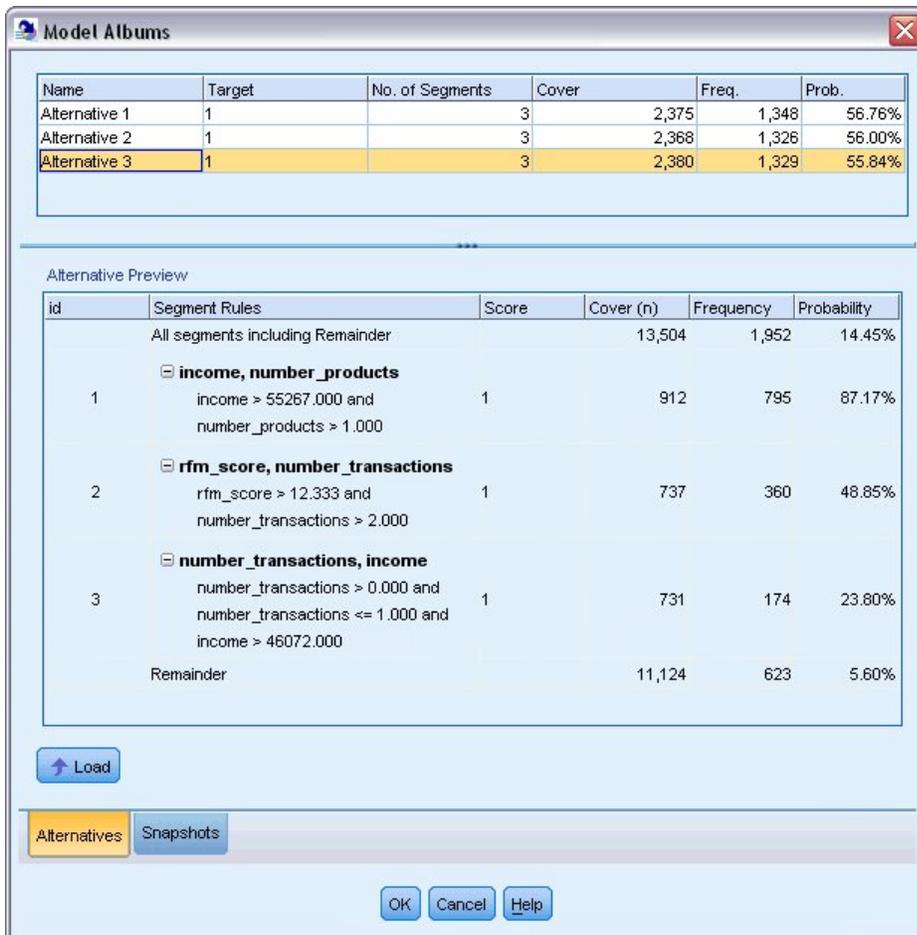


Abbildung 124. Verfügbare alternative Modelle

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

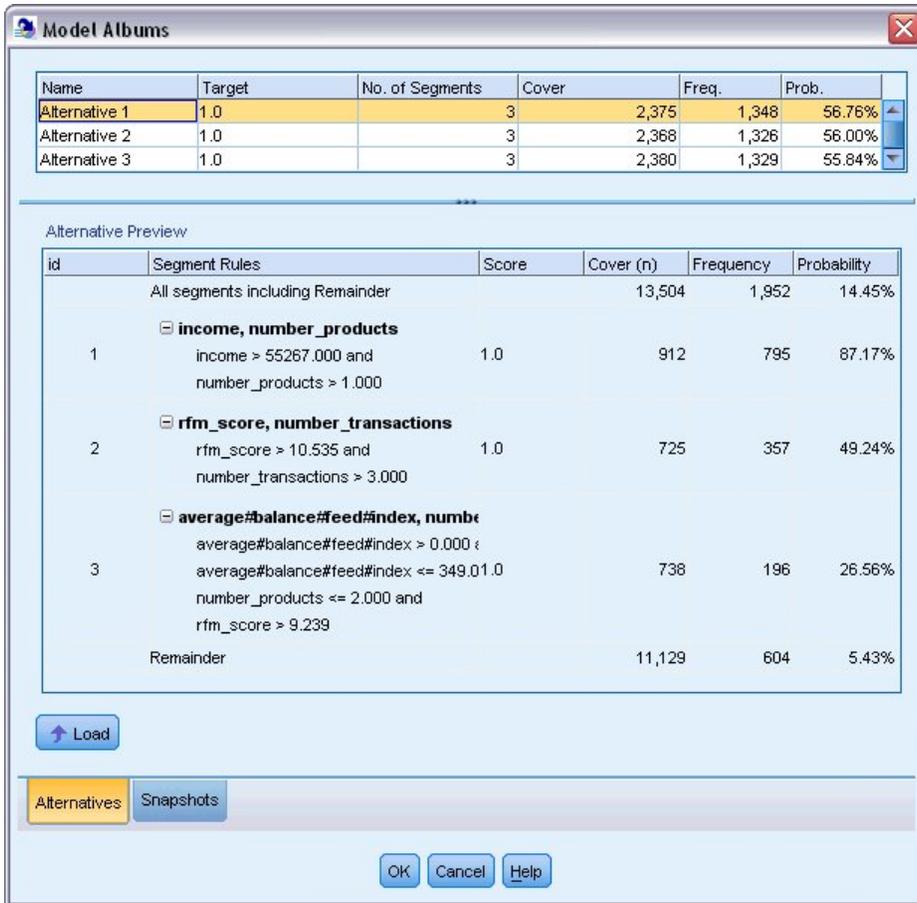


Abbildung 125. Ausgewähltes alternatives Modell

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.

Hinweis: Um das Modell besser sehen zu können, können Sie den Bereich "Alternative Vorschau" innerhalb des Dialogfelds, wie hier dargestellt, vergrößern. Dies können Sie 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:

Tools > Einstellungen

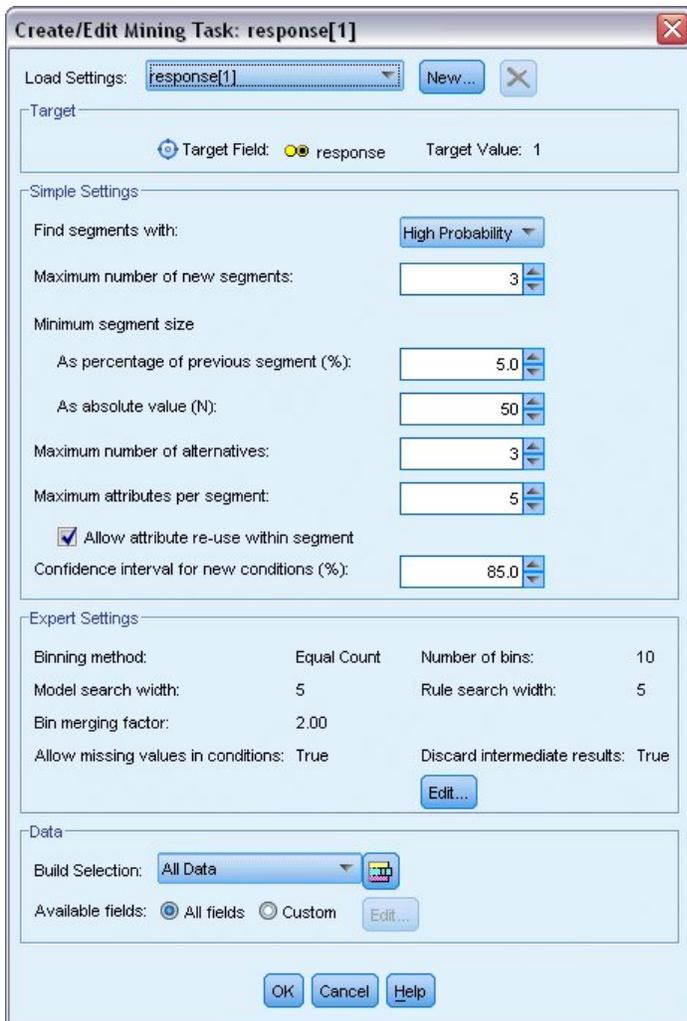


Abbildung 126. Dialogfeld "Mining-Aufgabe erstellen/bearbeiten"

11. 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.

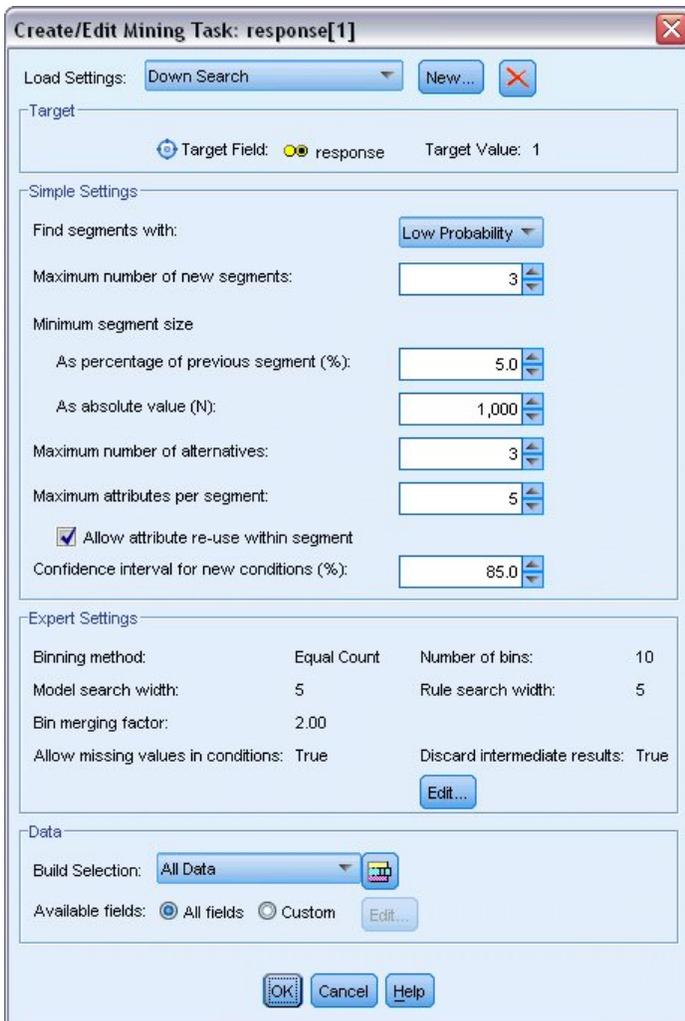


Abbildung 127. Dialogfeld "Mining-Aufgabe erstellen/bearbeiten"

12. 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.
13. Erhöhen Sie die minimale Segmentgröße auf 1.000. Klicken Sie auf **OK**, um in den Viewer "Interaktive Liste" zurückzukehren.
14. Stellen Sie im Viewer "Interaktive Liste" sicher, dass der Bereich *Segmentsuche* die Details der neuen Aufgabe anzeigt, und klicken Sie auf **Segmente suchen**.

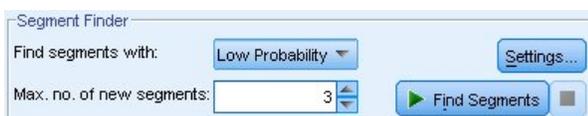


Abbildung 128. Segmente in einer neuen Mining-Aufgabe suchen

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

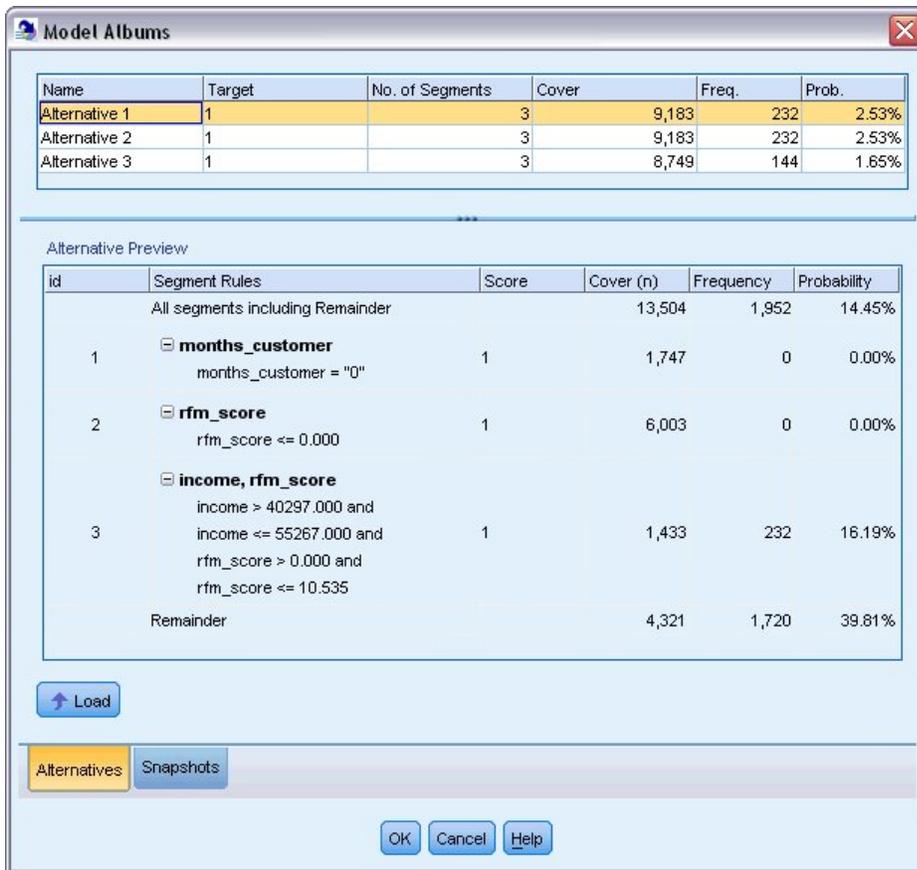


Abbildung 129. Modellergebnisse für Abwärtssuche

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 einer Suche des Typs "Hohe Wahrscheinlichkeit" - können Sie dieses Ergebnis eventuell verbessern.

15. 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.

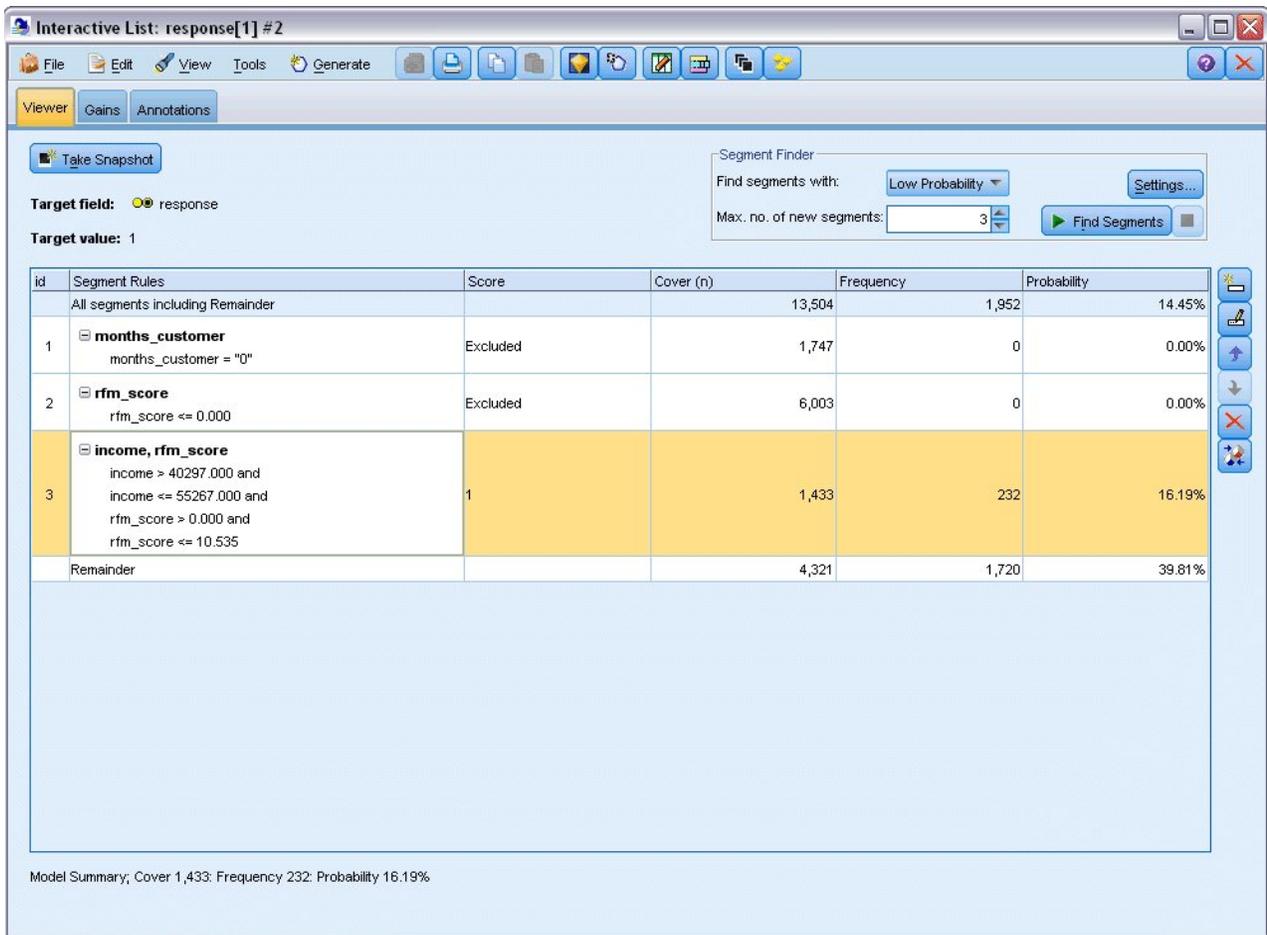


Abbildung 130. Ausschließen von Segmenten

16. Klicken Sie mit der rechten Maustaste auf jedes der ersten beiden Segmente und wählen Sie die Option **Segment ausschließen** aus. Zusammen erfassen diese Segmente fast 8.000 Datensätze, die insgesamt 0 Treffer aufweisen. Es ist also sinnvoll, sie aus zukünftigen Angeboten auszuschließen. (Ausgeschlossene Segmente erhalten den Score null, um dies anzuzeigen.)
17. 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 Basistrefferquote 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, während 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.
18. Klicken Sie in der Tabelle auf die Restzeile, um diese auszuwählen, sodass die nächste Mining-Aufgabe nur auf den Rest angewendet wird.

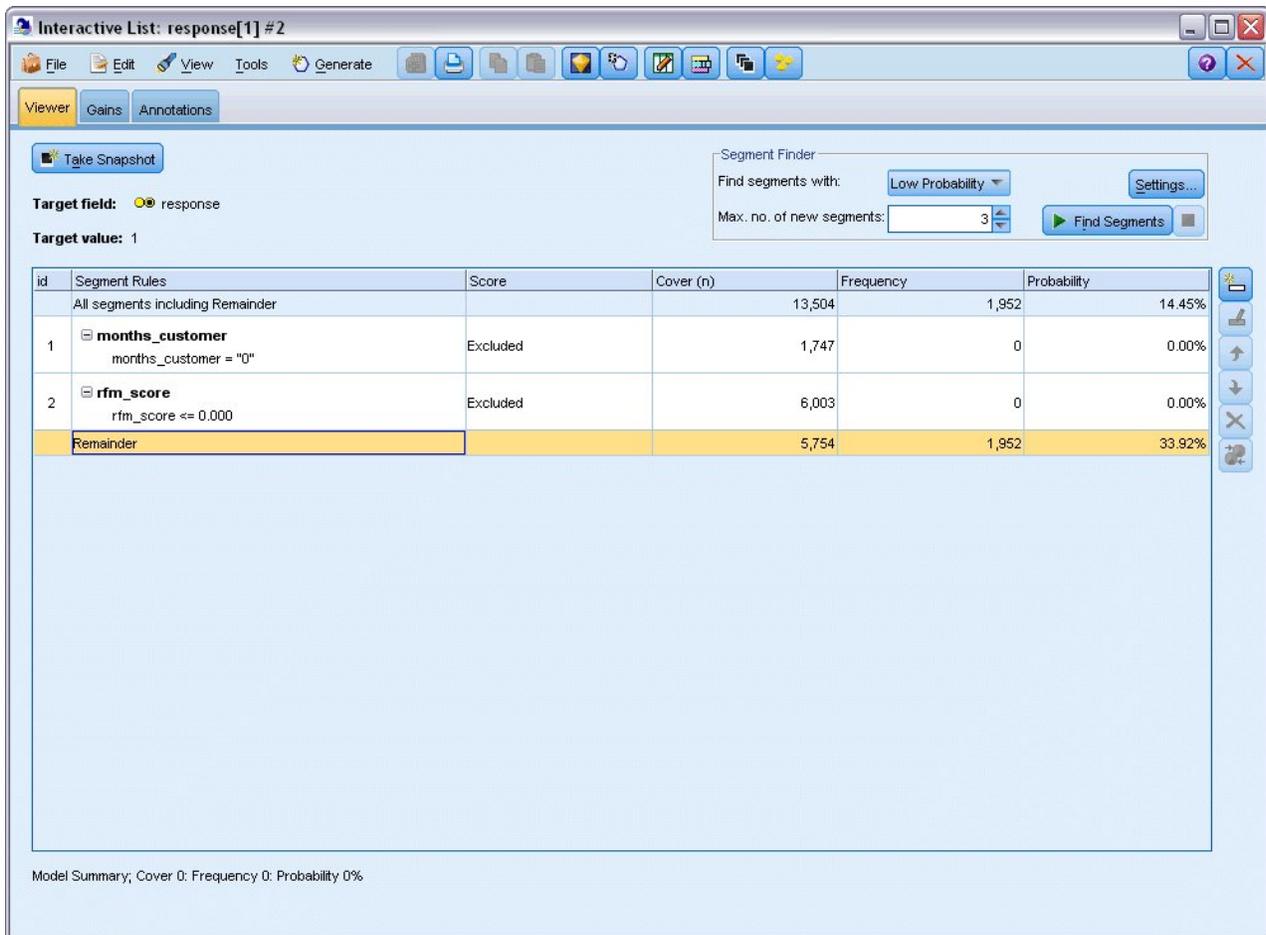


Abbildung 131. Auswählen von Segmenten

19. Klicken Sie bei ausgewähltem Rest auf **Einstellungen**, um das Dialogfeld "Mining-Aufgaben erstellen/bearbeiten" erneut zu öffnen.
20. Wählen Sie am oberen Rand unter **Einstellungen laden** die Standard-Mining-Aufgabe aus: **response[1]**.
21. Bearbeiten Sie die Daten unter **Einfache Einstellungen**, um die Anzahl der neuen Segmente auf 5 und die minimale Segmentgröße auf 500 zu erhöhen.
22. Klicken Sie auf **OK**, um in den Viewer "Interaktive Liste" zurückzukehren.

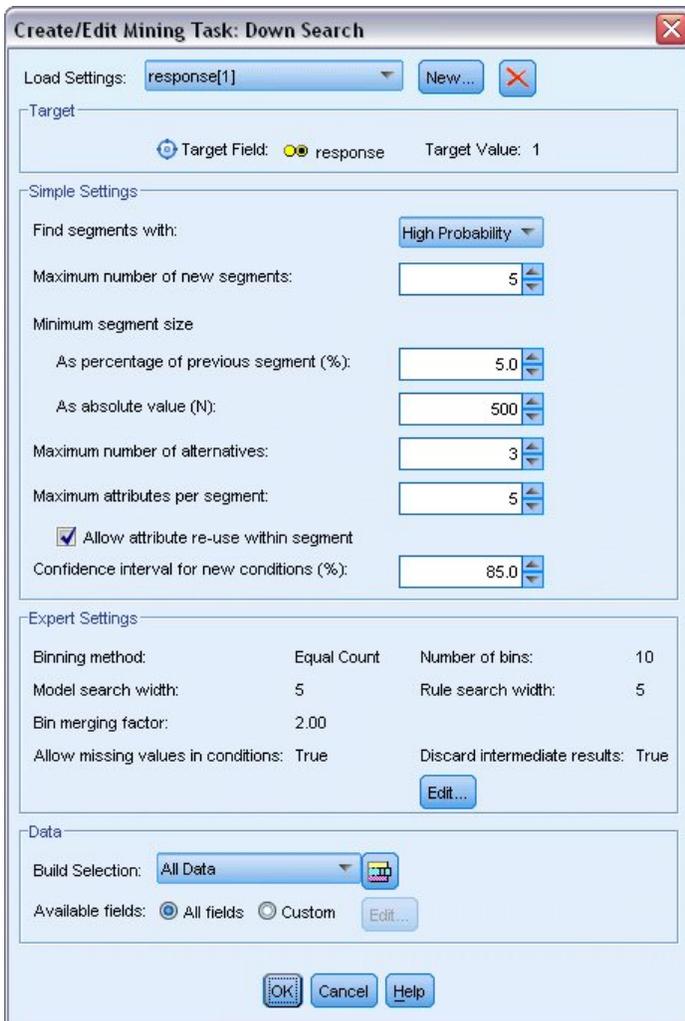


Abbildung 132. Auswählen der Standard-Mining-Aufgabe

23. 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, während 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.

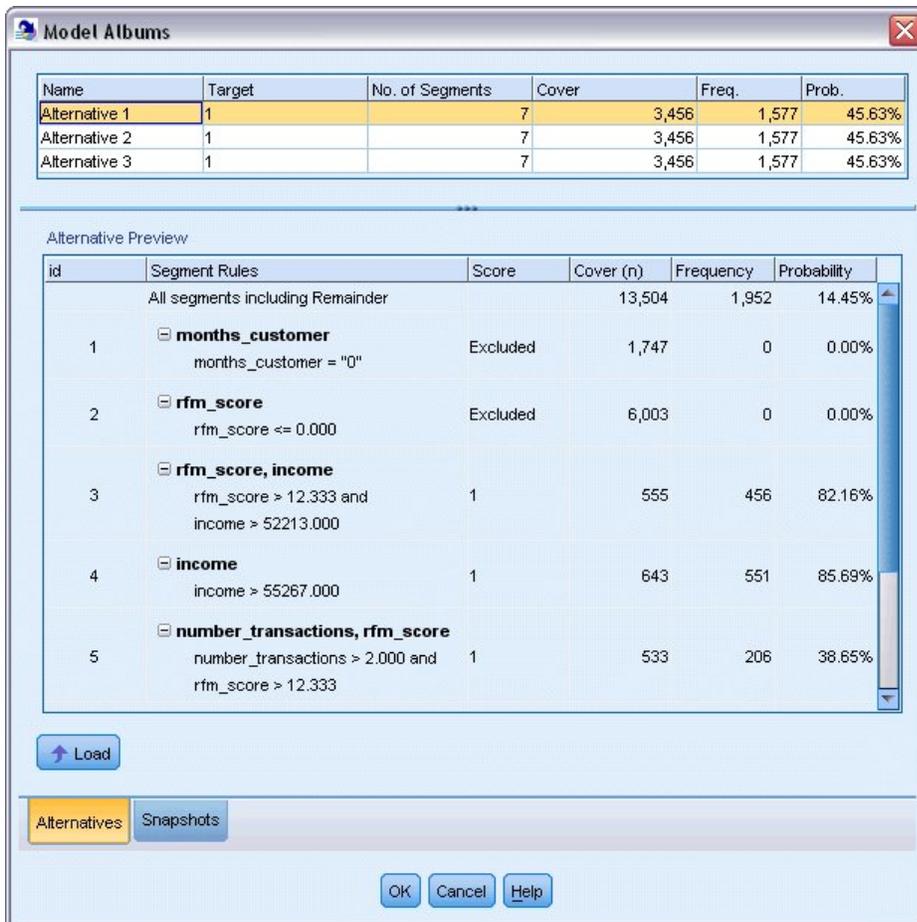


Abbildung 133. Alternativen für kombiniertes Modell

24. 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

1. Um weitere Einblicke in die Funktionsweise des Modells in der Praxis zu gewinnen, wählen Sie im Menü "Extras" die Option **Modellmaße organisieren** aus.

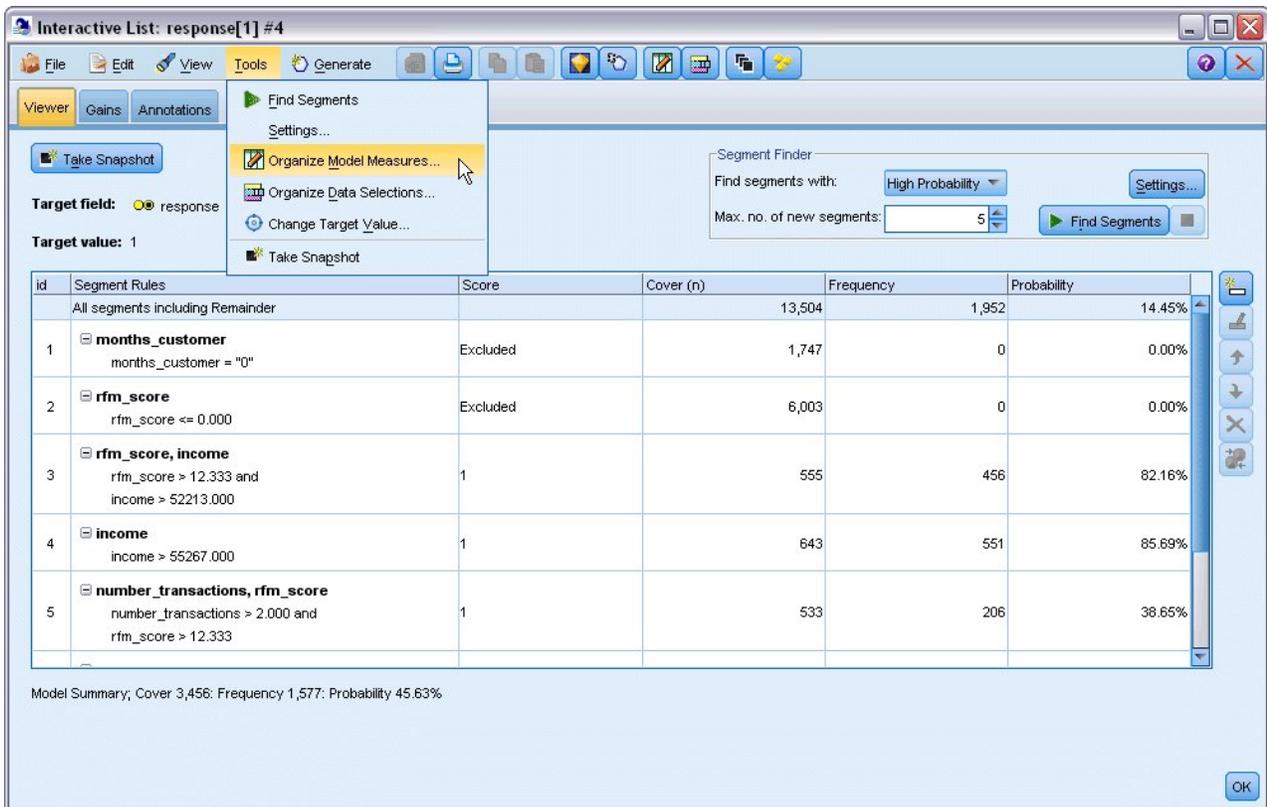


Abbildung 134. Organisieren von Modellmaßen

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

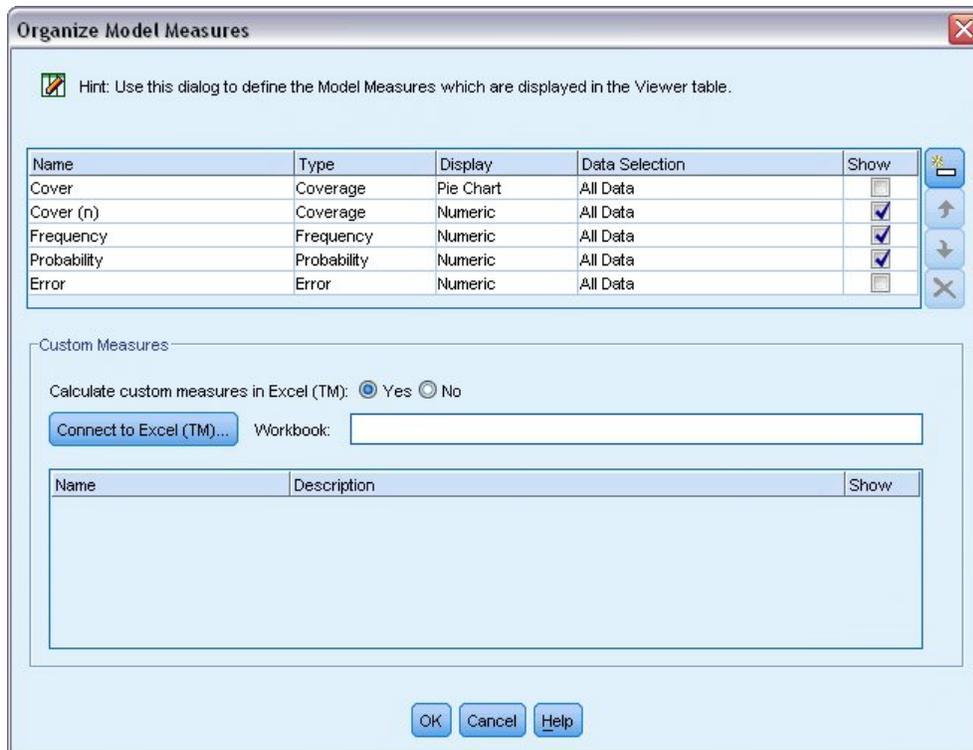


Abbildung 135. Dialogfeld "Modellmaße organisieren"

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.

2. Setzen Sie **Benutzerdefinierte Maße in Excel (TM) berechnen** im Dialogfeld "Modellmaße organisieren" auf **Ja**.
3. Klicken Sie auf die Schaltfläche **Verbindung mit Excel (TM) herstellen**.
4. 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.

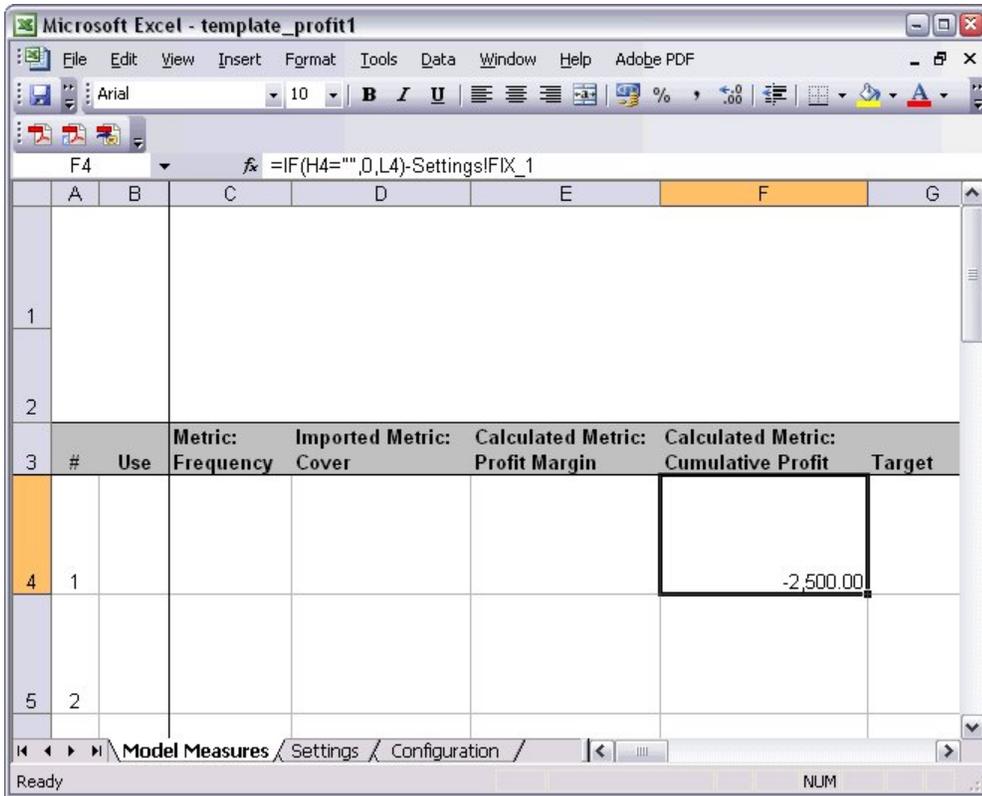


Abbildung 136. Excel-Arbeitsblatt "Modellmaße"

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 Nettoertrag aus dem Segment
- **Cumulative Profit (Kumulierter Profit)**. Der Gesamtprofit aus der Kampagne

Durch folgende Formeln definiert:

Profit Margin = Frequency * Revenue per respondent - Cover * Variable cost (Profitspanne = Häufigkeit * Ertrag pro Teilnehmer - Abdeckung * variable Kosten)

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

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

Die Parameter für Kosten und Ertrag werden vom Benutzer im Arbeitsblatt "Settings" (Einstellungen) angegeben.

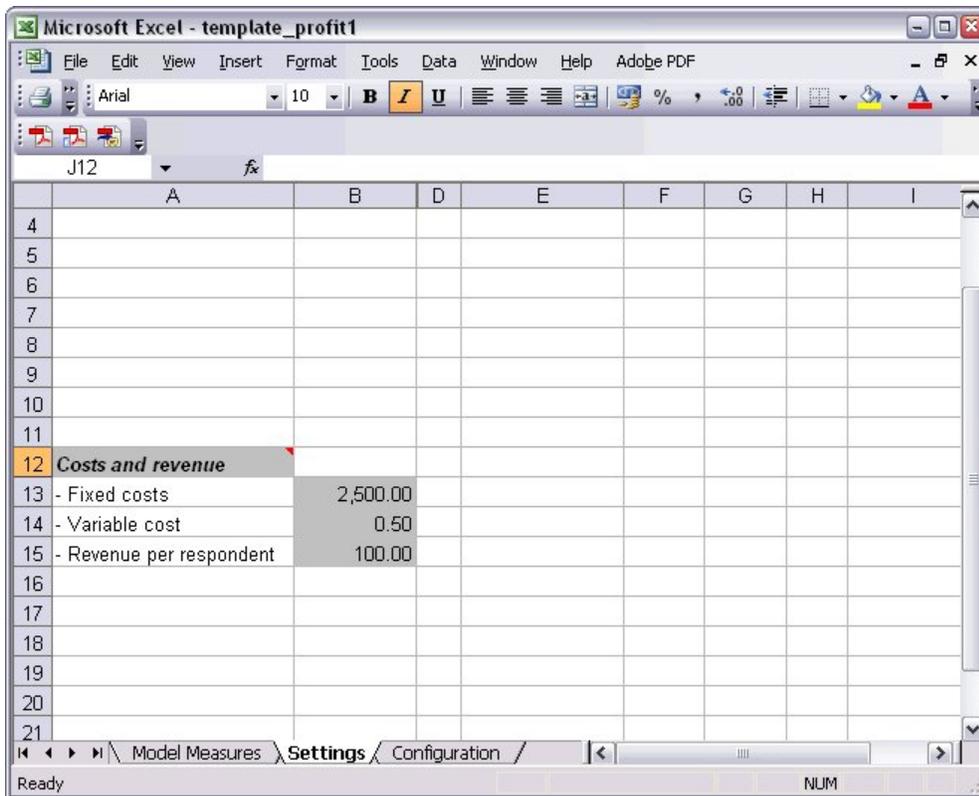


Abbildung 137. Excel-Arbeitsblatt "Einstellungen"

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 (Ertrag pro Teilnehmer) ist der Nettoertrag für einen Kunden, der auf das Angebot reagiert.

- Um die Verknüpfung zurück zum Modell abzuschließen, wechseln Sie mithilfe der Windows-Taskleiste (oder durch Drücken der Tastenkombination Alt+Tabulator) zurück zum Viewer "Interaktive Liste".

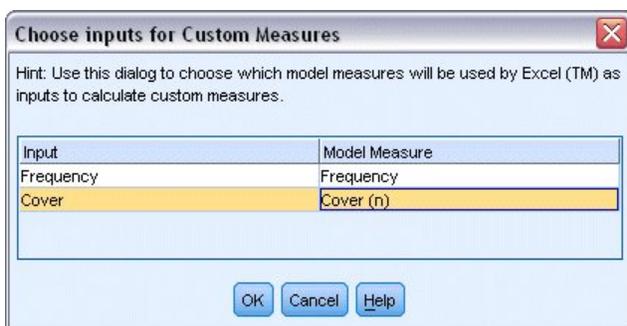


Abbildung 138. Auswählen von Eingaben für benutzerdefinierte Maße

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.

6. 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.
7. Klicken Sie im Dialogfeld "Modellmaße organisieren" auf **OK**, um die Anzeige des Viewers "Interaktive Liste" zu aktualisieren.

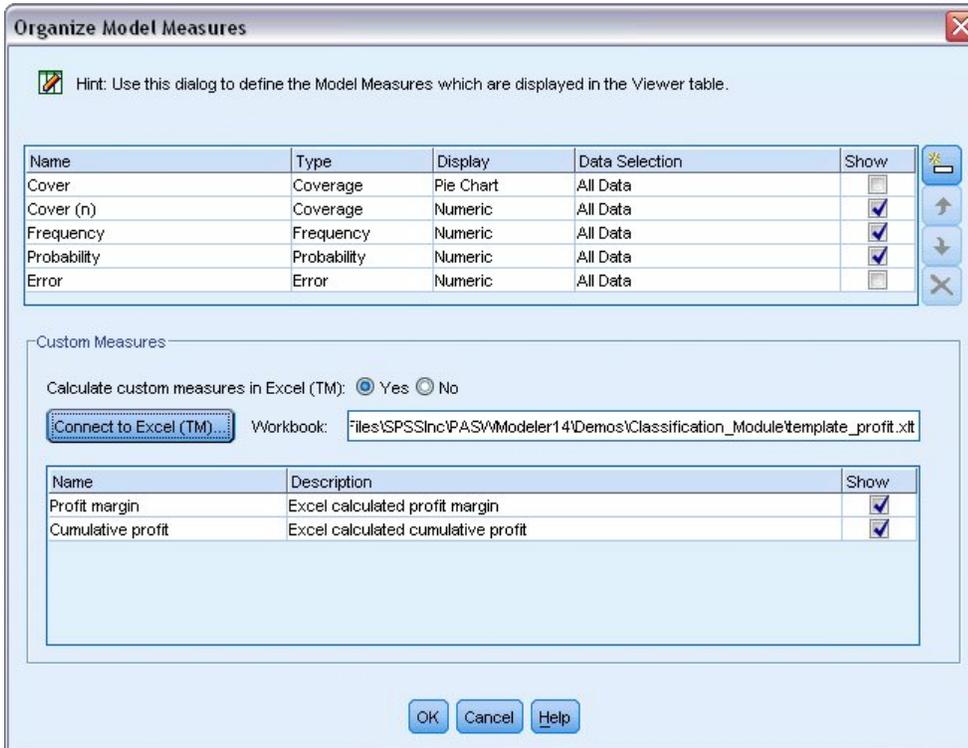


Abbildung 139. Dialogfeld "Modellmaße organisieren" mit benutzerdefinierten Maßen aus Excel

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

id	Segment Rules	Score	Cover (n)	Frequency	Probability	Profit margin	Cumulative...
	All segments including Remainder		13,504	1,952	14.45%	0	0
1	months_customer months_customer = "0"	Excluded	1,747	0	0.00%	-873.5	-2,500
2	rfm_score rfm_score <= 0.000	Excluded	6,003	0	0.00%	-3,001.5	-2,500
3	rfm_score, income rfm_score > 12.333 and income > 52213.000	1	555	456	82.16%	45,322.5	42,822.5
4	income income > 55267.000	1	643	551	85.69%	54,778.5	97,601
5	number_transactions, rfm_score number_transactions > 2.000 and rfm_score > 12.333	1	533	206	38.65%	20,333.5	117,934.5

Model Summary, Cover 3,456, Frequency 1,577, Probability 45.63%

Abbildung 140. 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 Ertrag und den Viewer "Interaktive Liste" mit den neuen Werten aktualisieren:

1. Wählen Sie im Viewer "Interaktive Liste" im Menü "Extras" die Option **Modellmaße organisieren** aus.
2. Klicken Sie im Dialogfeld "Modellmaße organisieren" auf **Verbindung mit Excel™ herstellen**.
3. Wählen Sie die Arbeitsmappe *template_profit.xlt* aus und klicken Sie auf **Öffnen**, um das Arbeitsblatt anzuzeigen.
4. Wählen Sie das Arbeitsblatt "Einstellungen" aus.
5. Ändern Sie den Wert für **Fixed costs** (Feste Kosten) in 3.250,00 und den Wert für **Revenue per respondent** (Ertrag pro Teilnehmer) in 150,00.

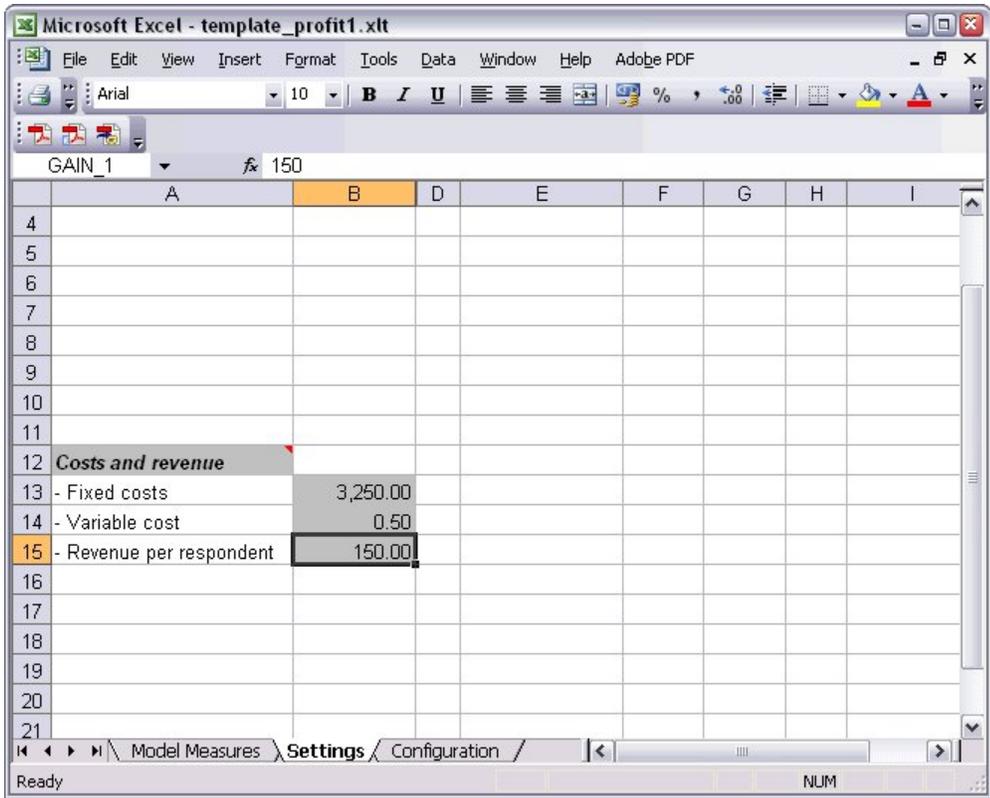


Abbildung 141. Geänderte Werte im Excel-Arbeitsblatt "Einstellungen"

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

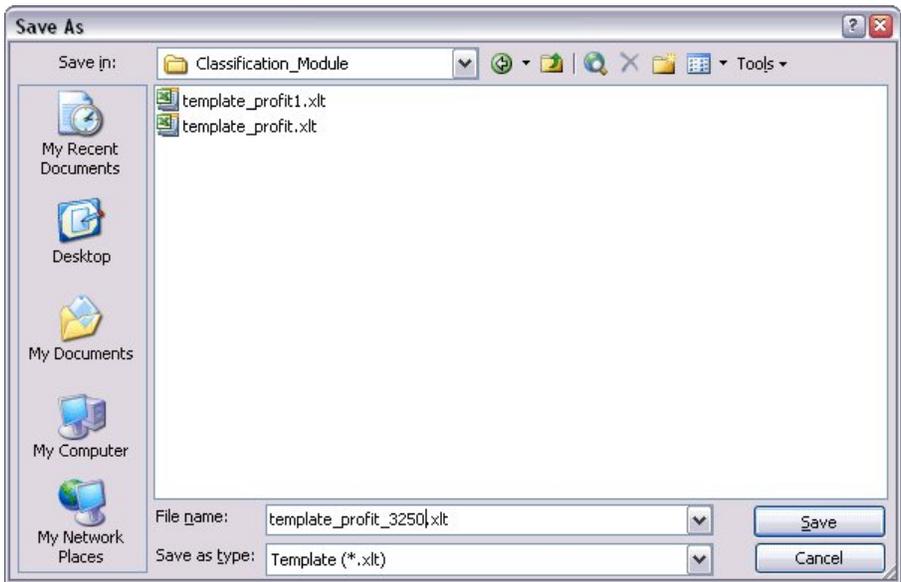


Abbildung 142. Speichern der geänderten Excel-Vorlage

7. Wechseln Sie mithilfe der Windows-Taskleiste (oder durch Drücken der Tastenkombination Alt+Tabulator) zurück zum Viewer "Interaktive Liste".

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

8. 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.

The screenshot shows the 'Interactive List' window with the following data table:

id	Segment Rules	Score	Cover (n)	Frequency	Probability	Profit margin	Cumulative ...
	All segments including Remainder		13,504	1,952	14.45%	0	0
1	months_customer months_customer = "0"	Excluded	1,747	0	0.00%	-873.5	-3,250
2	rfm_score rfm_score <= 0.000	Excluded	6,003	0	0.00%	-3,001.5	-3,250
3	rfm_score, income rfm_score > 12.333 and income > 52213.000	1	555	456	82.16%	68,122.5	64,872.5
4	income income > 55267.000	1	643	551	85.69%	82,328.5	147,201
5	number_transactions, rfm_score number_transactions > 2.000 and rfm_score > 12.333	1	533	206	38.65%	30,633.5	177,834.5

Model Summary; Cover 3,456; Frequency 1,577; Probability 45.63%

Abbildung 143. 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 eine Momentaufnahme des Modells erstellen. Dieser wird dann auf der Registerkarte "Momentaufnahmen" aufgeführt. Sie können jederzeit während der interaktiven Sitzung zu jeder beliebigen gespeicherten Momentaufnahme 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 gesort werden soll.

Wenn Sie mit den Ergebnissen zufrieden sind, können Sie mit dem Menü "Generieren" ein Modell erzeugen, das 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, Modellmomentaufnahmen, 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.

Kapitel 12. 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 entspricht der linearen Regression, außer dass statt eines numerischen Zielfelds ein kategoriales verwendet wird.

Nehmen wir beispielsweise an, dass ein Telekommunikationsanbieter seinen Kundenstamm nach Service-nutzungsmustern 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 auf die Datendatei *telco.sav* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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	Beschriftung
1	Basic Service (Basisservice)
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. Weitere Informationen finden Sie unter im Thema Kapitel 13, „Kundenabwanderung bei Telekommunikationsunternehmen (binomiale logistische Regression)“, auf Seite 139.

Erstellen des Streams

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

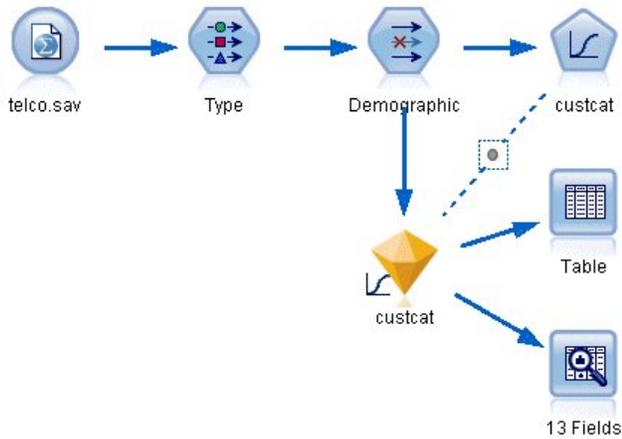


Abbildung 144. Beispielstream zur Klassifizierung von Kunden mithilfe der multinomialen logistischen Regression

- a. 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.

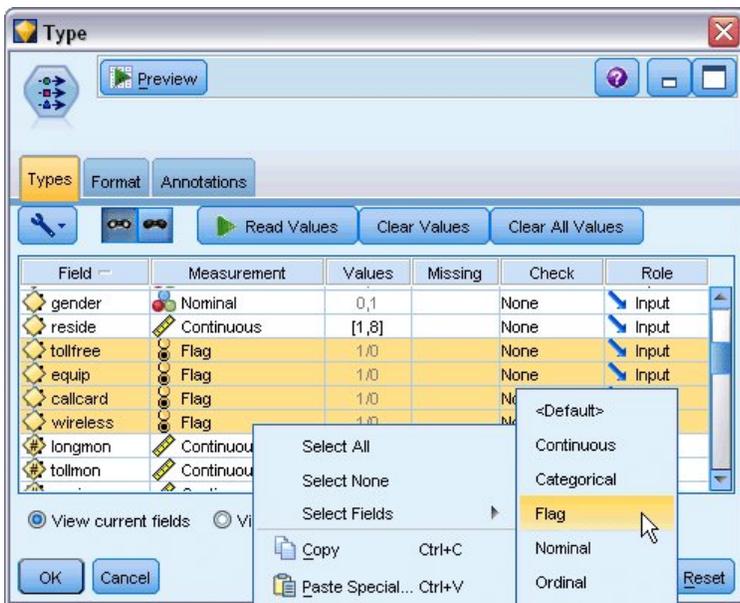


Abbildung 145. Festlegen des Messniveaus für mehrere Felder

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. Übernehmen Sie also **Nominal** für sein Messniveau.

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

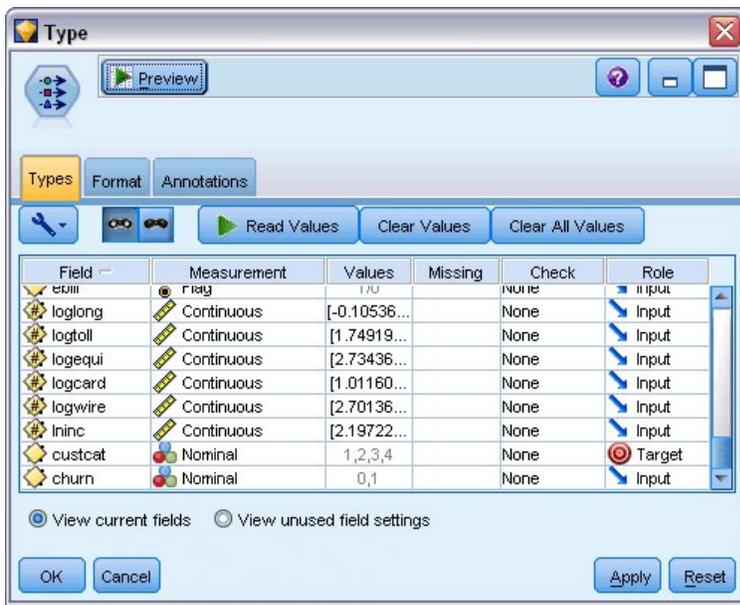


Abbildung 146. Festlegen der Feldrolle

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.

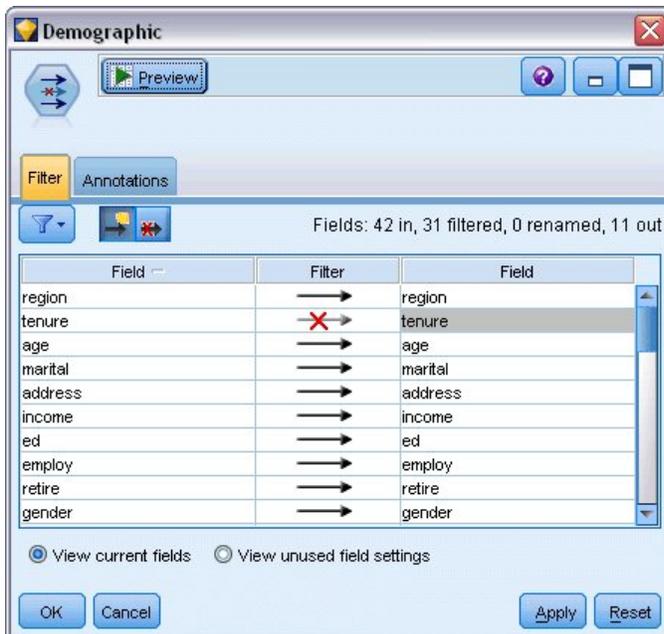


Abbildung 147. Filtern nach demografischen Feldern

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

2. 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.

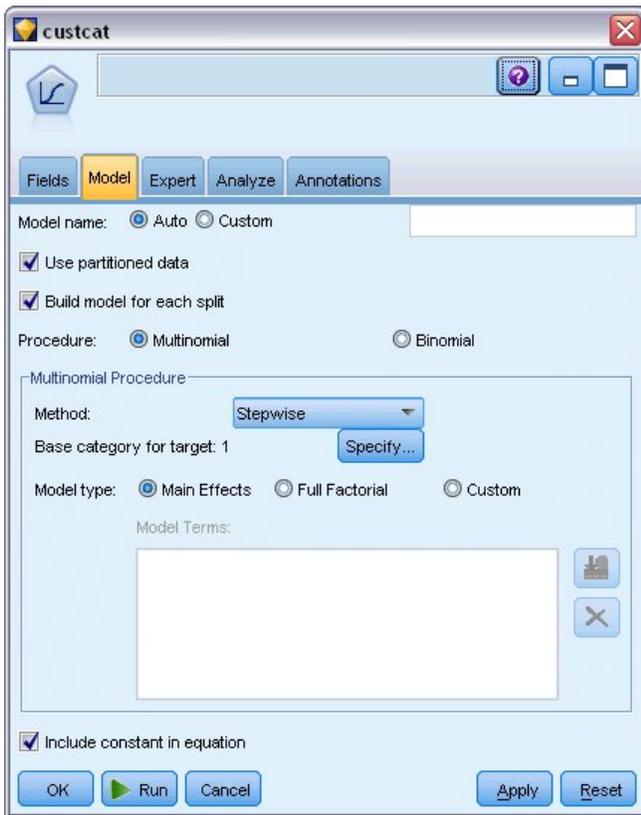


Abbildung 148. Auswählen der Modelloptionen

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

3. Wählen Sie auf der Registerkarte "Experten" den Modus **Experten** aus und wählen Sie die Option **Ausgabe** aus; wählen Sie dann auf der Registerkarte "Erweiterte Ausgabe" die Option **Klassifikationstabelle** aus.

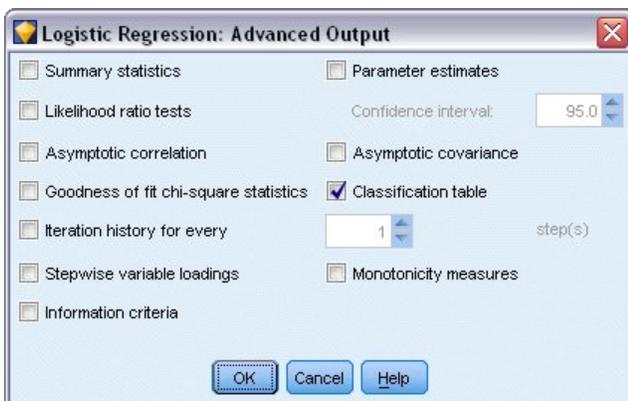


Abbildung 149. Auswahl der Ausgabeoptionen

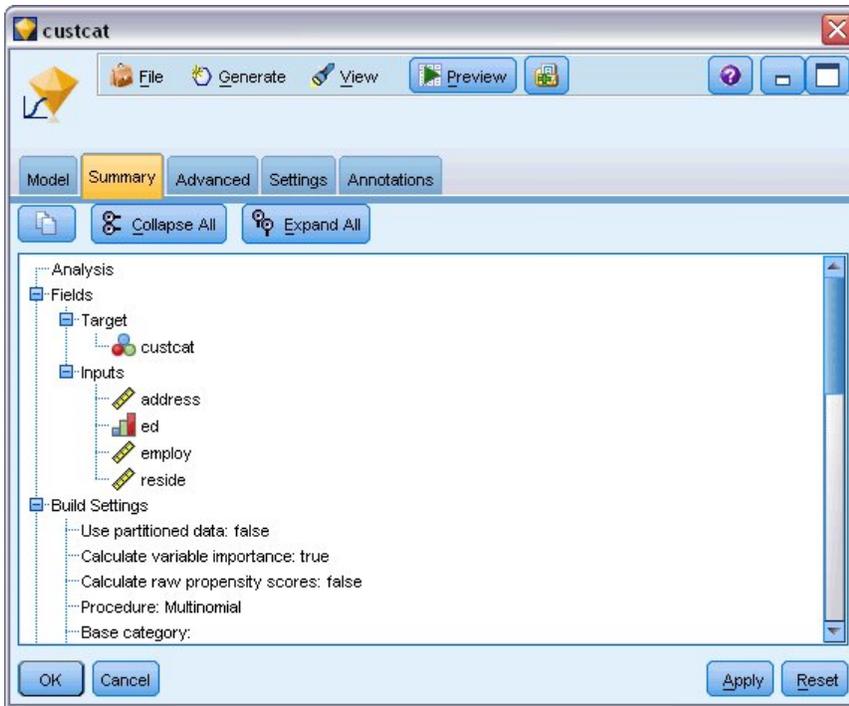


Abbildung 151. Modellübersicht mit Ziel- und Eingabefeldern

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.

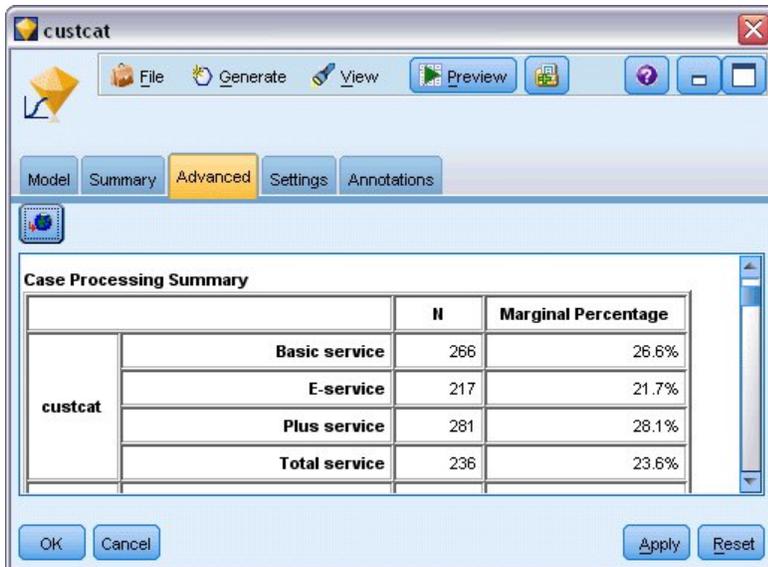


Abbildung 152. Zusammenfassung der Fallverarbeitung

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 ungünstig 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.

Observed	Predicted				Percent Correct
	Basic service	E-service	Plus service	Total service	
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%

Abbildung 153. Klassifikationsmatrix

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 der höchste Return-on-Investment (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 Diagnosetools zur Verfügung. Erläuterungen der mathematischen Grundlagen für die in IBM SPSS Modeler verwendeten Modellierungsmethoden finden Sie im *IBM 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 ein Subset der Datensätze für Test- und Validierungszwecke zurückhalten.

Kapitel 13. 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 entspricht der 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 auf die Datendatei *telco.sav* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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. Weitere Informationen finden Sie unter im Thema Kapitel 12, „Klassifizieren von Kunden im Telekommunikationsbereich (multinomiale logistische Regression)“, auf Seite 131.

Erstellen des Streams

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

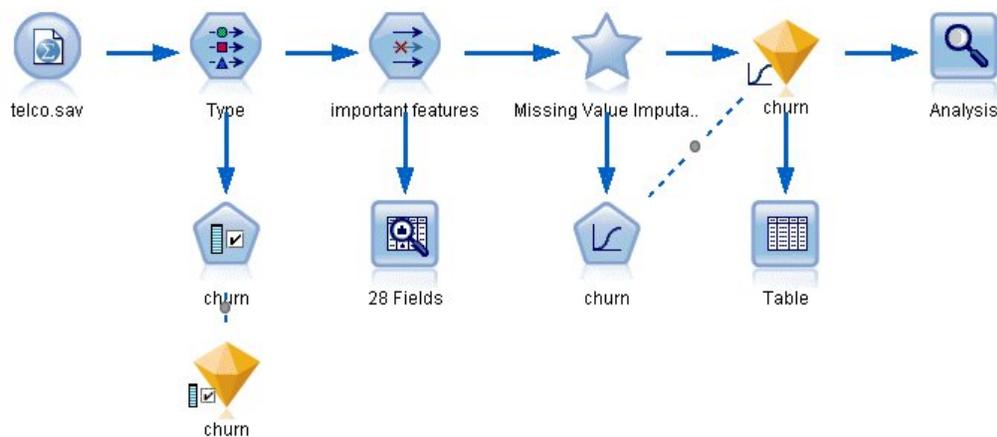


Abbildung 154. Beispielstream zur Klassifizierung von Kunden mithilfe der binomialen logistischen Regression

2. 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.

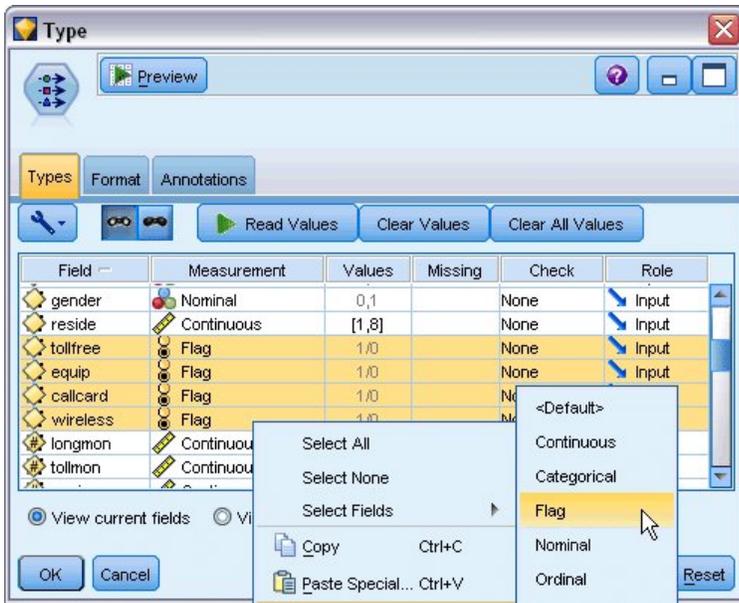


Abbildung 155. Festlegen des Messniveaus für mehrere Felder

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.

3. 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.

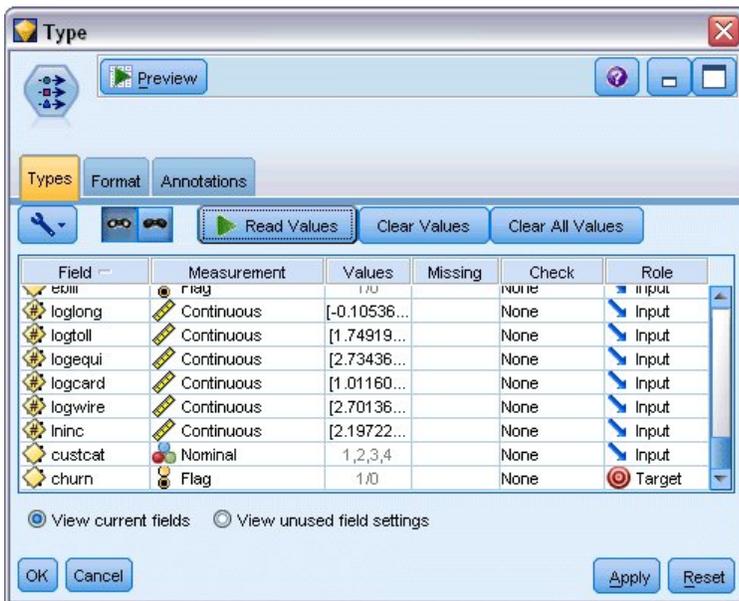


Abbildung 156. Festlegen von Messniveau und Rolle für das Feld "churn"

4. Fügen Sie dem Typknoten einen Merkmalauswahlmodellierungsknoten hinzu.
Mit einem Merkmalauswahlknoten 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 Ergebnismodellnugget und wählen Sie aus dem Menü **Generieren** die Option **Filter**, um einen Filterknoten zu erstellen.

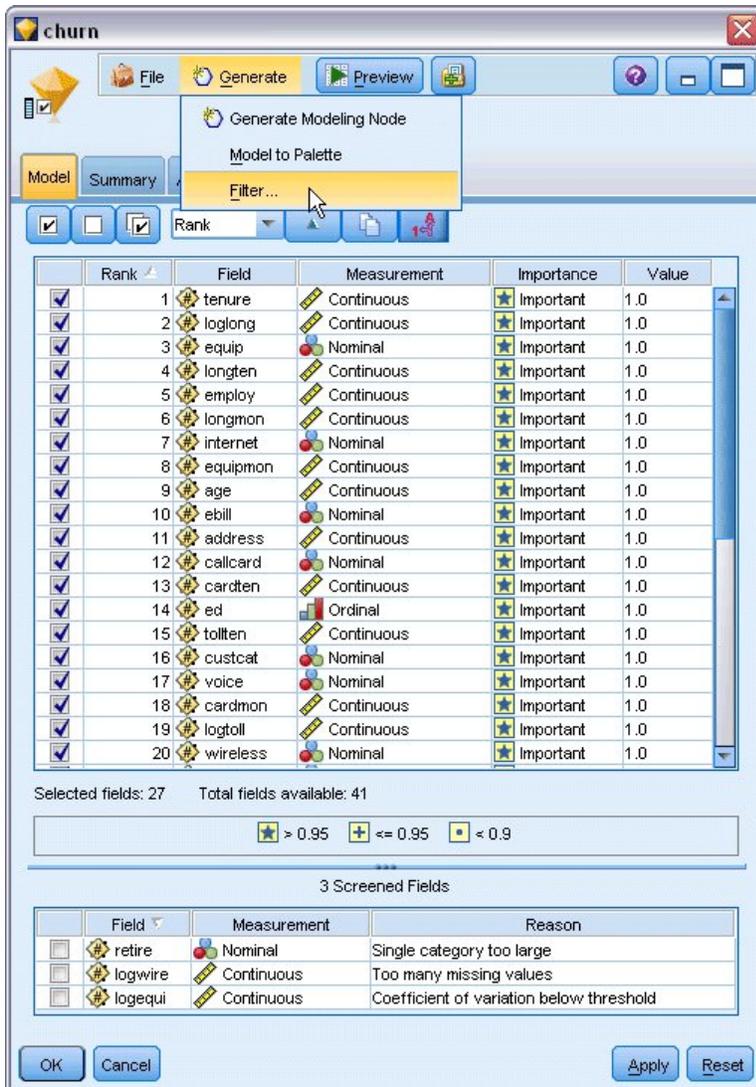


Abbildung 157. Generieren eines Filterknotens aus einem Merkmalauswahlknoten

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** aus und klicken Sie auf **OK**.
- Verbinden Sie den generierten Filterknoten mit dem Typknoten.

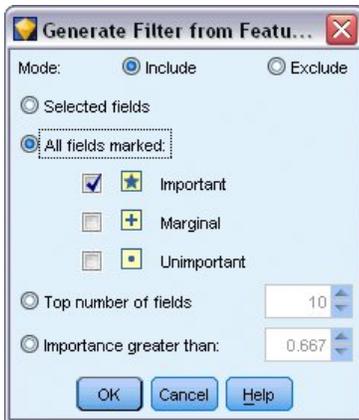


Abbildung 158. Auswählen bedeutsamer Felder

9. Fügen Sie dem generierten Filterknoten einen Data Audit-Knoten hinzu.
Öffnen Sie den Data Audit-Knoten und klicken Sie auf **Ausführen**.
10. 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.
11. Klicken Sie in der Spalte *Fehlende Werte imputieren* für *logtoll* auf **Angeben**.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid
logtoll	Continuous	2	0	None	Never	Fixed	47.5	
tenure	Continuous	0	0	None	Never	Fixed	100	
age	Continuous	0	0	None	Blank Values	Fixed	100	
address	Continuous	12	0	None	Null Values	Fixed	100	
income	Continuous	9	6	None	Blank & Null Value	Fixed	100	
ed	Ordinal	--	--	--	Condition...	Fixed	100	
employ	Continuous	8	0	None	Specify...	Fixed	100	
equip	Flag	--	--	--	never	Fixed	100	
callcard	Flag	--	--	--	Never	Fixed	100	
wireless	Flag	--	--	--	Never	Fixed	100	
longmon	Continuous	18	4	None	Never	Fixed	100	
tollmon	Continuous	9	1	None	Never	Fixed	100	
equipmon	Continuous	2	0	None	Never	Fixed	100	
cardmon	Continuous	11	3	None	Never	Fixed	100	
wiremon	Continuous	8	1	None	Never	Fixed	100	
longten	Continuous	20	4	None	Never	Fixed	100	
tollten	Continuous	18	2	None	Never	Fixed	100	
cardten	Continuous	11	6	None	Never	Fixed	100	
voice	Flag	--	--	--	Never	Fixed	100	

Abbildung 159. Imputieren fehlender Werte für "logtoll"

12. Wählen Sie für **Imputieren**, wenn die Option **Leere Werte und Nullwerte** aus. Wählen Sie für **Festgelegt als** die Option **Mittelwert** aus 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.

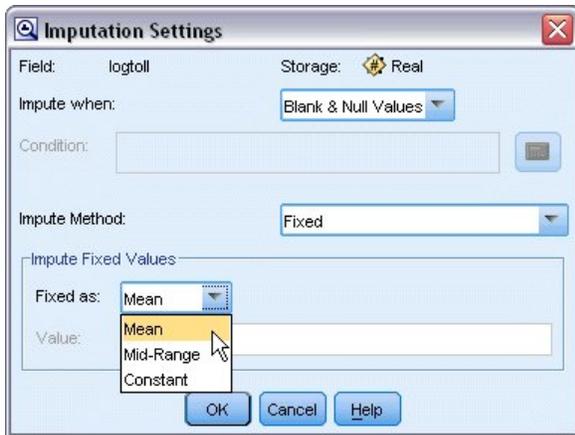


Abbildung 160. Auswahl der Imputationseinstellungen

13. 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:

Generieren > Superknoten für fehlende Werte

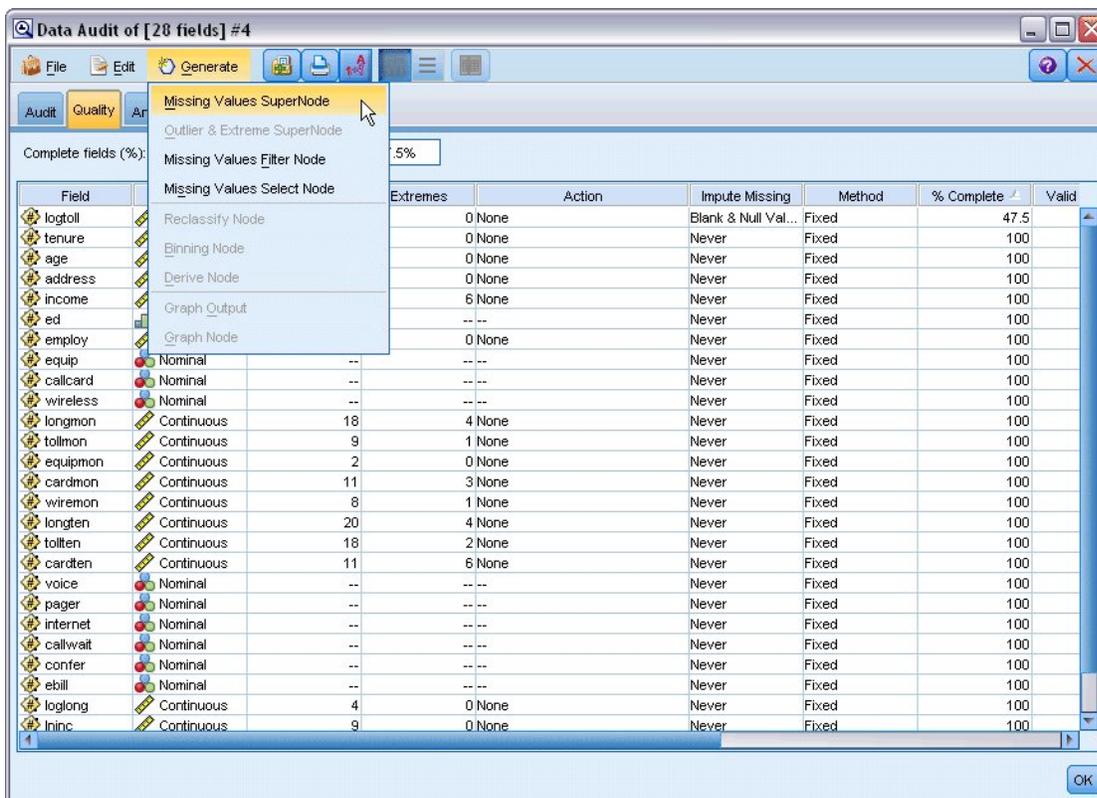


Abbildung 161. Generieren eines Superknotens für fehlende Werte

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

Der Superknoten wird im Streamerstellungsbereich angezeigt und trägt den Titel: *Imputation fehlender Werte*.

14. Verbinden Sie den Superknoten mit dem Filterknoten.

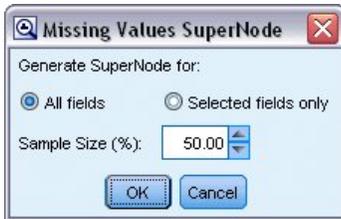


Abbildung 162. Angeben des Stichprobenumfangs

15. Fügen Sie dem Superknoten einen Logistikknoten hinzu.
16. 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.

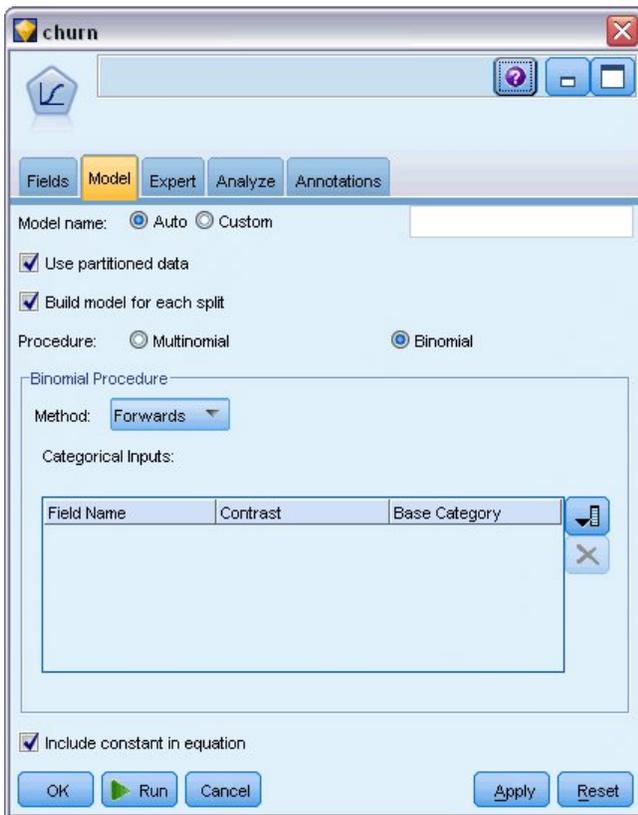


Abbildung 163. Auswählen der Modelloptionen

17. Wählen Sie auf der Registerkarte "Experten" den Modus **Experten** aus und klicken Sie dann auf **Ausgabe**. Das Dialogfeld "Erweiterte Ausgabe" wird angezeigt.
18. Wählen Sie im Dialogfeld "Erweiterte Ausgabe" die Option **Bei jedem Schritt** als Typ für *Anzeige* aus. Wählen Sie **Iterationsverlauf** und **Parameterschätzungen** aus und klicken Sie auf **OK**.

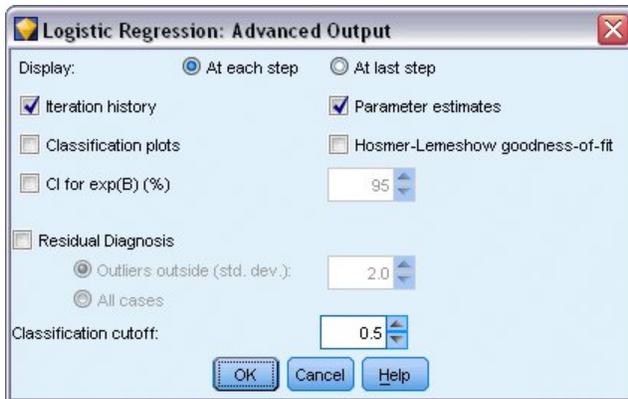


Abbildung 164. Auswahl der Ausgabeoptionen

Durchsuchen des Modells

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

Das generierte Modellnugget wird dem Streamerstellungsbereich und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die zugehörigen Details anzuzeigen, klicken Sie mit der rechten Maustaste auf das Modellnugget und wählen **Bearbeiten** oder **Durchsuchen** aus.

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.

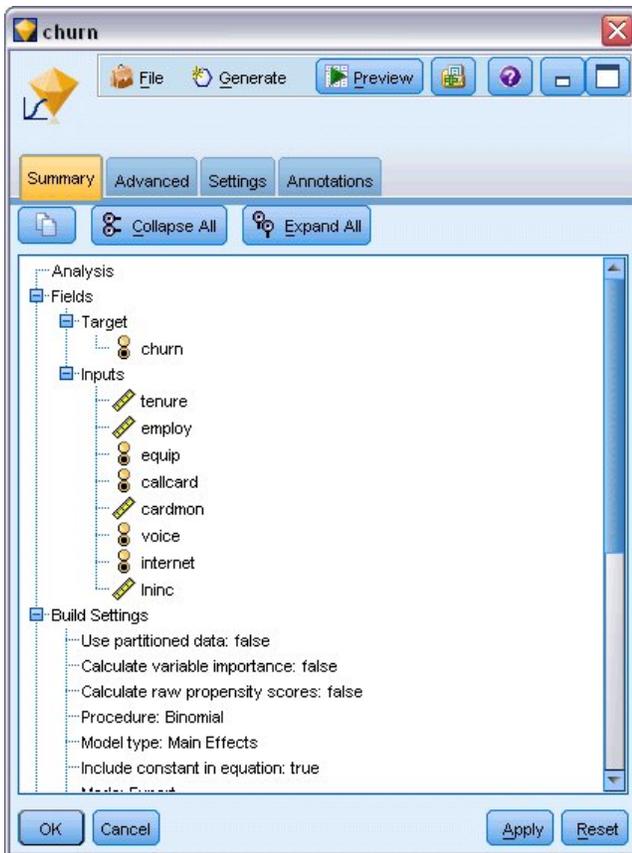


Abbildung 165. Modellübersicht mit Ziel- und Eingabefeldern

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 Zusammenfassung der Fallverarbeitung, die die Anzahl und den Prozentsatz der in der Analyse einbezogenen Datensätze anzeigt. Außerdem wird gegebenenfalls die Anzahl der fehlenden Fälle aufgeführt, bei denen ein oder mehrere Eingabefelder nicht verfügbar sind, und alle Fälle, die nicht ausgewählt wurden.

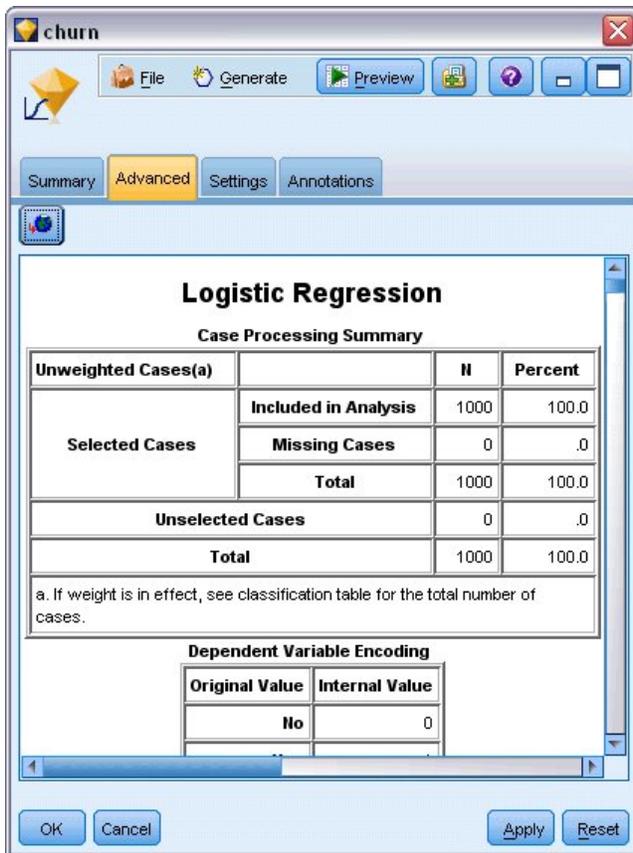


Abbildung 166. Zusammenfassung der Fallverarbeitung

2. Führen Sie in der Zusammenfassung der Fallverarbeitung einen Bildlauf nach unten durch, um die Klassifikationstabelle unter Block 0, dem 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 ganz und gar nicht richtig vorhergesagt.

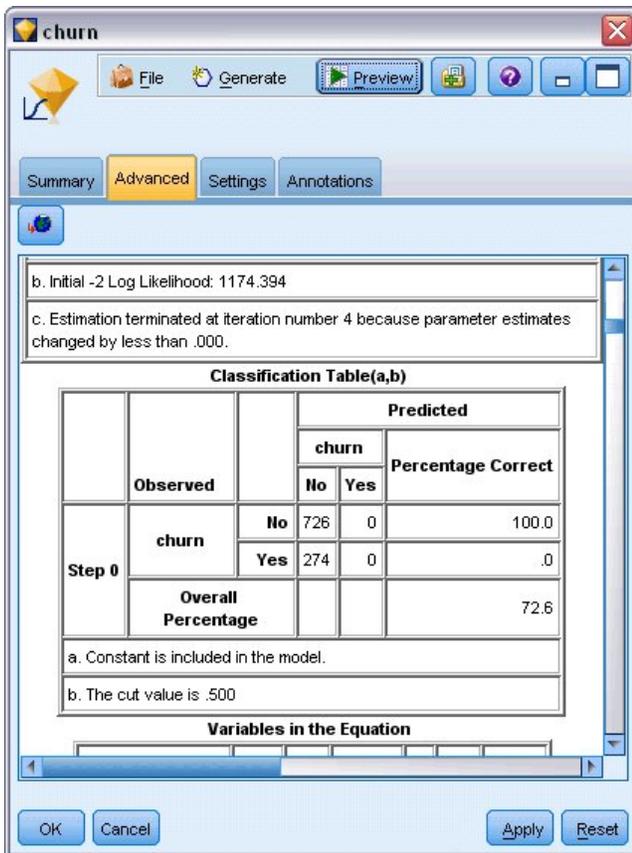


Abbildung 167. Anfangsklassifikationstabelle - Block 0

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

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.

The screenshot shows a software window titled 'churn' with a menu bar (File, Generate, Preview) and tabs (Summary, Advanced, Settings, Annotations). The main area displays a 'Classification Table(a)' with the following data:

	Observed	Predicted			Percentage Correct
		churn	churn		
			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

Abbildung 168. Klassifikationstabelle - Block 1

4. Führen Sie einen Bildlauf zum Ende dieser Klassifikationstabelle durch.

Die Klassifikationstabelle zeigt, dass der letzte Schritt der 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.

The screenshot shows the 'churn' dialog box in IBM SPSS Modeler. The 'Advanced' tab is selected. The main area contains classification tables for Step 7 and Step 8, and a 'Variables in the Equation' table for Step 1(a).

Step 7 Classification Table:

Overall Percentage				78.7
Step 7	churn	No	657 69	90.5
		Yes	144 130	47.4
Overall Percentage				78.7

Step 8 Classification Table:

Step 8	churn	No	662 64	91.2
		Yes	145 129	47.1
Overall Percentage				79.1

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	tenure	-.046	.004	123.346	1	.000	.955
	Constant	.462	.136	11.574	1	.001	1.587

Abbildung 169. 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 Diagnosetools zur Verfügung. Erläuterungen der mathematischen Grundlagen für die in IBM SPSS Modeler verwendeten Modellierungsmethoden finden Sie im *IBM 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 ein Subset der Datensätze für Test- und Validierungszwecke zurückhalten.

Kapitel 14. Vorhersage der Bandbreitennutzung (Zeitreihen)

Vorhersageerstellung mit dem Zeitreihenknoten

Ein Analyst eines Breitbandproviders soll Vorhersagen über die Vertragsabschlüsse mit Kunden erstellen, um die Auslastung der Bandbreite prognostizieren 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 auf die Datendatei *broadband_1.sav* nimmt. Die Dateien stehen im Ordner *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. Die Datei *broadband_create_models.str* befindet sich im Ordner *streams*.

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

In IBM 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 Zeichenfolge in IBM SPSS Modeler eingelesen. Um das Feld in IBM SPSS Modeler verwenden zu können, müssen Sie jedoch den Speichertyp mithilfe eines Füllerknotens in das numerische Datumsformat konvertieren.

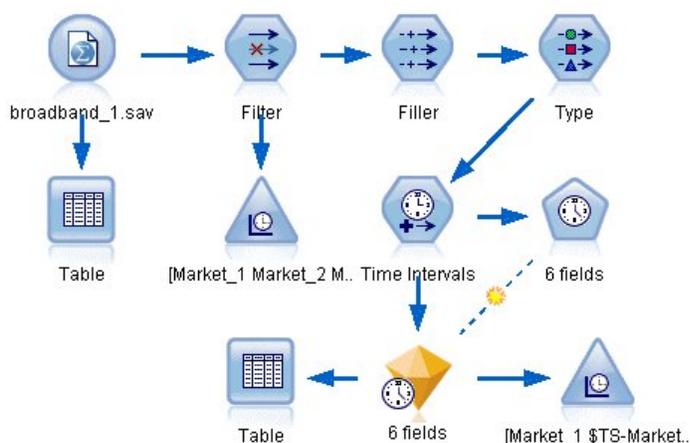


Abbildung 170. Beispielstream zur Anzeige der Zeitreihenmodellierung

Für den Zeitreihenknoten ist es erforderlich, dass sich jede Zeitreihe in einer separaten Spalte befindet und dabei jeweils eine Zeile für jedes Intervall vorliegt. IBM SPSS Modeler bietet Methoden, mit denen

die Daten, falls erforderlich, in dieses Format umgewandelt werden können.

	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	5041
2	3846	11984	12228	4825	2301	5672	6390	2404	5160
3	3894	12266	12897	5041	2352	5802	6670	2469	5230
4	4010	12801	13716	5211	2490	5899	6929	2574	5400
5	4147	13291	14647	5383	2534	6017	7312	2654	5540
6	4335	13828	15419	5496	2664	6137	7493	2699	5770
7	4554	14273	16108	5747	2738	6250	7702	2786	5900
8	4744	14664	16958	5885	2754	6439	7965	2847	6030
9	4885	15130	17642	6053	2874	6701	8107	2967	6150
10	5020	15851	18453	6229	2975	6957	8366	3099	6340
11	5208	16509	19181	6320	3042	7111	8684	3195	6630
12	5379	17225	19885	6499	3095	7275	8997	3341	6760
13	5574	18173	20565	6593	3199	7380	9326	3376	7020
14	5828	19287	21155	6680	3207	7633	9543	3443	7330
15	5942	20171	21655	6757	3298	7985	9673	3617	7490
16	6139	21379	21964	6804	3387	8236	9934	3732	7710
17	6244	22067	22756	6915	3450	8464	10211	3831	7940
18	6274	23074	23464	7035	3528	8575	10440	3886	8290
19	6347	23729	24324	7151	3546	8817	10763	3938	8580
20	6399	24803	25351	7304	3604	9041	11012	3953	8710

Abbildung 171. Monatliche Abonnementdaten für lokale Breitbandmärkte

Erstellen des Streams

1. Erstellen Sie einen neuen Stream und fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *broadband_1.sav* verweist.
2. 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 Steuertaste gedrückt und klicken Sie auf die Felder *MONTH_* und *YEAR_*.

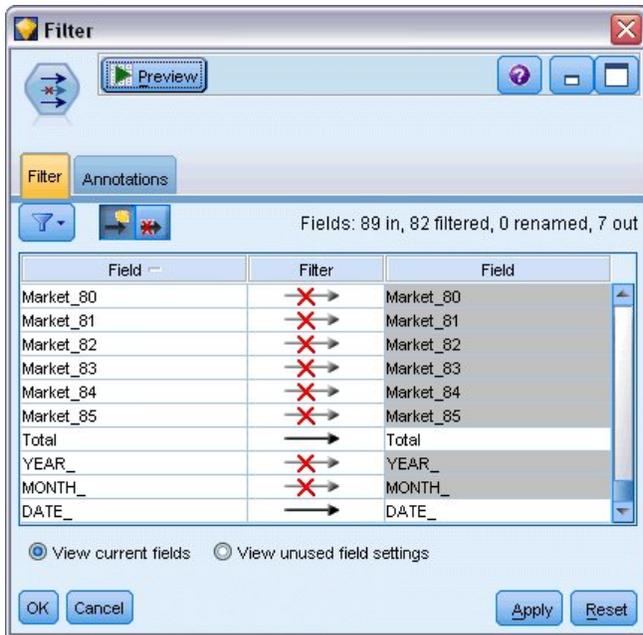


Abbildung 172. Vereinfachen des Modells

Untersuchen der Daten

Es empfiehlt sich grundsätzlich, sich einen Eindruck über die Beschaffenheit der Daten zu verschaffen, bevor Sie das Modell erstellen. Weisen die Daten saisonale Schwankungen auf? Der Expert Modeler kann zwar automatisch das beste saisonale oder nicht saisonale Modell für jede Zeitreihe ermitteln, Sie können jedoch häufig schnellere Ergebnisse erzielen, indem Sie die Suche auf nicht saisonale 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.

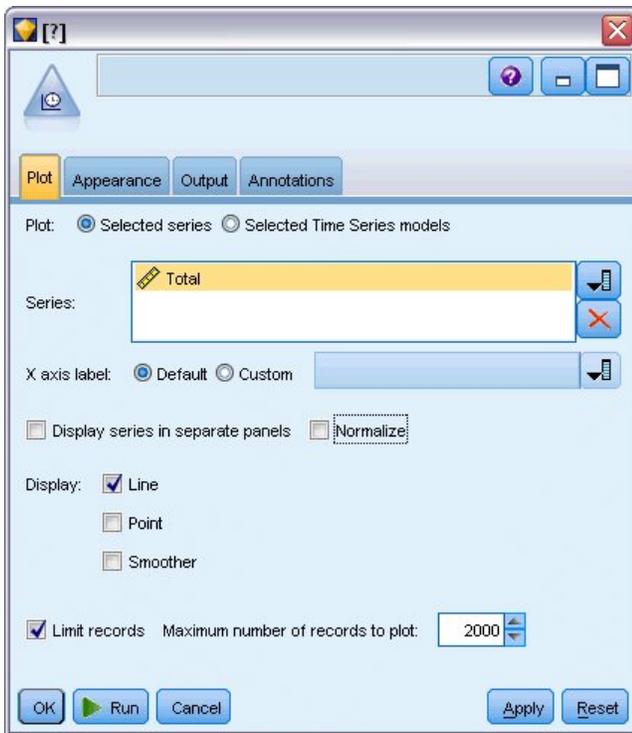


Abbildung 173. Plotten der Gesamtzahl an Abonnenten

1. Fügen Sie auf der Diagrammpalette einen Zeitdiagrammknoten an den Filterknoten an.
2. Fügen Sie der Liste "Series" (Reihe) das Feld *Total* (Gesamt) hinzu.
3. Inaktivieren Sie die Kontrollkästchen **Reihen in gesonderten Fenstern anzeigen** und **Normalisieren**.
4. Klicken Sie auf **Ausführen**.

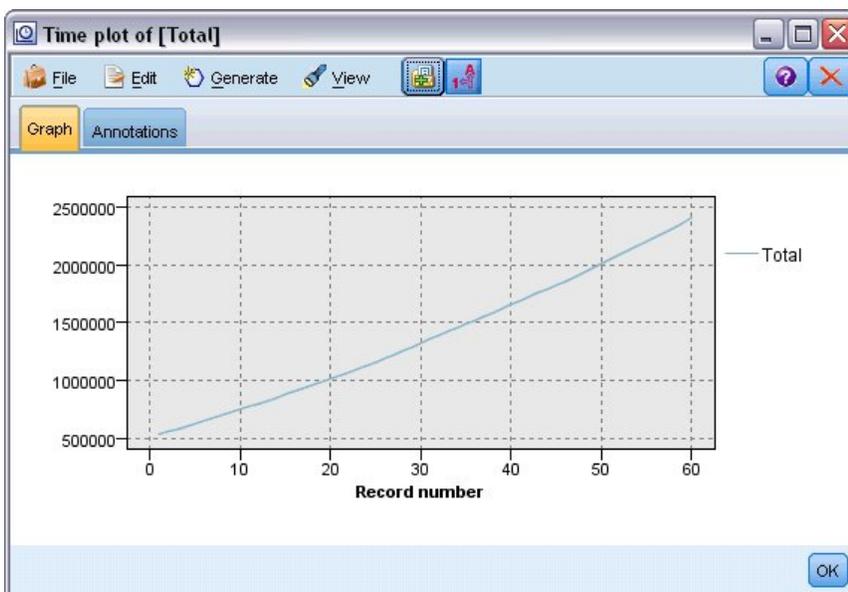


Abbildung 174. Zeitdiagramm des Felds "Total" (Gesamt)

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.

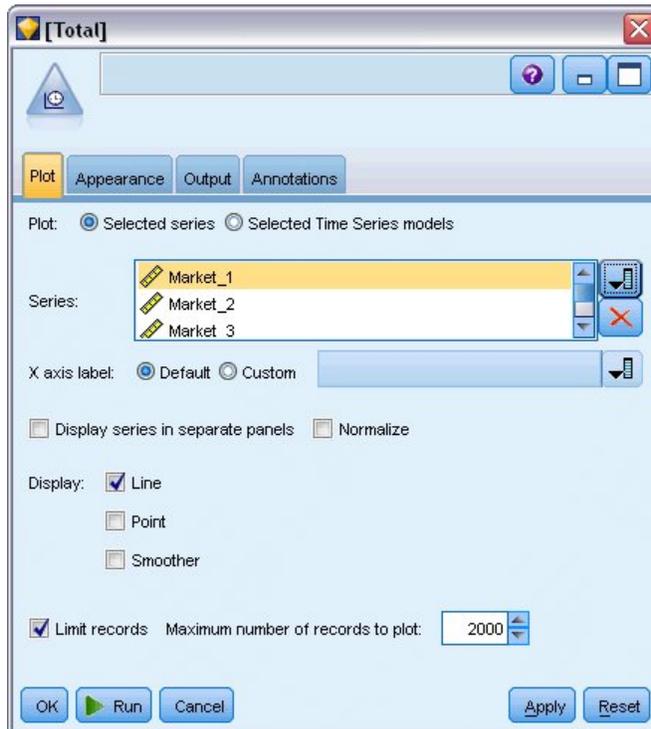


Abbildung 175. Plotten mehrerer Zeitreihen

5. Öffnen Sie den Zeitdiagrammknoten erneut.
6. Entfernen Sie das Feld *Total* (Gesamt) aus der Liste "Series" (Reihe) (wählen Sie es aus und klicken Sie auf das rote X).
7. Fügen Sie der Liste die Felder *Market_1* bis *Market_5* hinzu.
8. Klicken Sie auf **Ausführen**.

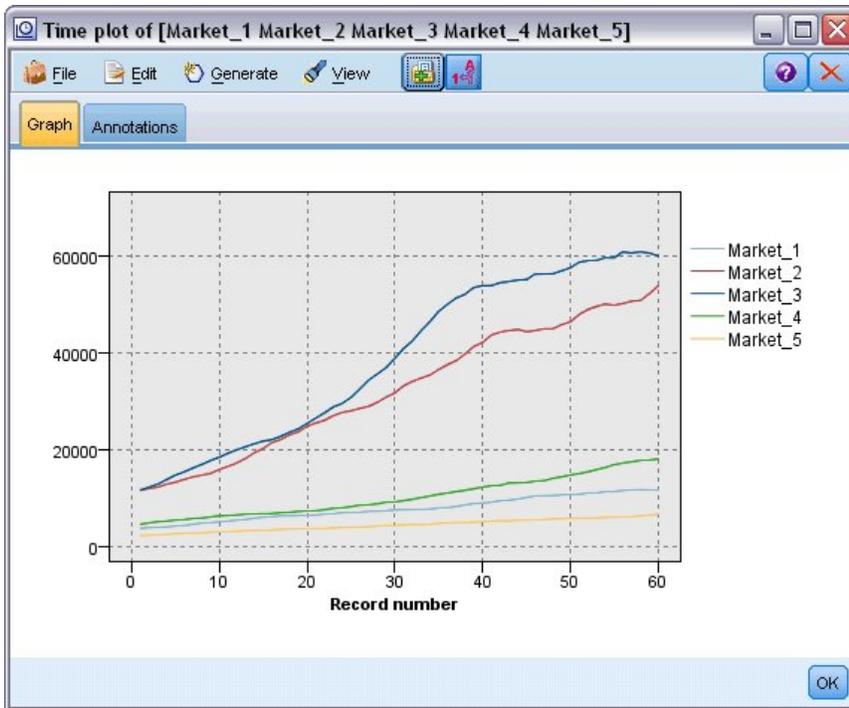


Abbildung 176. Zeitdiagramm mehrerer Felder

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.

1. Fügen Sie einen Füllerknoten an den Filterknoten an.
2. Öffnen Sie den Füllerknoten und klicken Sie auf die Feldauswahlschaltfläche.
3. Wählen Sie `DATE_` aus und fügen Sie das Feld zu **Felder ausfüllen** hinzu.
4. Setzen Sie die Bedingung **Ersetzen** auf **Immer**.
5. Setzen Sie den Wert von **Ersetzen durch** auf `to_date(DATE_)`.

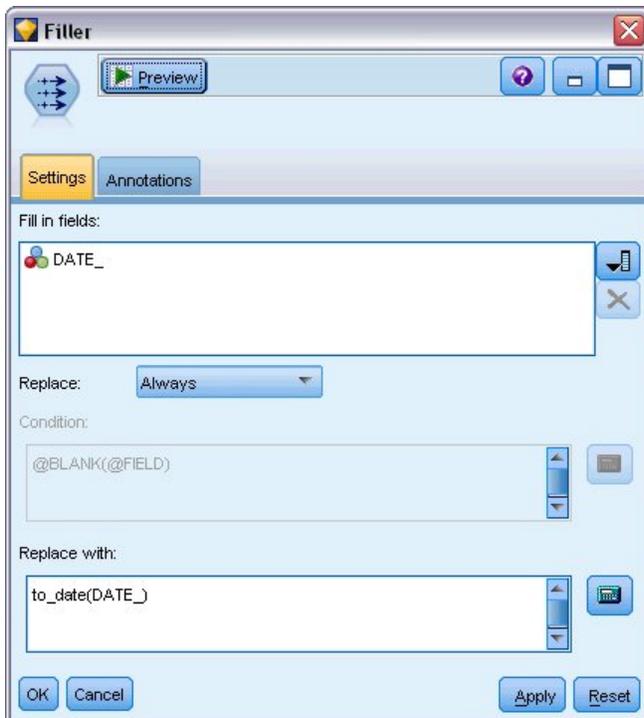


Abbildung 177. Einrichten des Datenspeichertyps

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.

6. Wählen Sie im Menü die Optionsfolge **Tools > Streameigenschaften > Optionen** aus, um das Dialogfeld für die Streamoptionen anzuzeigen.
7. Setzen Sie den Standardwert für **Datumsformat** auf **MON JJJJ** .

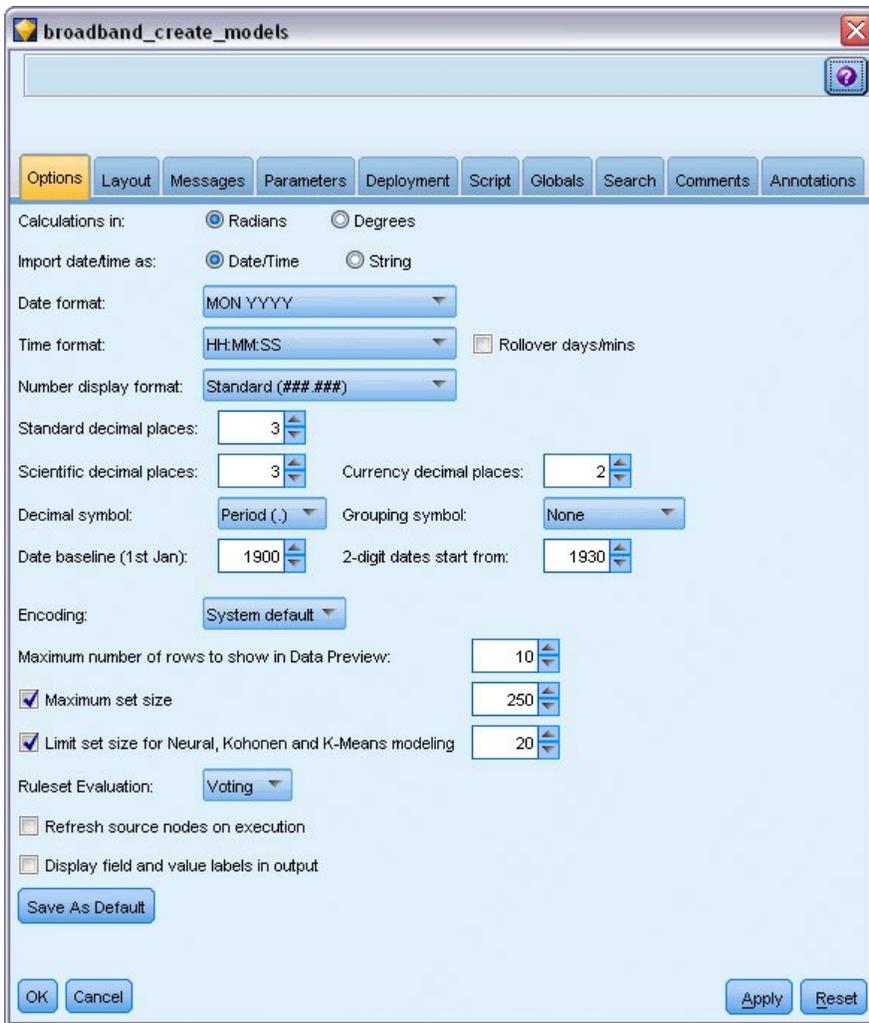


Abbildung 178. Einrichten des Datumsformats

Definieren der Ziele

1. Fügen Sie einen Typknoten hinzu und setzen Sie die Rolle für das Feld *DATE_auf* auf **Keine**. Setzen Sie die Rolle für alle anderen Felder (die Felder *Market_n* und das Feld *Total* (Gesamt)) auf **Ziel**.
2. Klicken Sie auf die Schaltfläche **Werte lesen**, um die Spalte "Werte" auszufüllen.

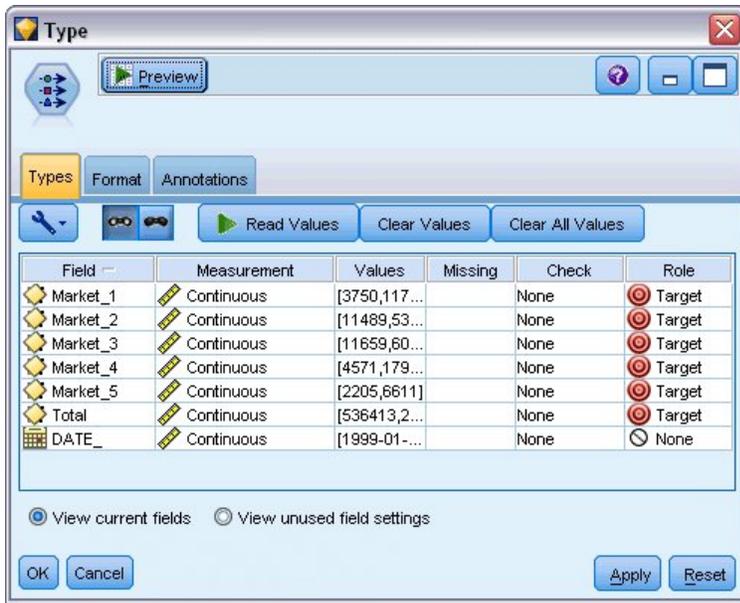


Abbildung 179. Festlegen der Rolle für mehrere Felder

Festlegen der Zeitintervalle

1. Fügen Sie einen Zeitintervallknoten (aus der Palette "Feldfunktionen) hinzu.
2. Wählen Sie auf der Registerkarte "Intervalle" als Zeitintervall die Option **Monate** aus.
3. Wählen Sie die Option **Aus Daten erstellen** aus.
4. Wählen Sie **DATE_** als Erstellungsfeld aus.

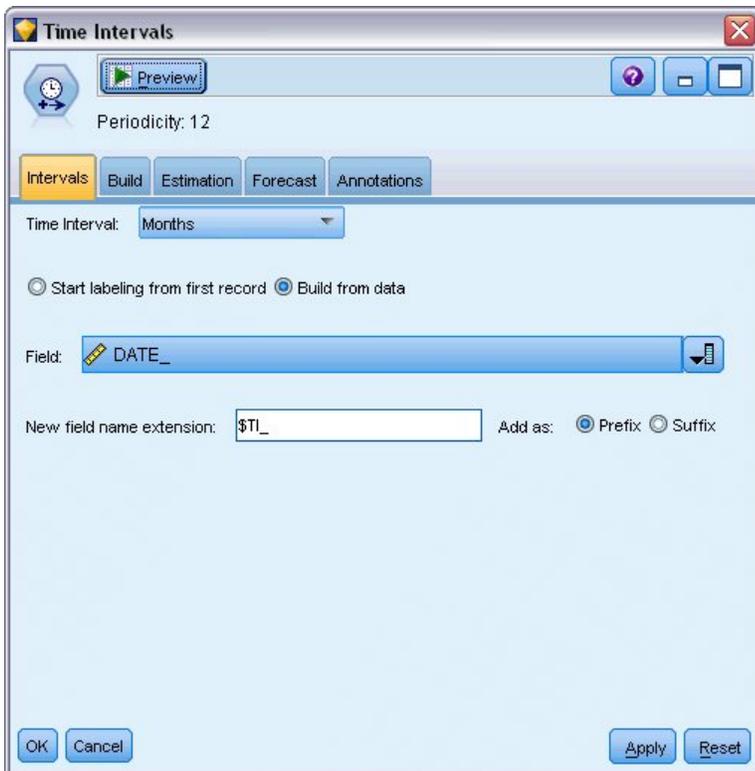


Abbildung 180. Festlegen des Zeitintervalls

5. Wählen Sie auf der Registerkarte "Vorhersage" die Option **Datensätze auf die Zukunft ausdehnen** aus.
6. Setzen Sie den Wert auf 3.
7. Klicken Sie auf **OK**.

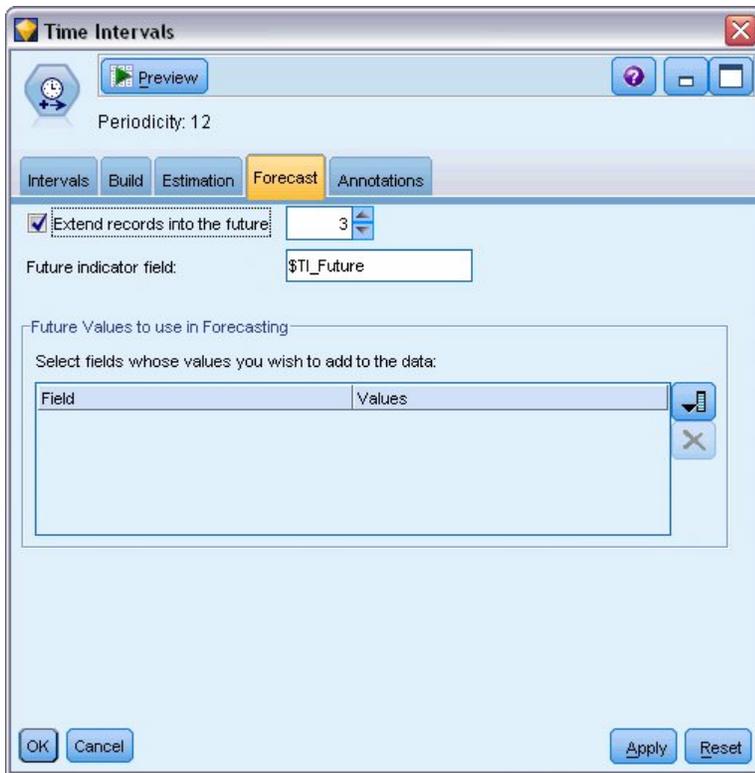


Abbildung 181. Festlegen der Vorhersageperiode

Erstellen des Modells

1. Fügen Sie aus der Modellierungspalette dem Stream einen Zeitreihenknoten hinzu und verbinden Sie ihn mit dem Zeitintervallknoten.
2. 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.

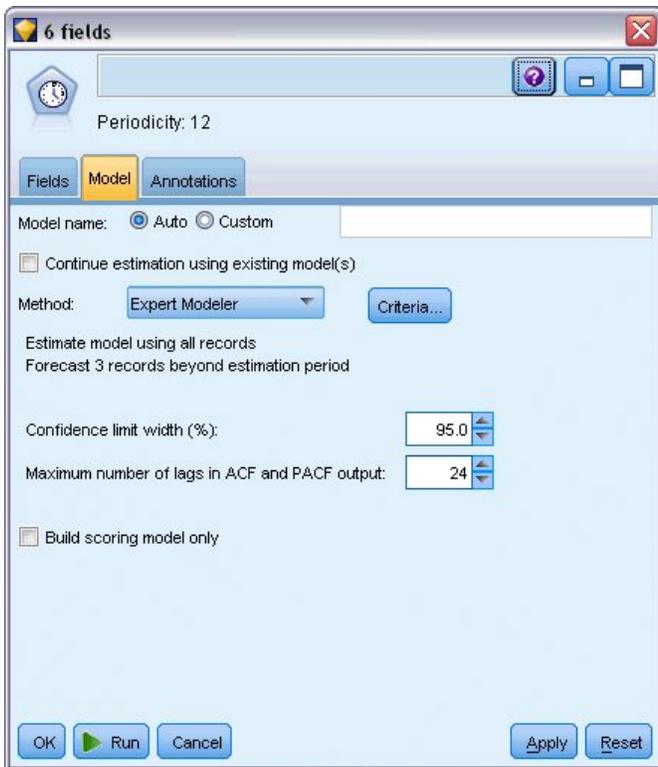


Abbildung 182. Auswählen des Expert Modelers für Zeitreihen

3. Verbinden Sie das Modellanugget vom Typ "Zeitreihe" mit dem Zeitintervallknoten.
4. Verbinden Sie einen Tabellenknoten mit dem Zeitreihenmodell und klicken Sie auf **Ausführen**.

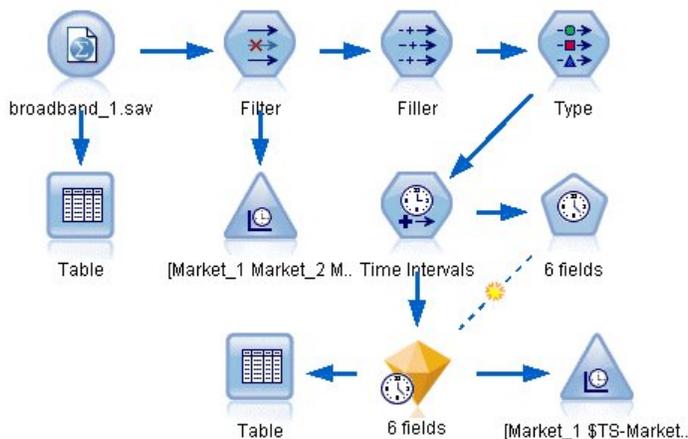


Abbildung 183. Beispielstream zur Anzeige der Zeitreihenmodellierung

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):

Spalte	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 Vorhersagedaten 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 Vorhersagedaten für die Benutzerabonnements und die Konfidenzintervalle für die einzelnen lokalen Märkte.

Untersuchen des Modells

1. Doppelklicken Sie auf das Modellnugget 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.

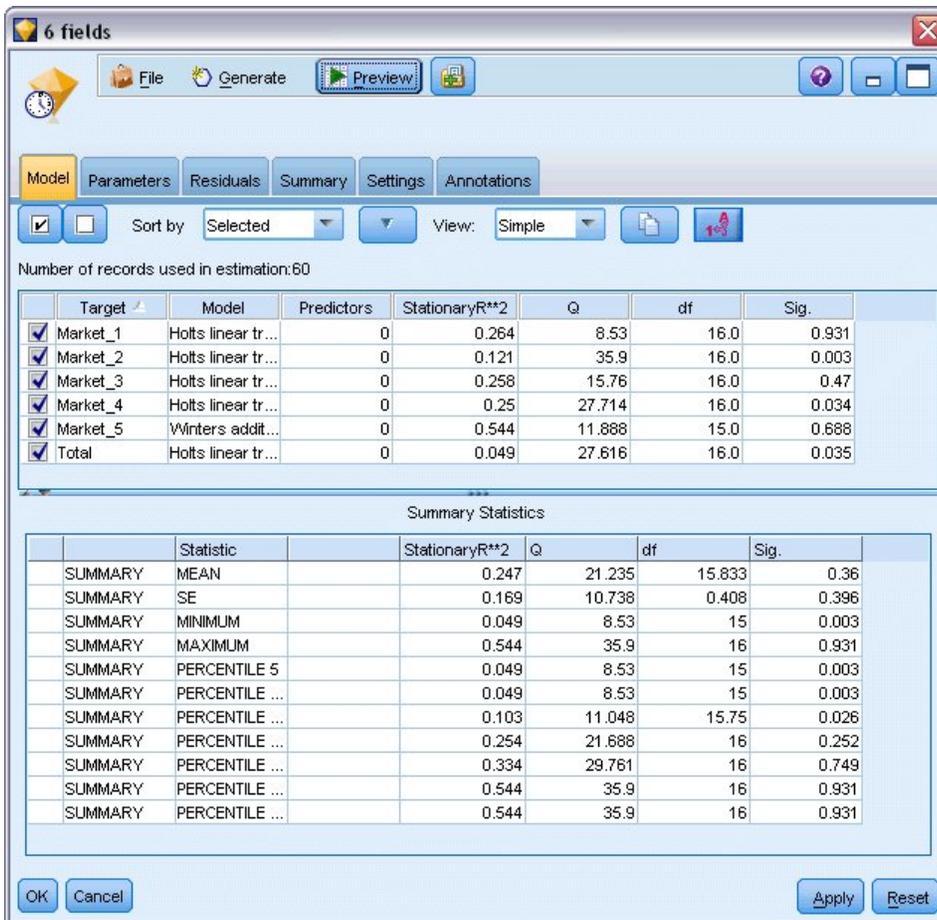


Abbildung 184. Für die Märkte generierte Zeitreihenmodelle

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 Modell: je zufälliger die Fehler, desto besser ist das Modell voraussichtlich. **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.

2. Klicken Sie auf die Dropdown-Liste "Ansicht" und wählen Sie die Option **Erweitert** aus.

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.

Number of records used in estimation:60

	MAPE	MAE	MaxAPE	MaxAE	Norm. BIC	Q	df	Sig.
47	0.94	73.869	2.147	224.517	9.15	8.53	16.0	0.931
76	0.94	314.721	1.867	927.949	12.059	35.9	16.0	0.003
33	0.776	306.877	1.918	1,030.105	12.1	15.76	16.0	0.47
38	0.78	79.49	1.942	233.544	9.329	27.714	16.0	0.034
32	0.936	39.963	2.481	137.633	8.114	11.888	15.0	0.688
74	0.094	1,326.071	0.299	7,062.662	15.243	27.616	16.0	0.035

Summary Statistics

MAPE	MAE	MaxAPE	MaxAE	Norm. BIC	Q	df	Sig.
0.744	356.832	1.776	1,602.735	10.999	21.235	15.833	0.36
0.328	490.119	0.758	2,702.397	2.641	10.738	0.408	0.396
0.094	39.963	0.299	137.633	8.114	8.53	15	0.003
0.94	1,326.071	2.481	7,062.662	15.243	35.9	16	0.931
0.094	39.963	0.299	137.633	8.114	8.53	15	0.003
0.094	39.963	0.299	137.633	8.114	8.53	15	0.003
0.605	65.393	1.475	202.796	8.891	11.048	15.75	0.026
0.858	193.183	1.93	580.747	10.694	21.688	16	0.252
0.94	567.559	2.231	2,538.245	12.886	29.761	16	0.749
0.94	1,326.071	2.481	7,062.662	15.243	35.9	16	0.931
0.94	1,326.071	2.481	7,062.662	15.243	35.9	16	0.931

Abbildung 185. Zeitreihenmodell - erweiterte Anzeige

RMSE steht für Root Mean Square Error (Wurzel des mittleren quadratischen Fehlers), 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.

3. 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.

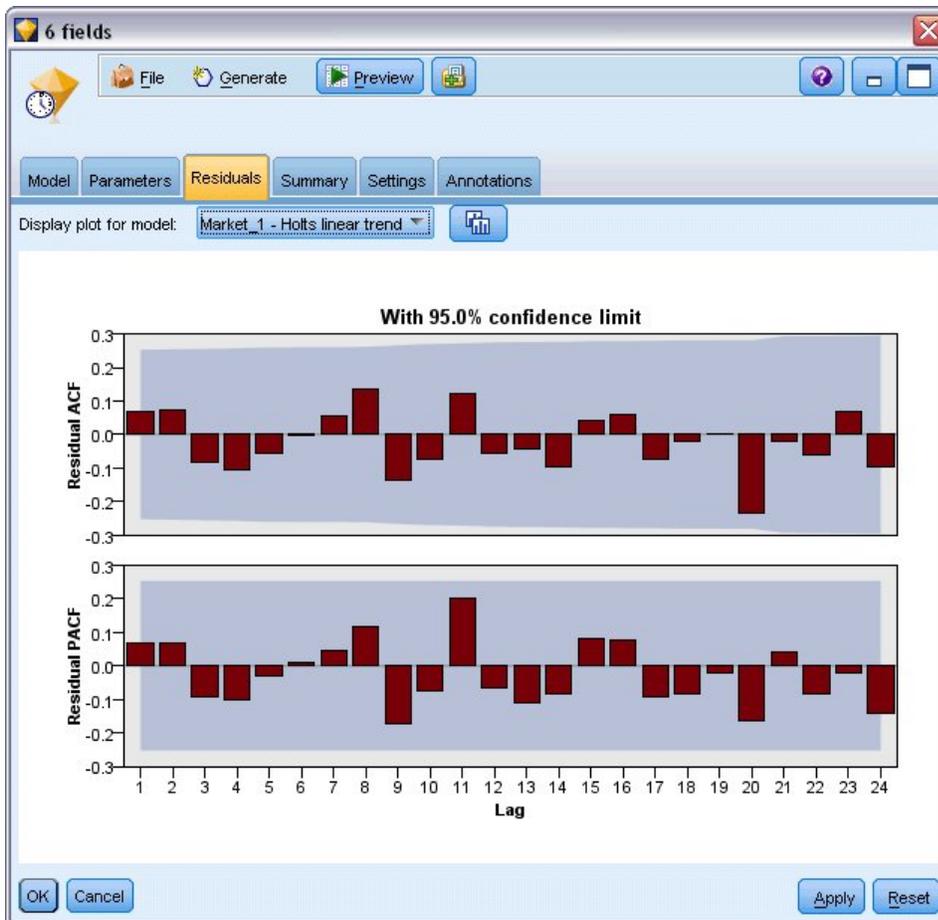


Abbildung 186. ACF- und PACF-Werte für die Märkte

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 festzustellen, 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.

4. 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 festzustellen, 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.

5. Fügen Sie auf der Diagrammpalette einen Zeitdiagrammknoten an das Modellnugget vom Typ "Zeitreihe" an.
6. Inaktivieren Sie auf der Registerkarte "Plot" das Kontrollkästchen **Reihen in gesonderten Fenstern anzeigen**.

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

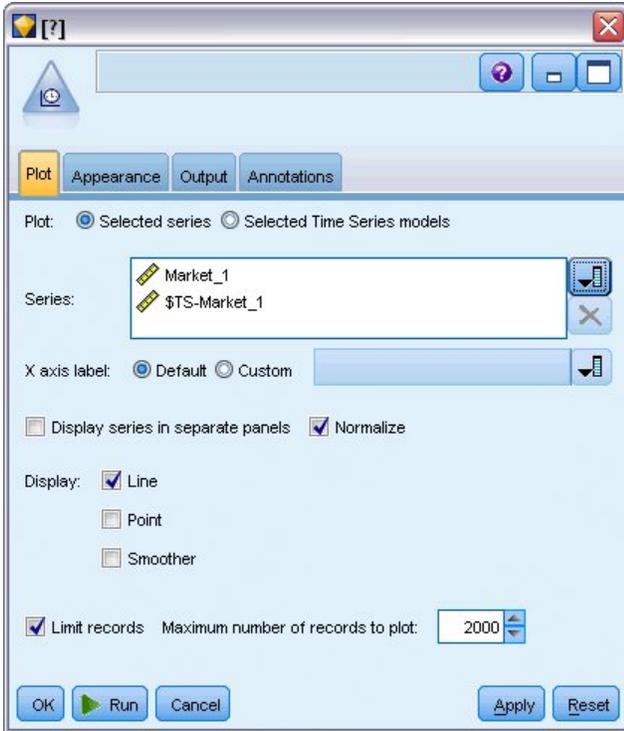


Abbildung 187. Auswählen der zu plottenden Felder

Achten Sie darauf, wie die Vorhersagelinie (*\$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.

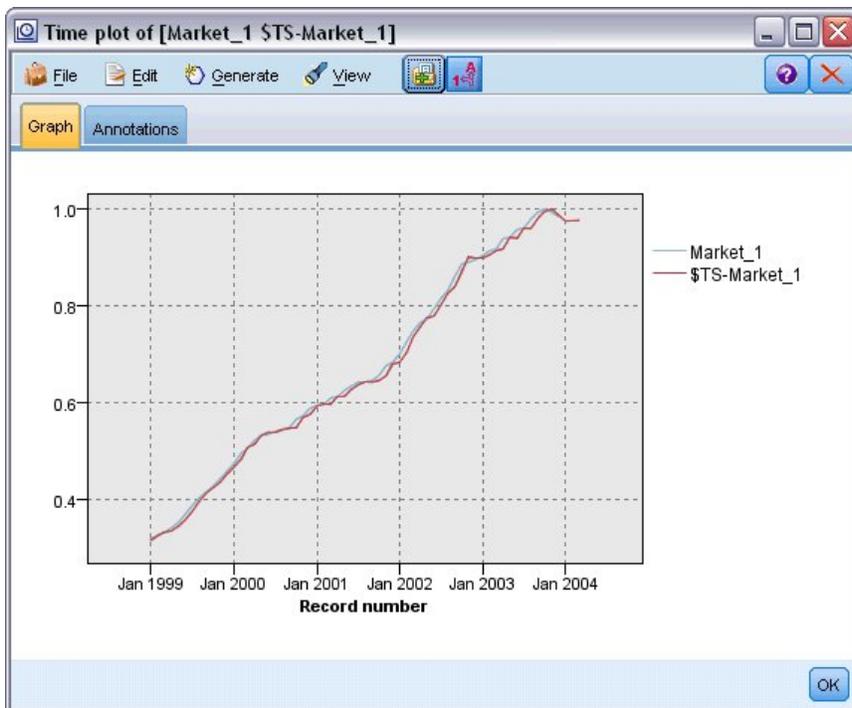


Abbildung 188. Zeitdiagramm der Ist-Daten und der vorhergesagten Daten für Market_1

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

9. Klicken Sie auf **OK**, um das aktuelle Diagramm zu schließen.
10. Öffnen Sie das Modellnugget vom Typ "Zeitreihe".
11. Wählen Sie die Optionsfolge **Datei > Knoten speichern** und geben Sie den Speicherort für die Datei an.
12. 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.
13. 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.
14. Klicken Sie auf die Feldauswahlschaltfläche und fügen Sie der Liste **Series** (Reihe) die Felder **\$TSL-CI-Market_1** und **\$TSUCI-Market_1** hinzu.
15. Klicken Sie auf **Ausführen**.

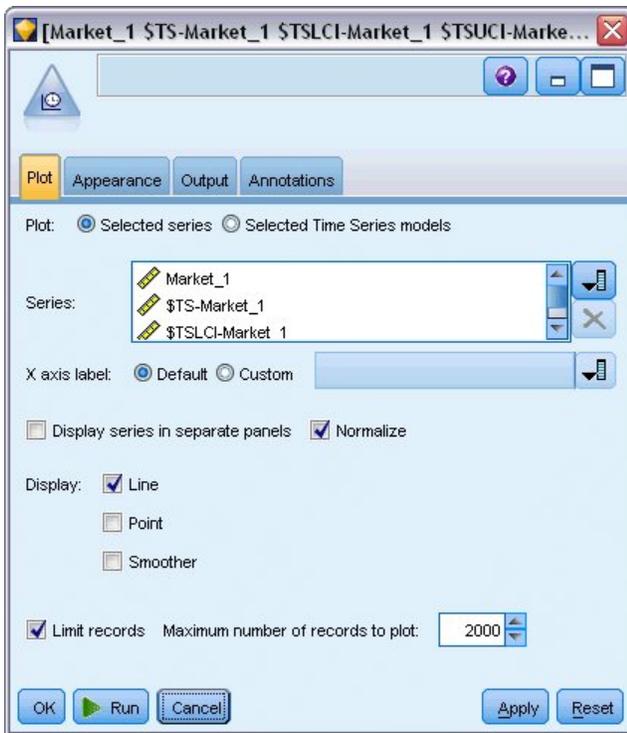


Abbildung 189. Hinzufügen weiterer zu plottender Felder

Sie erhalten dasselbe Diagramm wie zuvor, allerdings diesmal um die obere (*\$TSUCI*) und untere (*\$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. Weitere Informationen finden Sie unter im Thema „Erneutes Anwenden eines Zeitreihenmodells“ auf Seite 171.

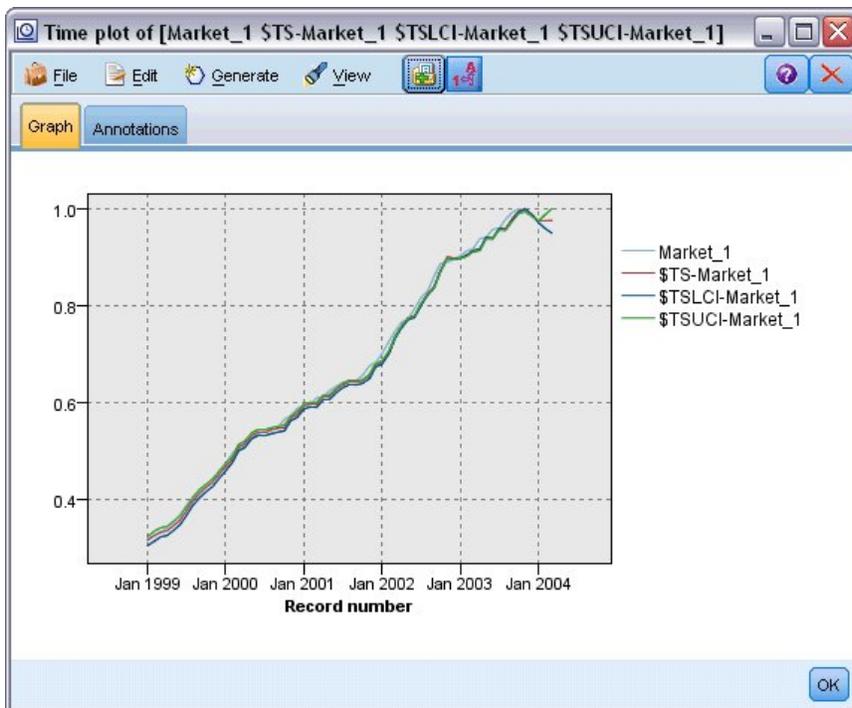


Abbildung 190. Zeitdiagramm, um Konfidenzintervall ergänzt

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. Weitere Informationen finden Sie unter im Thema „Vorhersageerstellung mit dem Zeitreihenknoten“ auf Seite 151.

Wie im ursprünglichen Szenario muss ein Analyst eines nationalen Breitbandproviders monatliche Vorhersagen der Benutzerabonnements für eine Reihe von lokalen Märkten erstellen, um einen 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 Vorhersagehorizont um weitere drei Monate auszuweiten.

In diesem Beispiel wird ein Stream namens *broadband_apply_models.str* verwendet, der Bezug auf die Datendatei *broadband_2.sav* nimmt. Die Dateien stehen im Ordner *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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. Ein eigenes gespeichertes Modell ist nicht erforderlich, da Ihnen im Ordner *Demos* ein Modell zur Verfügung gestellt wurde.

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

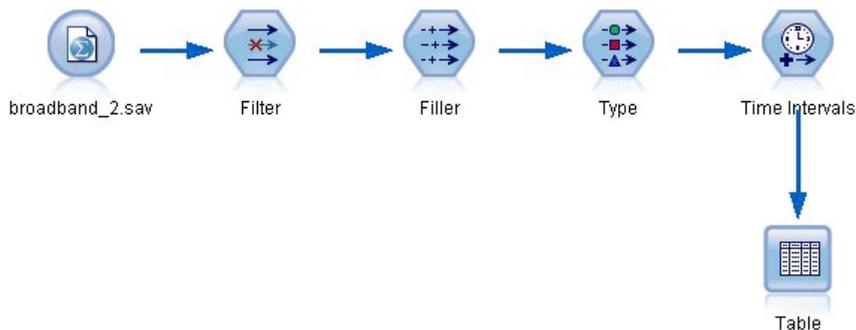


Abbildung 191. Öffnen des Streams

	#1	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	

Abbildung 192. Aktualisierte Umsatzdaten

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

2. 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).

3. Öffnen Sie den Zeitintervallknoten im Stream.
4. Klicken Sie auf die Registerkarte **Vorhersage**.
5. Vergewissern Sie sich, dass **Datensätze auf die Zukunft ausdehnen** auf **3** gesetzt ist.

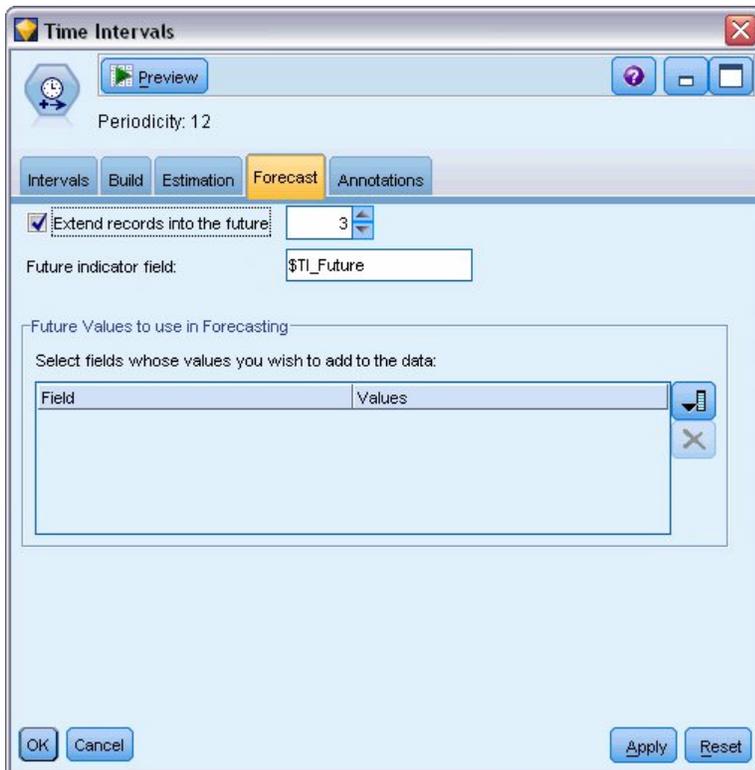


Abbildung 193. Überprüfen der Einstellung für die Vorhersageperiode

Abrufen des gespeicherten Modells

1. Wählen Sie im IBM SPSS Modeler-Menü die Optionsfolge **Einfügen > Knoten aus Datei** und wählen Sie die Datei *TSmodel.nod* im Ordner *Demos* aus (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 Zeitreihenmodellnugget im Erstellungsbereich.

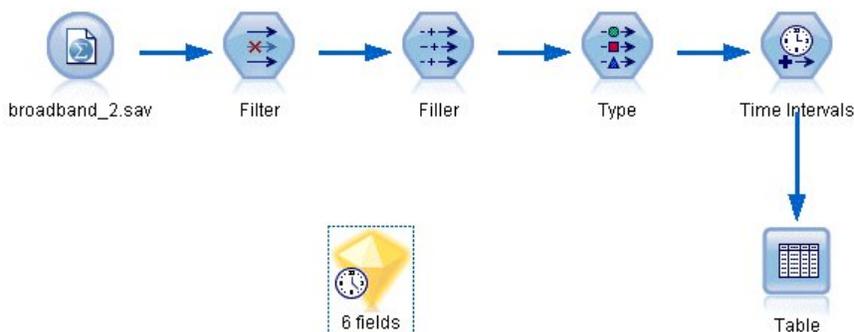


Abbildung 194. Hinzufügen des Modellnuggets

Generieren eines Modellknotens

1. Öffnen Sie das Modellnugget vom Typ "Zeitreihe" und wählen Sie die Optionsfolge **Generieren > Modellierungsknoten erzeugen**.

Dadurch wird ein Zeitreihenmodellierungsknoten im Erstellungsbereich platziert.

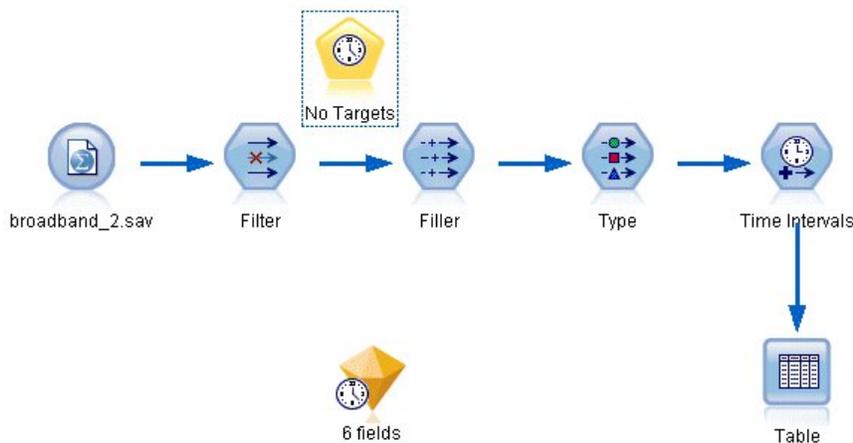


Abbildung 195. Generieren eines Modellierungsknotens aus dem Modellnugget

Generieren eines neuen Modells

1. Schließen Sie das Zeitreihenmodellnugget und löschen Sie es aus dem Erstellungsbereich.
Das alte Modell wurde anhand von 60 Datenzeilen erstellt. Sie müssen ein neues Modell auf der Grundlage der aktualisierten Umsatzdaten erstellen (63 Zeilen).
2. Fügen Sie den neu erzeugten Zeitreihenerstellungsknoten dem Stream hinzu.

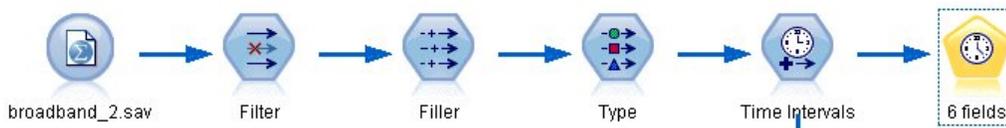


Abbildung 196. Hinzufügen des Modellierungsknotens zum Stream

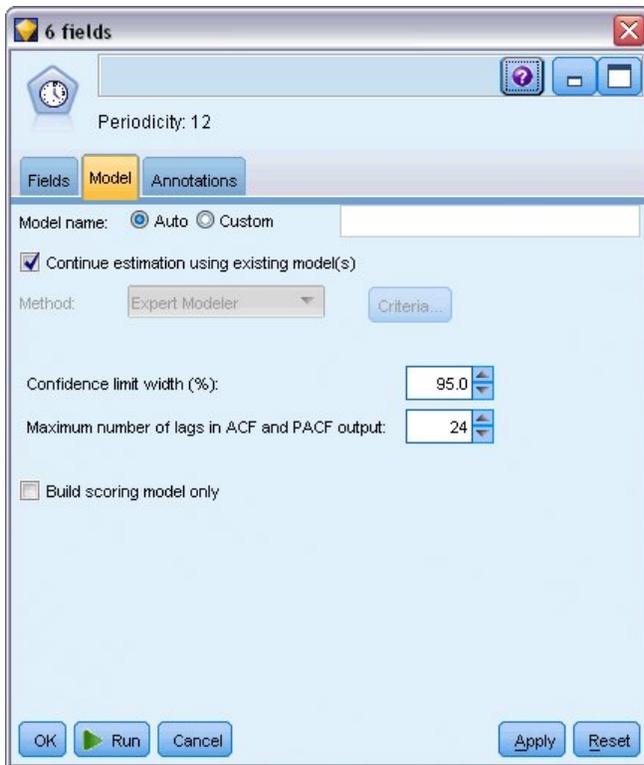


Abbildung 197. Wiederverwenden gespeicherter Einstellungen für das Zeitreihenmodell

3. Öffnen Sie den Zeitreihenknoten.
4. Vergewissern Sie sich auf der Registerkarte **Modell**, dass die Option **Schätzung unter Verwendung bestehender Modelle fortsetzen** aktiviert ist.
5. Klicken Sie auf **Ausführen**, um ein neues Modellnugget im Erstellungsbereich und in der Modellpalette zu platzieren.

Untersuchen des neuen Modells

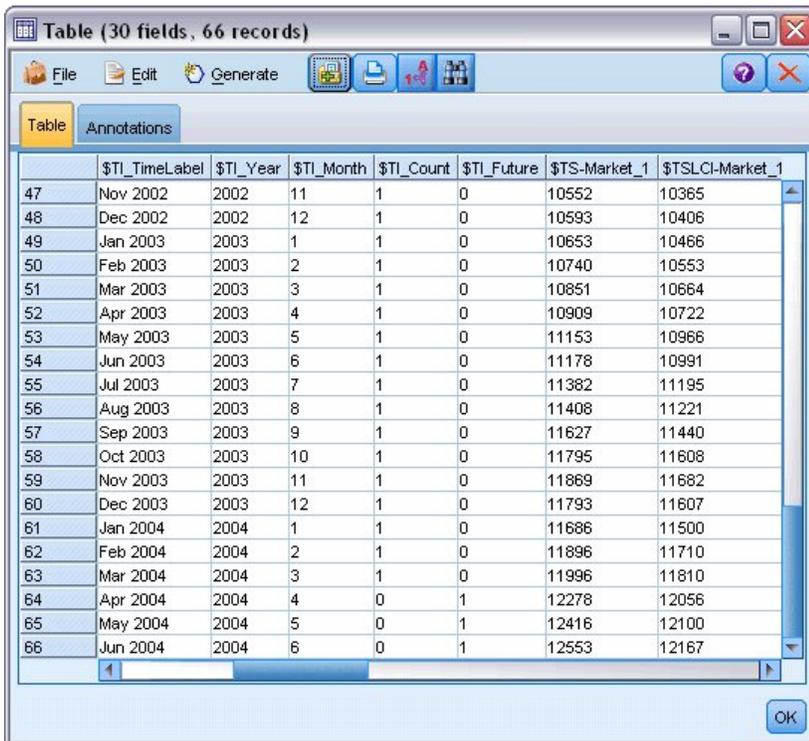


Table (30 fields, 66 records)

	\$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	Dec 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	Mar 2003	2003	3	1	0	10851	10664
52	Apr 2003	2003	4	1	0	10909	10722
53	May 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	Oct 2003	2003	10	1	0	11795	11608
59	Nov 2003	2003	11	1	0	11869	11682
60	Dec 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	Mar 2004	2004	3	1	0	11996	11810
64	Apr 2004	2004	4	0	1	12278	12056
65	May 2004	2004	5	0	1	12416	12100
66	Jun 2004	2004	6	0	1	12553	12167

Abbildung 198. Tabelle mit neuer Vorhersage

1. Verbinden Sie einen Tabellenknoten mit dem neuen Zeitreihenmodellnugget im Erstellungsbereich.
2. Öffnen Sie den Tabellenknoten und klicken Sie auf **Ausführen**.

Auch die Vorhersage 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.

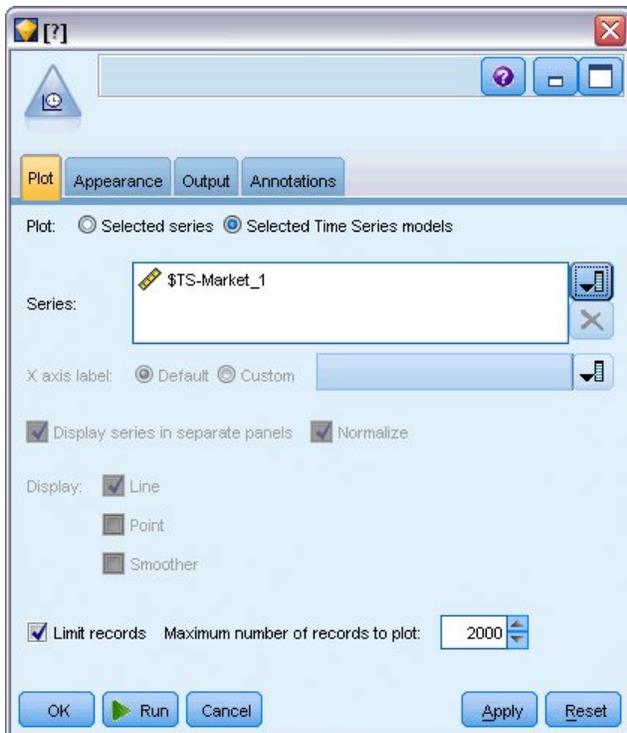


Abbildung 199. Angabe der zu plottenden Felder

3. Fügen Sie dem Zeitreihenmodellnugget einen Zeitdiagrammknoten hinzu.
Diesmal verwenden wir die speziell für Zeitreihenmodelle entworfene Zeitdiagrammanzeige.
4. Wählen Sie auf der Registerkarte "Plot" die Option **Ausgewählte Zeitreihenmodelle**.
5. 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 der Liste hinzuzufügen.
6. 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.

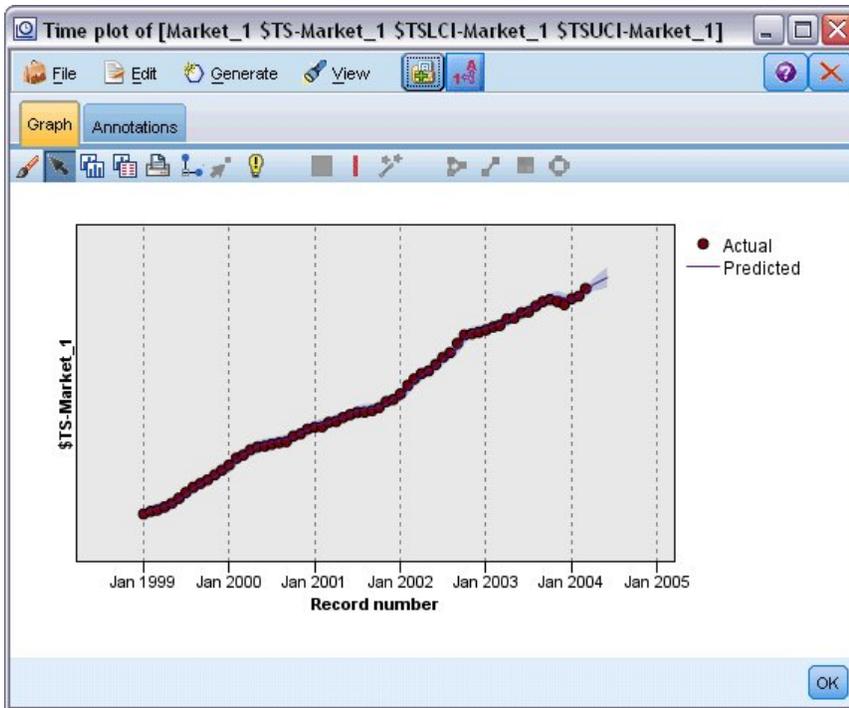


Abbildung 200. Bis Juni erweiterte Vorhersage

Zusammenfassung

Sie haben erfahren, wie Sie gespeicherte Modelle anwenden können, um Ihre vorherigen Vorhersagen 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 15. Vorhersage von Katalogverkäufen (Zeitreihen)

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

In diesem Beispiel wird ein Stream namens *catalog_forecast.str* verwendet, der Bezug auf die Datendatei *catalog_forecast.str* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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 ARI-MA.

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

1. Erstellen Sie einen neuen Stream und fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *catalog_seasfac.sav* verweist.

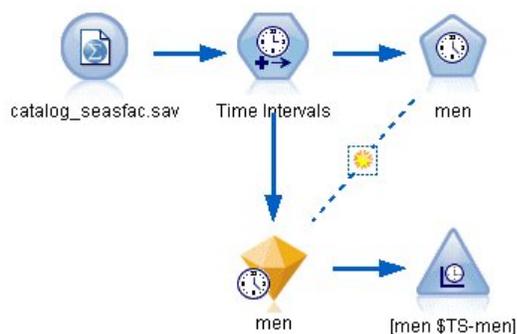


Abbildung 201. Vorhersage von Katalogverkäufen



Abbildung 202. Angabe des Zielfelds

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



Abbildung 203. Festlegen des Zeitintervalls

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

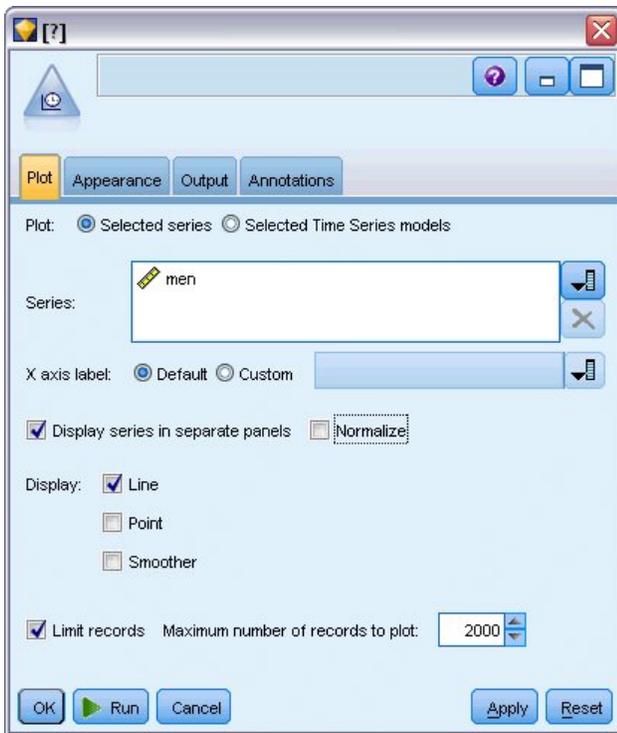


Abbildung 204. Plotten der Zeitreihe

10. Verbinden Sie einen Zeitdiagrammknoten mit dem Zeitintervallknoten.
11. Fügen Sie auf der Registerkarte "Plot" **men** (Herren) zur Liste "Reihe" hinzu.
12. Inaktivieren Sie das Kontrollkästchen **Normalisieren**.
13. Klicken Sie auf **Ausführen**.

Untersuchen der Daten

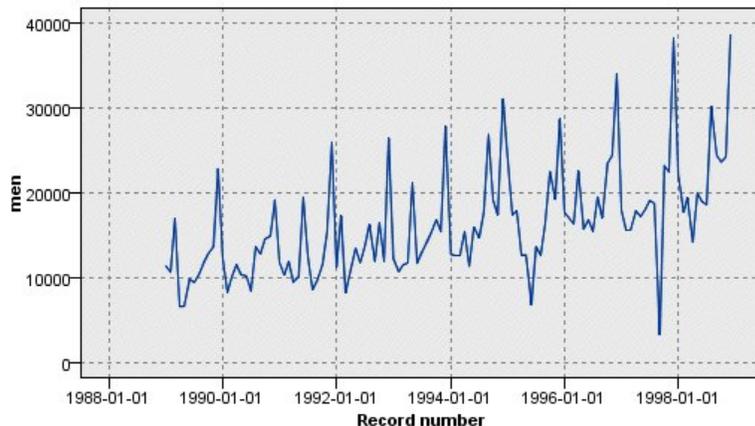


Abbildung 205. Tatsächlicher Umsatz bei der Herrenbekleidung

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.

1. 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 Vorhersage 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 gehören 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 der Herrenbekleidung im Laufe der Zeit hat ein Modell mit linearer Trendkomponente 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 sich darin üben, zu erkennen, wann ein Modell keine gute Anpassung an die Daten darstellt. Dies ist eine entscheidende Fähigkeit für die erfolgreiche Modellerstellung.

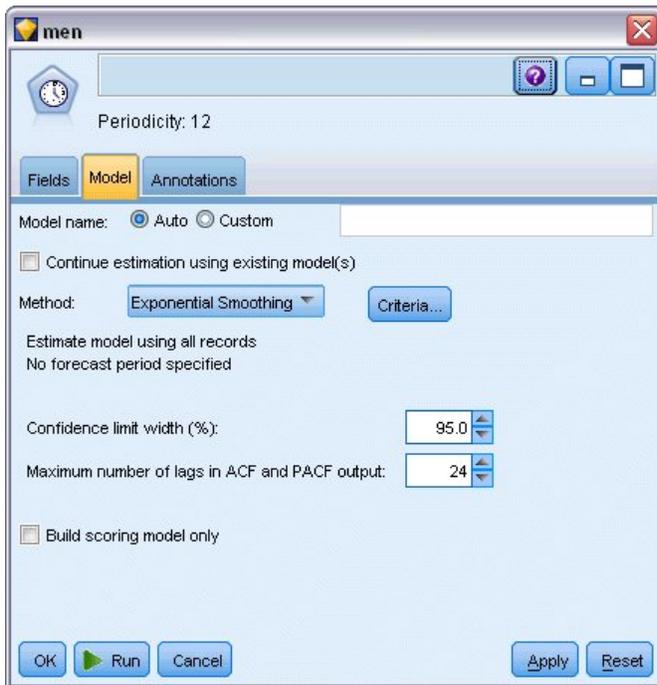


Abbildung 206. Angabe des exponentiellen Glättens

Wir beginnen mit einem einfachen Modell mit exponentiellem Glätten.

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

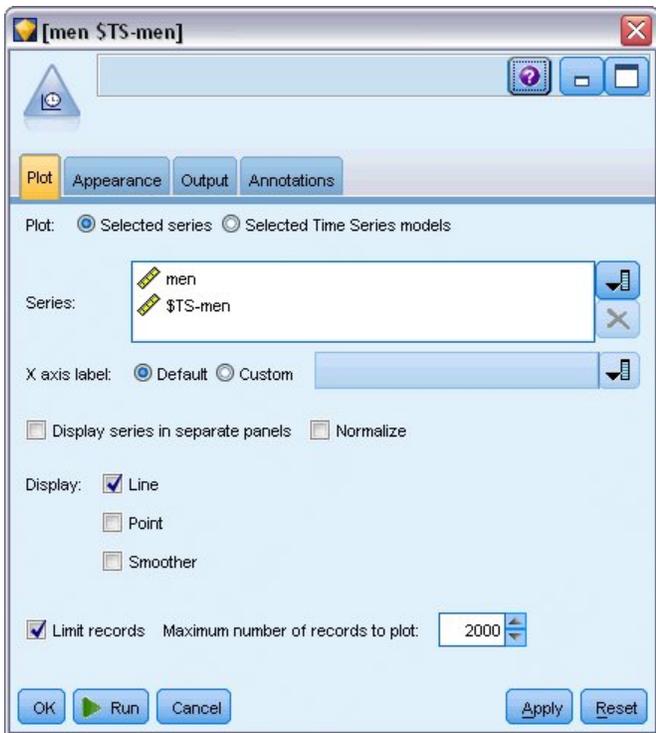


Abbildung 207. Plotten des Zeitreihenmodells

4. Verbinden Sie einen Zeitdiagrammknoten mit dem Modellnugget.
5. Fügen Sie auf der Registerkarte **Plot** **men** (Herren) und **\$TS-men** zur Liste **Reihe** hinzu.
6. Inaktivieren Sie die Kontrollkästchen **Reihen in gesonderten Fenstern anzeigen** und **Normalisieren**.
7. Klicken Sie auf **Ausführen**.

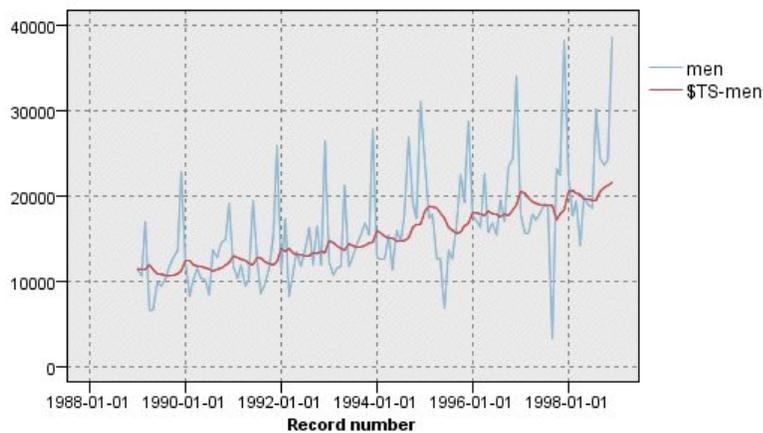


Abbildung 208. Einfaches Modell mit exponentiellem Glätten

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.

8. Klicken Sie auf **OK**, um das Zeitdiagrammfenster zu schließen.



Abbildung 209. 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.

9. Öffnen Sie den Zeitreihenknoten erneut.
10. Klicken Sie auf der Registerkarte **Modell** bei aktivierter Auswahl von **Exponentielles Glätten auf Kriterien**.
11. Wählen Sie im Dialogfeld "Kriterien für exponentielles Glätten" die Option **Linearer Trend nach Holt** aus.
12. Klicken Sie auf **OK**, um das Dialogfeld zu schließen.
13. Klicken Sie auf **Ausführen**, um das Modellnugget neu zu erstellen.
14. Öffnen Sie erneut den Zeitdiagrammknoten und klicken Sie auf **Ausführen**.

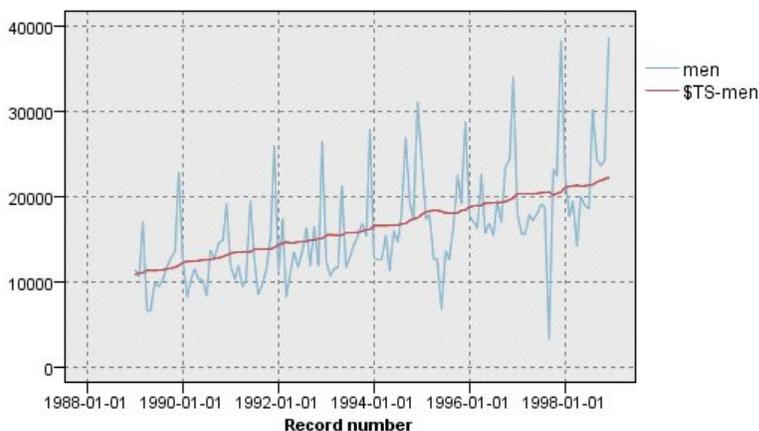


Abbildung 210. Modell für den linearen Trend nach Holt

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.

15. 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 211. Auswahl des Winter-Modells

16. Öffnen Sie den Zeitreihenknoten erneut.
17. Klicken Sie auf der Registerkarte **Modell** bei aktivierter Auswahl von **Exponentielles Glätten** auf **Kriterien**.
18. Wählen Sie im Dialogfeld "Kriterien für exponentielles Glätten" die Option **Multiplikatives Winters-Modell** aus.
19. Klicken Sie auf **OK**, um das Dialogfeld zu schließen.
20. Klicken Sie auf **Ausführen**, um das Modellnugget neu zu erstellen.
21. Öffnen Sie den Zeitdiagrammknoten und klicken Sie auf **Ausführen**.

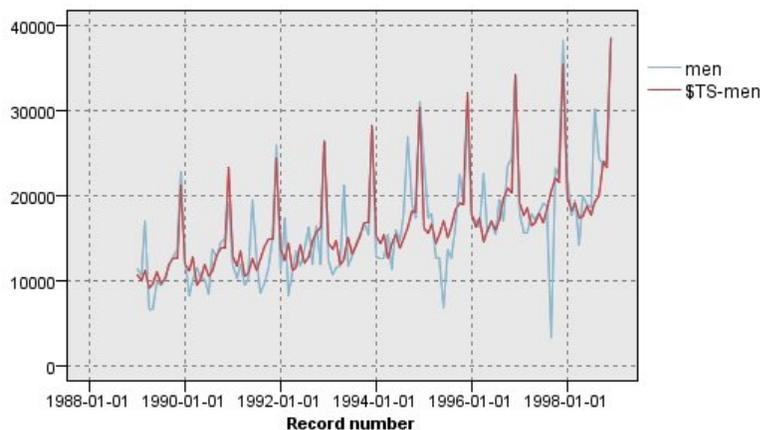


Abbildung 212. Multiplikatives Winters-Modell.

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

Das Dataset 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 Vorhersagemodell 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 Prädiktoren sinnvoll für die Prognostizierung? Ist ein Modell mit Prädiktoren wirklich besser als ein Modell ohne Prädiktoren? Mithilfe der Prozedur ARIMA können wir ein Vorhersagemodell 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 Dataset 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*).

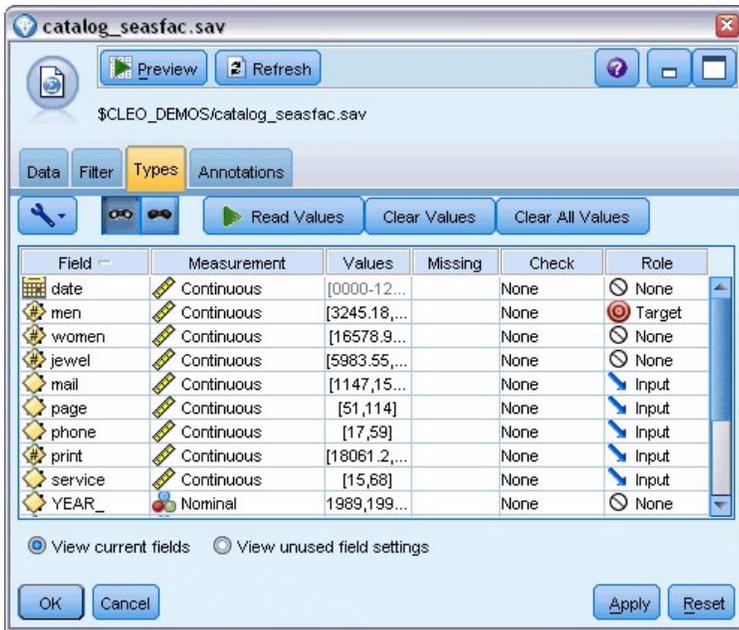


Abbildung 213. Festlegen der Prädiktorfelder

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

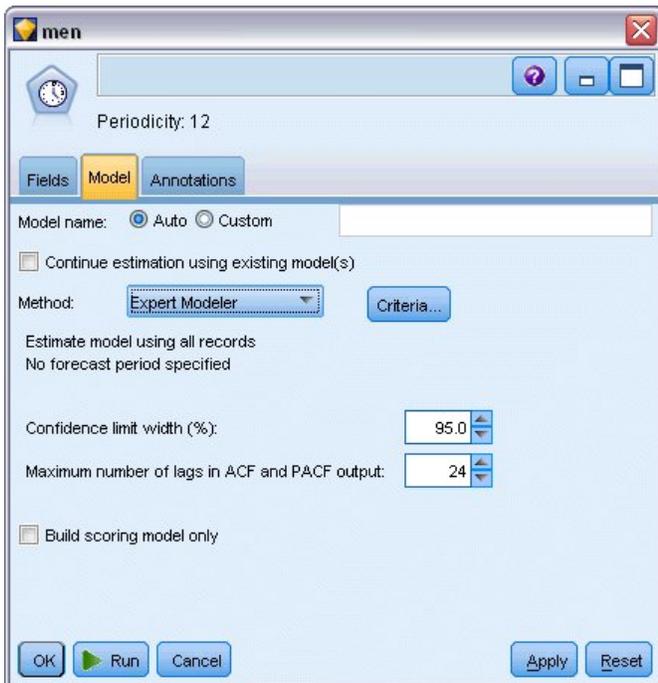


Abbildung 214. Auswahl des Expert Modeler

5. Öffnen Sie den Zeitreihenknoten.

- Legen Sie auf der Registerkarte "Modell" für **Methode** die Option **Expert Modeler** fest und klicken Sie auf **Kriterien**.

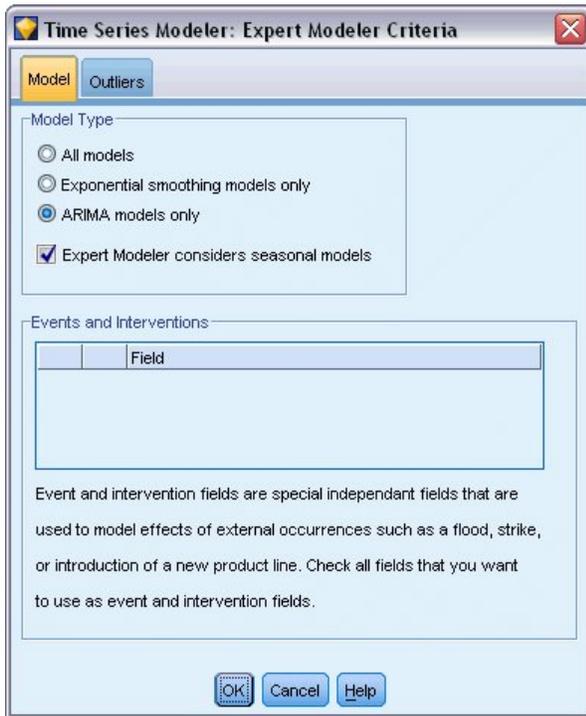


Abbildung 215. Ausschließliche Auswahl von ARIMA-Modellen

- 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 der Registerkarte "Modell" auf **Ausführen**, um das Modellnugget neu zu erstellen.

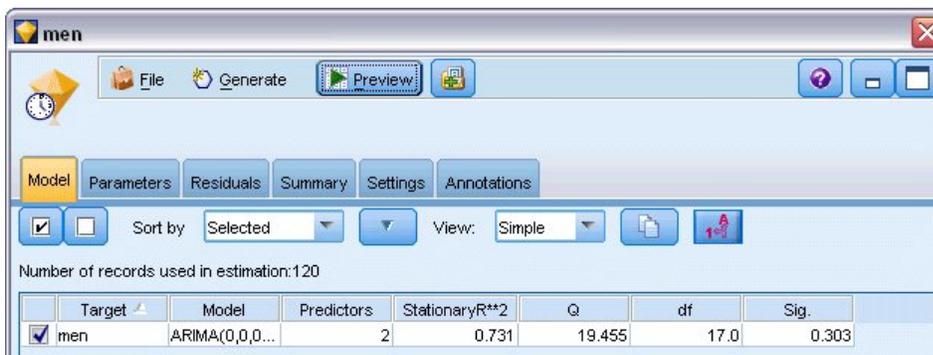


Abbildung 216. Der Expert Modeler wählt zwei Prädiktoren aus.

- Öffnen Sie das Modellnugget.
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 Modellnugget zu schließen.
- Öffnen Sie den Zeitdiagrammknoten und klicken Sie auf **Ausführen**.

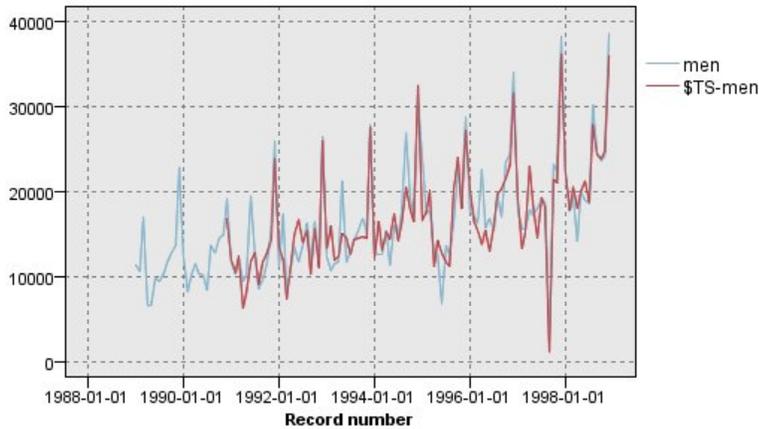


Abbildung 217. ARIMA-Modell mit angegebenen Prädiktoren

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 festgestellt, 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.

13. Klicken Sie auf **OK**, um das Zeitdiagrammfenster zu schließen.
14. Öffnen Sie den Zeitintervallknoten und wählen Sie die Registerkarte *Vorhersage* aus.
15. 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.

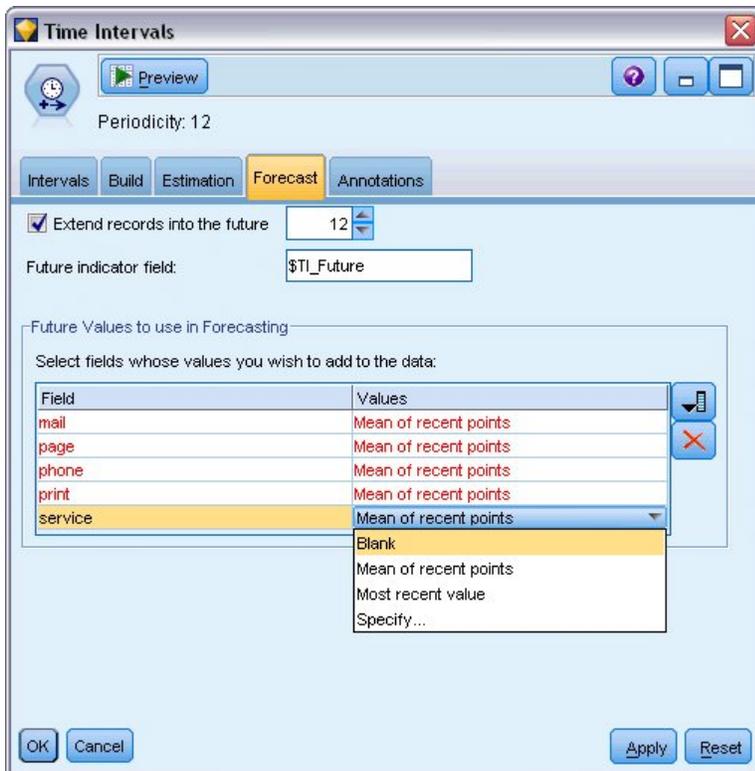


Abbildung 218. Angabe zukünftiger Werte für Prädiktorfelder

16. Klicken Sie in der Gruppe **Zukünftige Werte für Verwendung in Vorhersagen** auf die Feldausschaltfläche rechts neben der Spalte "Werte".
17. 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.)

18. 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.
19. Klicken Sie auf **OK**.
20. Öffnen Sie den Zeitreihenknoten und klicken Sie auf **Ausführen**, um das Modellnugget neu zu erstellen.
21. Öffnen Sie den Zeitdiagrammknoten und klicken Sie auf **Ausführen**.

Die Vorhersage 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 Vorjahres liegt.

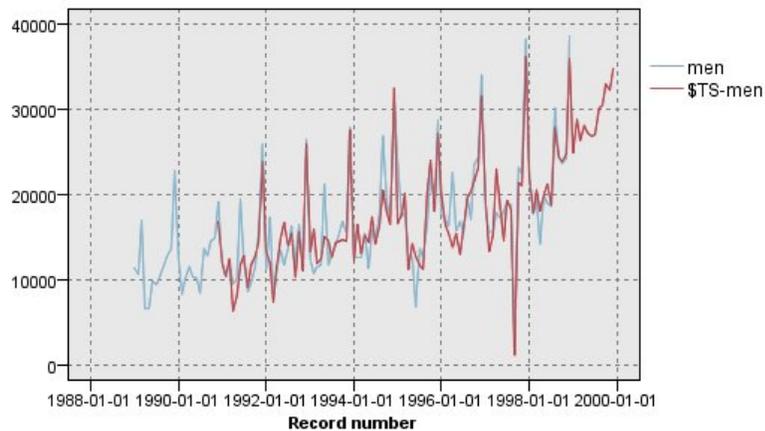


Abbildung 219. Umsatzvorhersage mit angegebenen Prädiktoren

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. Weitere Informationen finden Sie unter im Thema „Erneutes Anwenden eines Zeitreihenmodells“ auf Seite 171.

Kapitel 16. Erstellen von Angeboten für Kunden (Selbstlernfunktion)

Der SLRM-Knoten (Self-Learning Response Model - lernfähiges Antwortmodell) 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 lernfähiges 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 auf die Datendateien *pm_customer_train1.sav*, *pm_customer_train2.sav* und *pm_customer_train3.sav* nimmt. Die Dateien stehen im Ordner *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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.

Table (31 fields, 21,927 records)

	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

OK

Abbildung 220. Reaktionen auf frühere Angebote

Erstellen des Streams

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

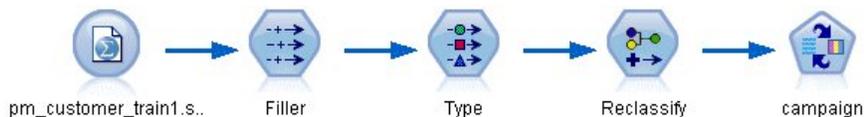


Abbildung 221. SLRM-Beispielstream

2. Fügen Sie einen Füllerknoten hinzu und wählen Sie *campaign* (Kampagne) als das Ausfüllfeld aus.
3. Wählen Sie den Ersetzungstyp **Immer** aus.
4. Geben Sie `to_string(campaign)` im Textfeld "Ersetzen durch" ein und klicken Sie auf **OK**.

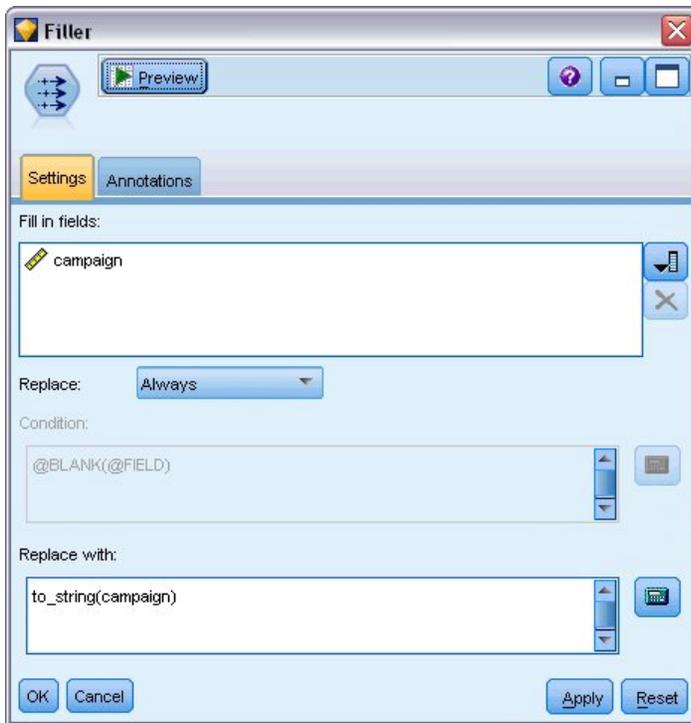


Abbildung 222. Ableiten eines Kampagnenfelds

5. 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 223. Ändern der Typknoteneinstellungen

6. 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**.
7. 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 umcodieren, um ihnen aussagekräftigere Titel zu geben.

8. Fügen Sie dem Typknoten einen Umcodierungsknoten hinzu.
9. Wählen Sie im Feld **Umcodieren** die Option **Vorhandenes Feld** aus.
10. Wählen Sie im Feld **Umcodieren** die Option **campaign** aus.
11. Klicken Sie auf die Schaltfläche **Ermitteln**. Die Kampagnenwerte werden der Spalte *Ursprünglicher Wert* hinzugefügt.
12. Geben Sie in der Spalte *Neuer Wert* folgende Kampagnennamen in die ersten vier Zeilen ein:
 - **Mortgage**
 - **Car loan**
 - **Savings**
 - **Pension**
13. Klicken Sie auf **OK**.



Abbildung 224. Umcodieren der Kampagnennamen

14. Verbinden Sie einen SLRM-Modellierungsknoten mit dem Umcodierungsknoten. Wählen Sie auf der Registerkarte "Felder" als Zielfeld **campaign** und als Zielantwortfeld die Option **response** aus.

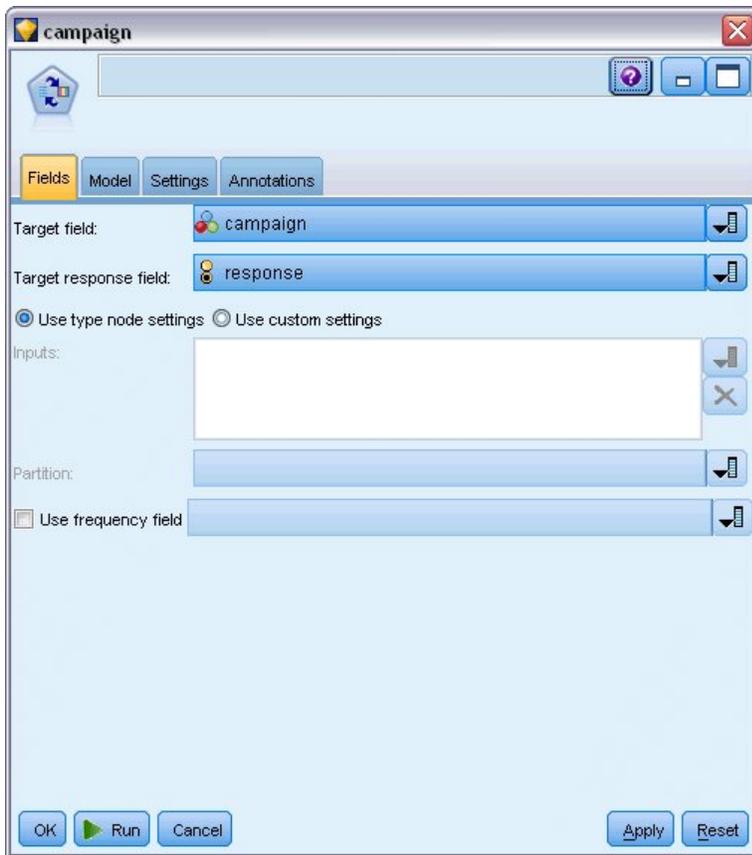


Abbildung 225. Auswahl von Ziel und Zielantwort

15. Reduzieren Sie auf der Registerkarte "Einstellungen" im Feld "Maximale Anzahl an Vorhersagen 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.
16. Vergewissern Sie sich, dass die Option **Reliabilität des Modells berücksichtigen** ausgewählt ist, und klicken Sie auf **Ausführen**.

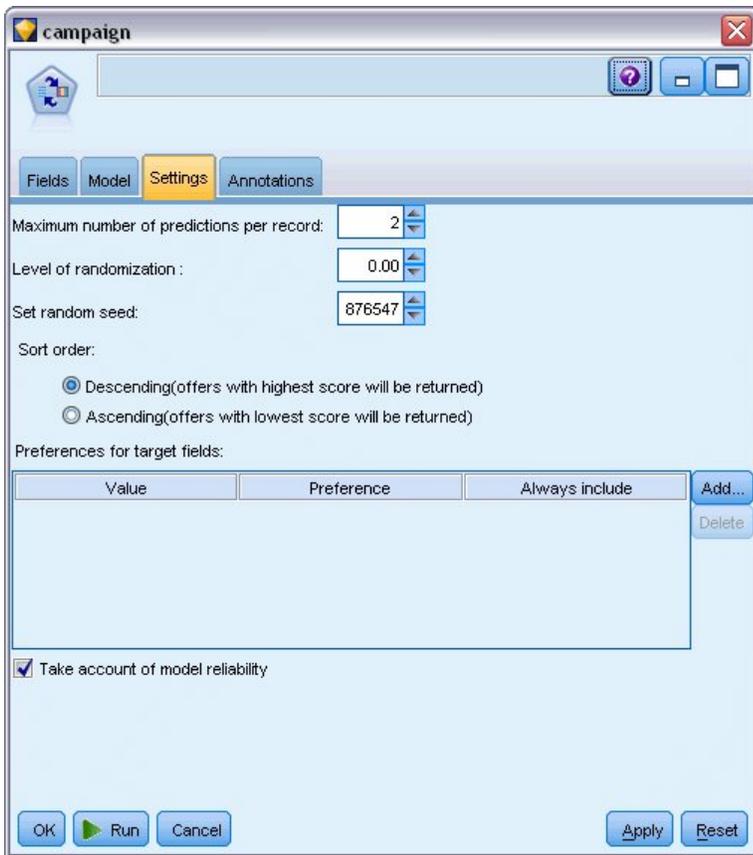


Abbildung 226. SLRM-Knoten - Einstellungen

Durchsuchen des Modells

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

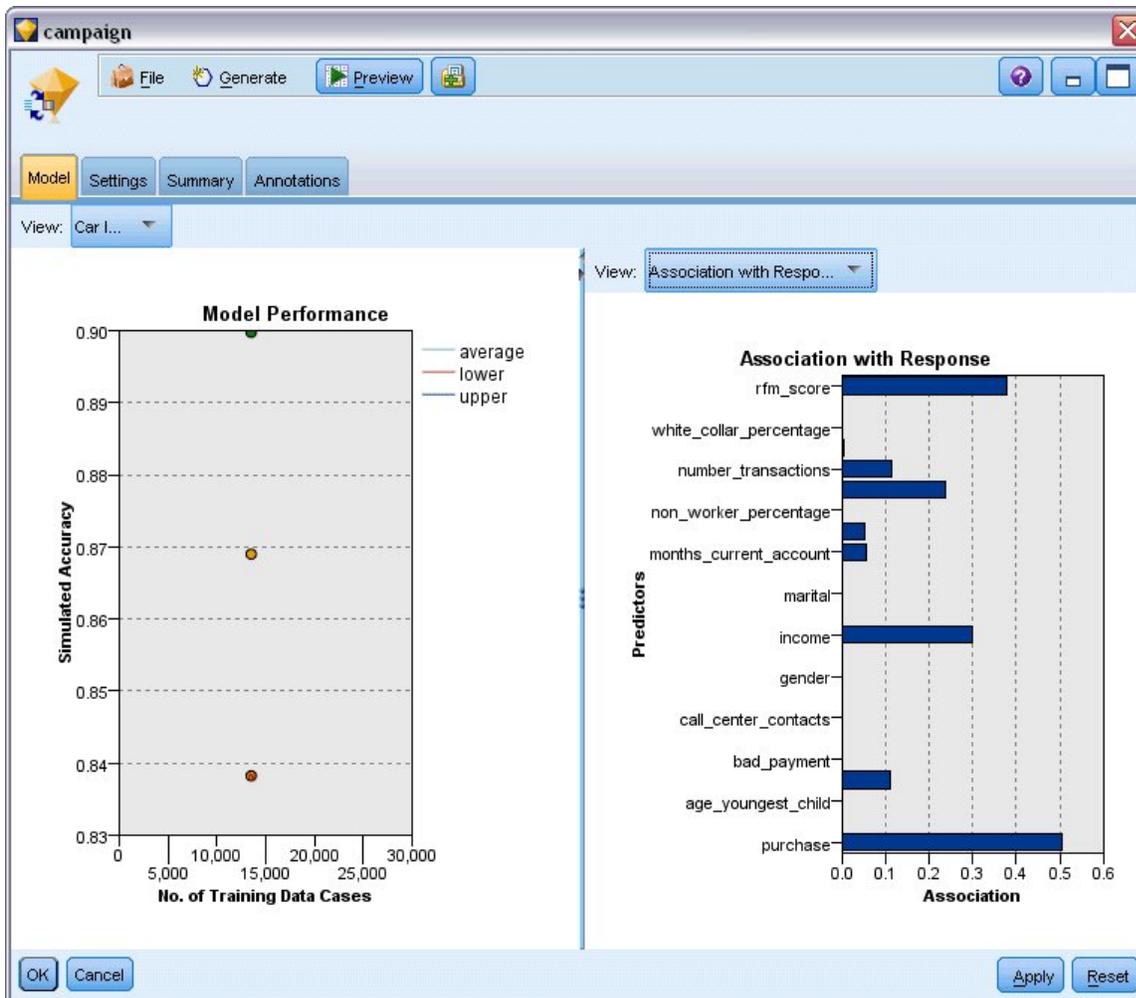


Abbildung 227. SLRM-Modellnugget

3. Schließen Sie das Modellnuggetfenster.
4. Trennen Sie im Streamerstellungsbereich den IBM SPSS Statistics-Dateiquellenknoten, der auf *pm_customer_train1.sav* verweist.
5. Fügen Sie einen Quellenknoten für Statistikdateien 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.

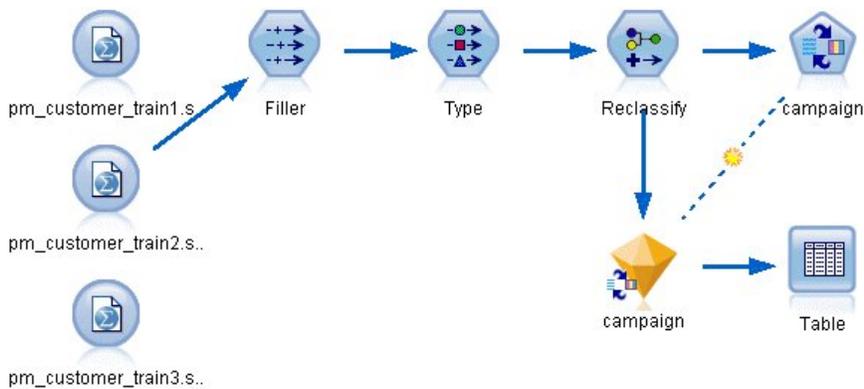


Abbildung 228. Hinzufügen einer zweiten Datenquelle zu einem SLRM-Stream

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

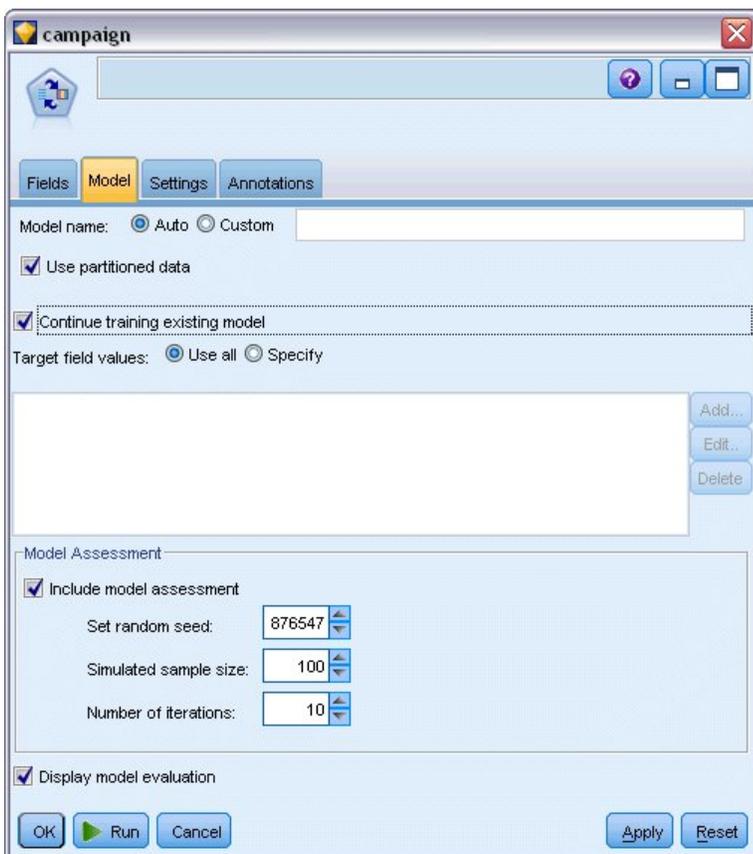


Abbildung 229. Fortsetzen des Trainings des Modells

- Klicken Sie auf **Ausführen**, um das Modellnugget neu zu erstellen. Um die zugehörigen Details anzuzeigen, können Sie auf das Nugget im Erstellungsbereich doppelklicken.
Auf der Registerkarte "Modell" werden nun die revidierten Schätzungen für die Genauigkeit der Vorhersagen für die einzelnen Angebote angezeigt.
- Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf die Datei *pm_customer_train3.sav* im Ordner *Demos* Ihrer IBM SPSS Modeler-Installation verweist, und verbinden Sie ihn mit dem Fül-

lerknoten.

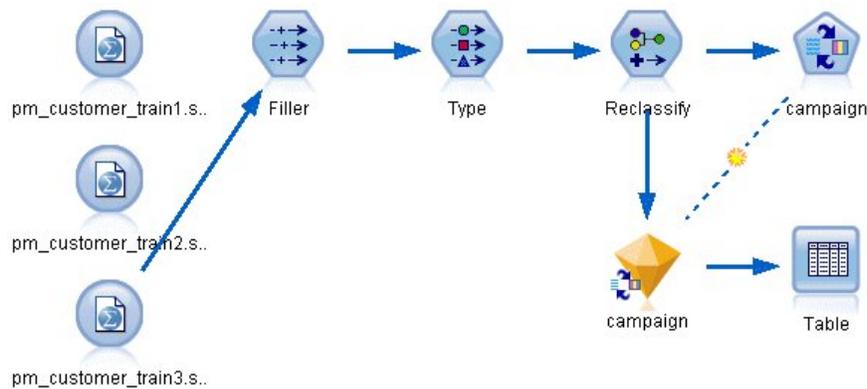


Abbildung 230. Hinzufügen einer dritten Datenquelle zu einem SLRM-Stream

9. Klicken Sie auf **Ausführen**, um das Modellnugget noch einmal neu zu erstellen. Um die zugehörigen Details anzuzeigen, können Sie auf das Nugget im Erstellungsbereich doppelklicken.
10. 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.

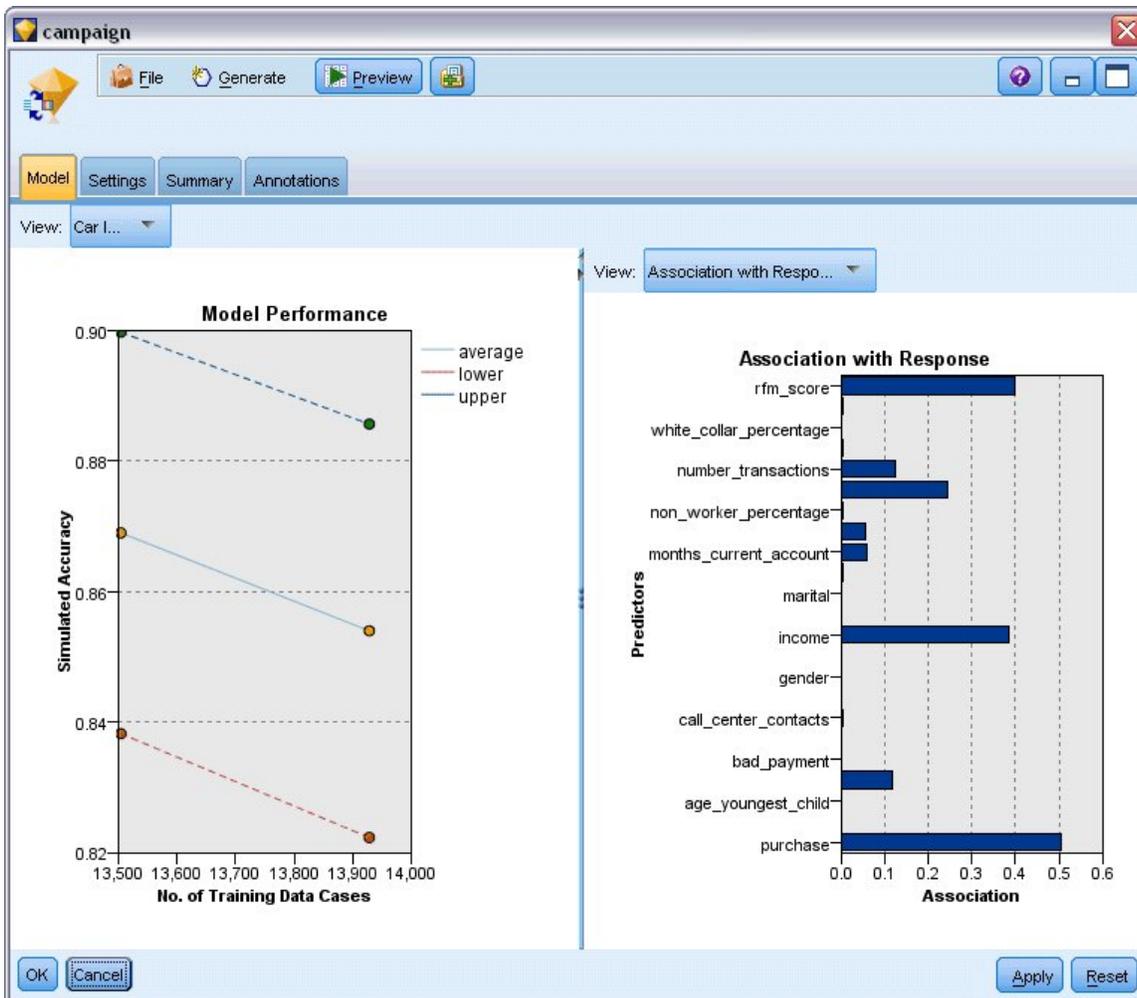


Abbildung 231. Aktualisiertes SLRM-Modellnugget

11. Verbinden Sie einen Tabellenknoten mit dem letzten (dritten) generierten Modell und führen Sie den Tabellenknoten aus.
12. 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 *SSC-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.

Table (35 fields, 27 records)

	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

Abbildung 232. Modellausgabe - vorhergesagte Angebote und Konfidenzwerte

Erläuterungen der mathematischen Grundlagen für die in IBM SPSS Modeler verwendeten Modellierungsmethoden finden Sie im *IBM SPSS Modeler-Algorithmushandbuch*, das sich im Verzeichnis *Documentation* auf der 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 ein Subset der Datensätze für Test- und Validierungszwecke zurückhalten. .

Kapitel 17. Vorhersage von Kreditausfällen (Bayes-Netz)

Mithilfe des Bayes-Netzknosens 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 auf die Datendatei *bankloan.sav* nimmt. 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-Netzmodellen untersucht, um zu ermitteln, welches in dieser Situation die besseren Vorhersagen bietet.

Erstellen des Streams

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *bankloan.sav* im Ordner *Demos* verweist.

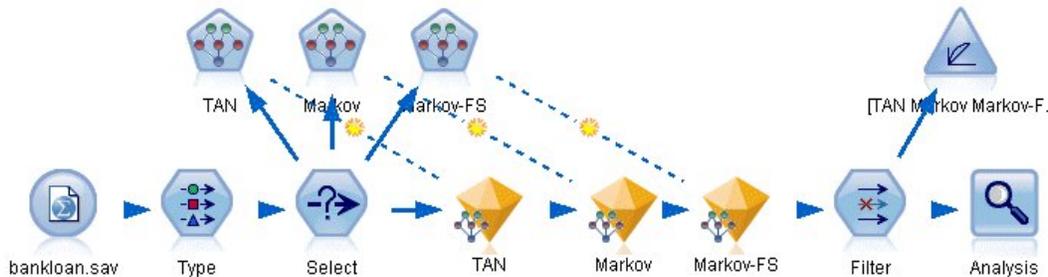


Abbildung 233. Beispielstream für Bayes-Netz

2. Fügen Sie einen Typknoten an den Quellenknoten an und setzen Sie die Rolle des Standardfelds auf **Ziel**. Für alle anderen Felder sollte als Rolle **Eingabe** festgelegt sein.
3. Klicken Sie auf die Schaltfläche **Werte lesen**, um die Spalte *Werte* auszufüllen.



Abbildung 234. Auswahl des Zielfelds

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

4. Fügen Sie dem Typknoten einen Auswahlknoten hinzu.
5. Wählen Sie als Modus die Option **Verwerfen** aus.
6. Geben Sie im Feld "Bedingung" **default = '\$null\$'** ein.

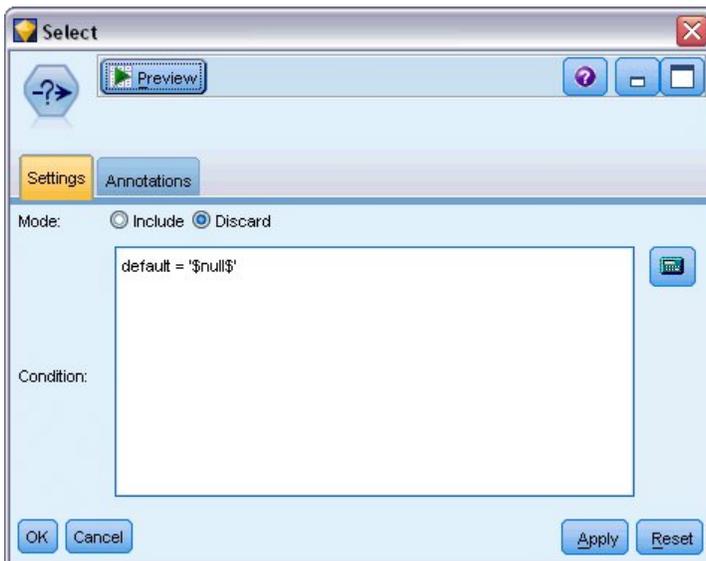


Abbildung 235. Verwerfen von Null-Zielen

Da Sie mehrere verschiedene Typen von Bayes-Netzen 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.

7. Fügen Sie einen Bayes-Netzknoden an den Auswahlknoden an.

8. Wählen Sie auf der Registerkarte "Modell" als Modellnamen **Benutzerdefiniert** (Angepasst) aus und geben Sie im Textfeld den Ausdruck TAN ein.
9. Wählen Sie als Strukturtyp **TAN** aus und klicken Sie auf **OK**.

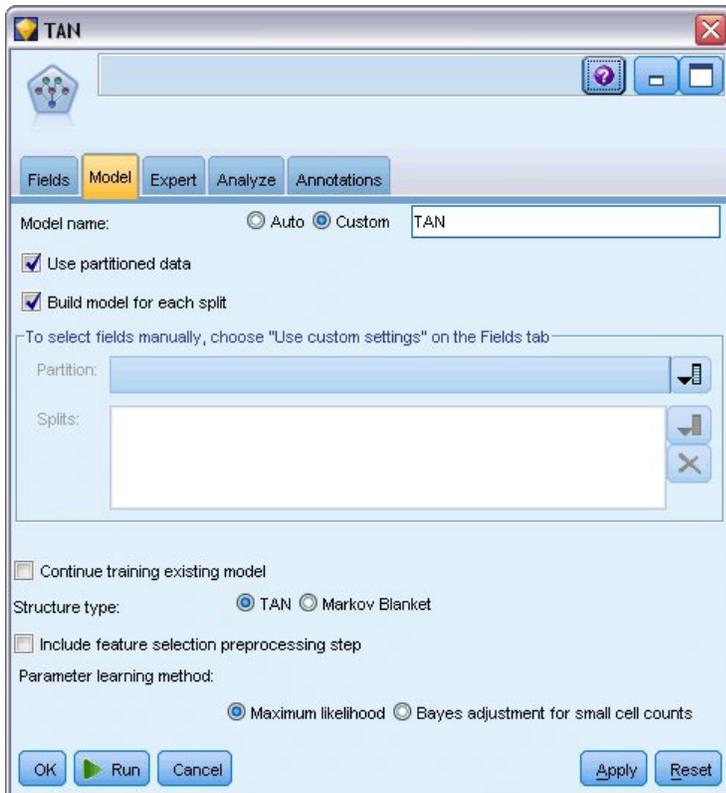


Abbildung 236. Erstellen eines Modells vom Typ "Tree Augmented Naïve Bayes"

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

10. Fügen Sie einen zweiten Bayes-Netzknoten an den Auswahlknoten an.
11. Wählen Sie auf der Registerkarte "Modell" als Modellnamen **Benutzerdefiniert** (Angepasst) aus und geben Sie im Textfeld den Ausdruck Markov ein.
12. Wählen Sie als Strukturtyp **Markov Blanket** aus und klicken Sie auf **OK**.

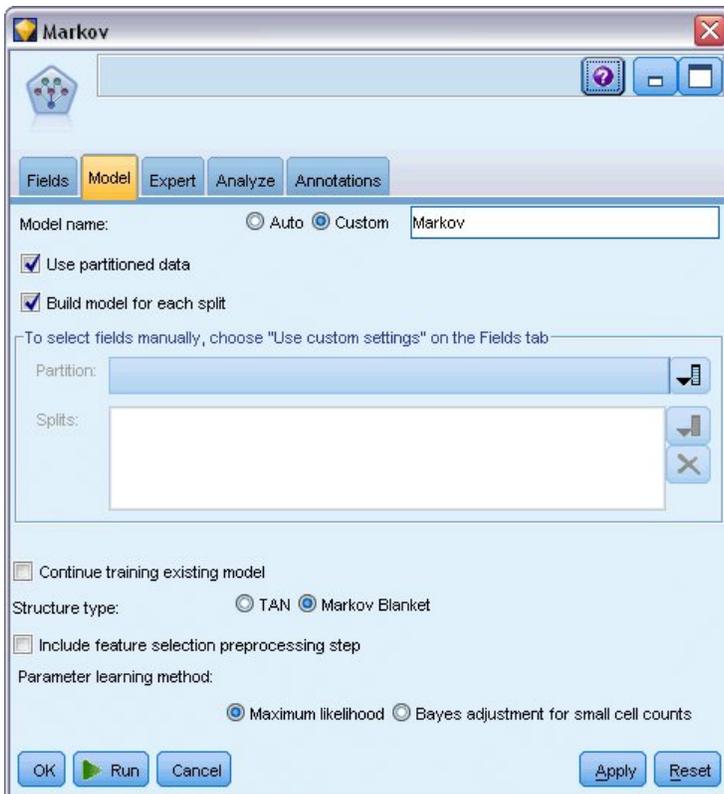


Abbildung 237. Erstellen eines Modells vom Typ "Markov-Decke"

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

13. Fügen Sie einen dritten Bayes-Netzknoden an den Auswahlknoden an.
14. Wählen Sie auf der Registerkarte "Modell" als Modellnamen **Benutzerdefiniert** (Angepasst) aus und geben Sie im Textfeld den Ausdruck Markov-FS ein.
15. Wählen Sie als Strukturtyp **Markov-Decke** aus.
16. Wählen Sie die Option **Vorbereitenden Merkmalauswahlschritt einschließen** aus und klicken Sie auf **OK**.

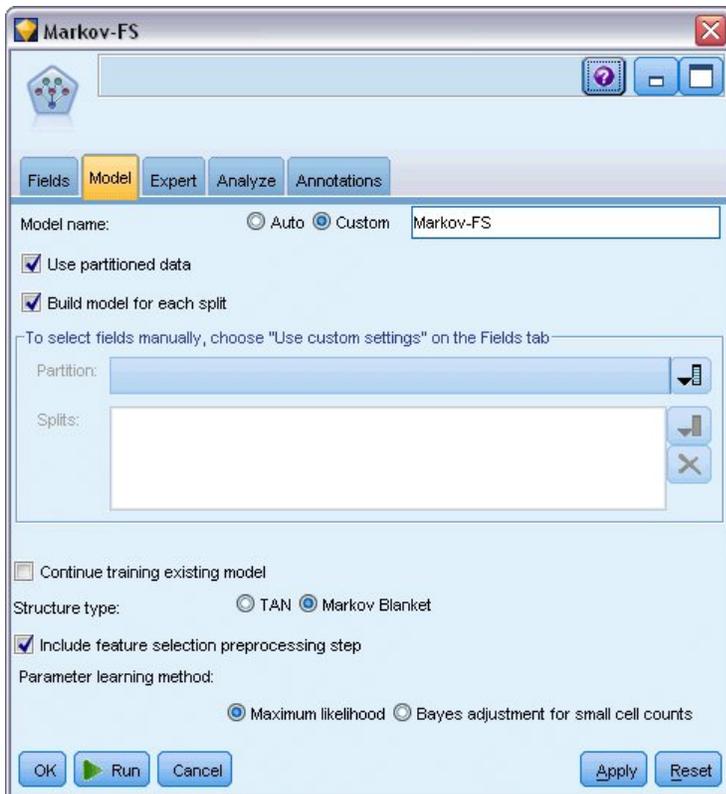


Abbildung 238. Erstellen eines Modells vom Typ "Markov-Decke" mit vorbereitender Merkmalauswahl

Durchsuchen des Modells

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

Die Registerkarte "Modell" des Modellnuggets gliedert sich in zwei Bereiche. Der linke Bereich enthält ein Netzdiagramm 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 *Bedingten Wahrscheinlichkeiten*, also den Wert der bedingten Wahrscheinlichkeit für die einzelnen Knoten und jede Kombination von Werten in ihren übergeordneten Knoten.

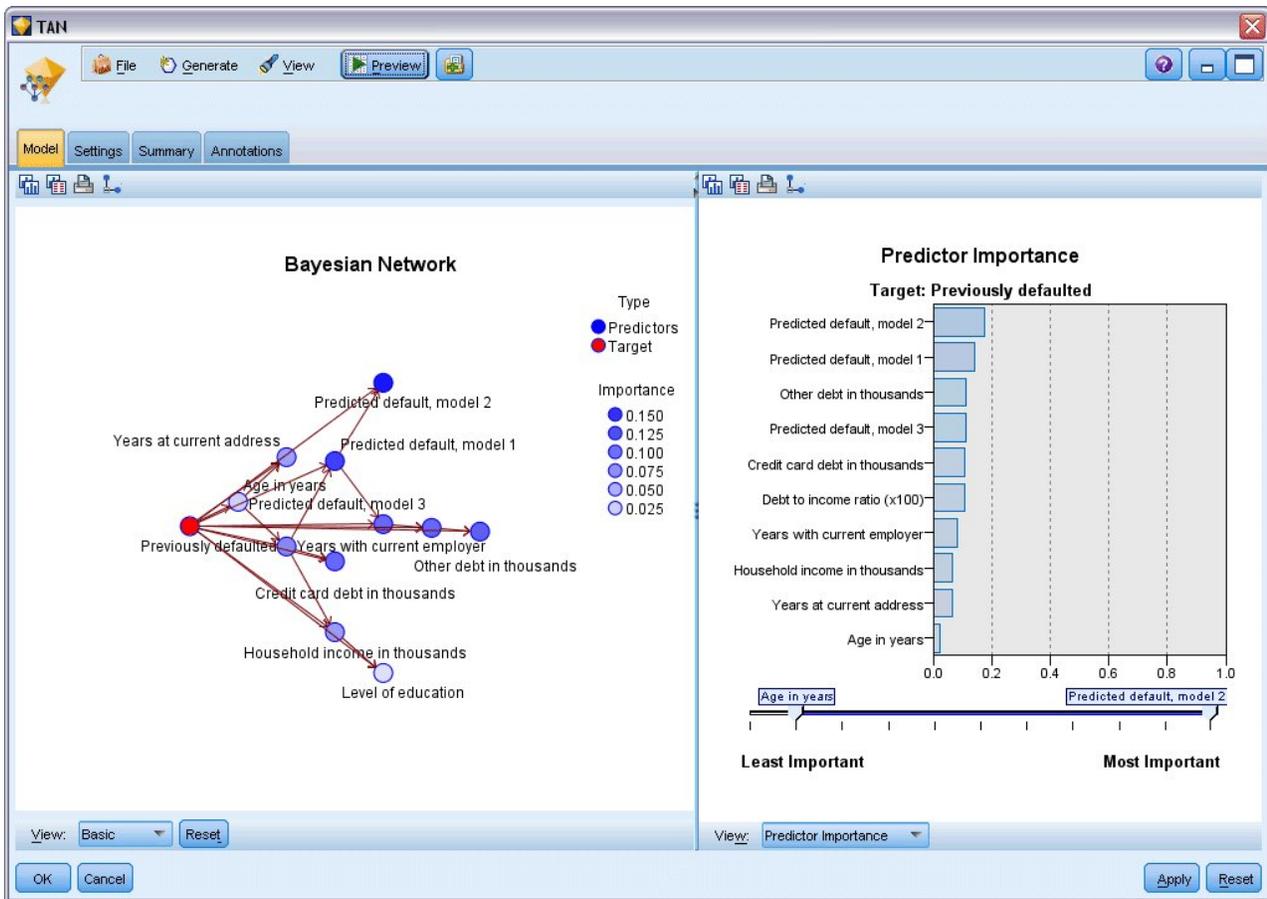


Abbildung 239. Anzeigen eines Modells vom Typ "Tree Augmented Naïve Bayes"

2. Verbinden Sie das TAN-Modellnugget mit dem Markov-Nugget (wählen Sie **Ersetzen** im Warnungsdialog).
3. Verbinden Sie das Markov-Nugget mit dem Markov-FS-Nugget (wählen Sie **Ersetzen** im Warnungsdialog).
4. Richten Sie für bessere Übersichtlichkeit die drei Nuggets mit dem Auswahlknoten aus.

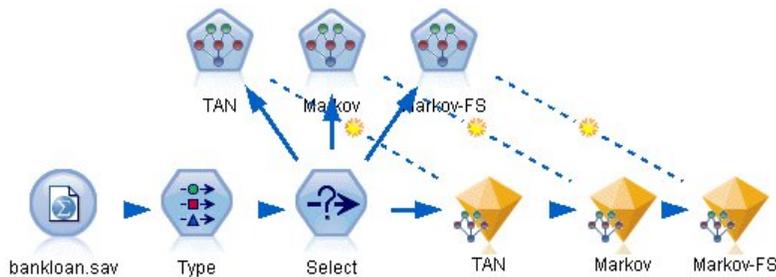


Abbildung 240. Ausrichten der Nuggets im Stream

5. Um die Modellausgaben zugunsten größerer Klarheit im Evaluierungsdiagramm (das Sie erstellen werden) umzubenennen, müssen Sie einen Filterknoten an das Markov-FS-Modellnugget anfügen.
6. Benennen Sie in der rechten Spalte *Feld* jeweils "\$B-default" in "TAN", "\$B1-default" in "Markov" und "\$B2-default" in Markov-FS um.

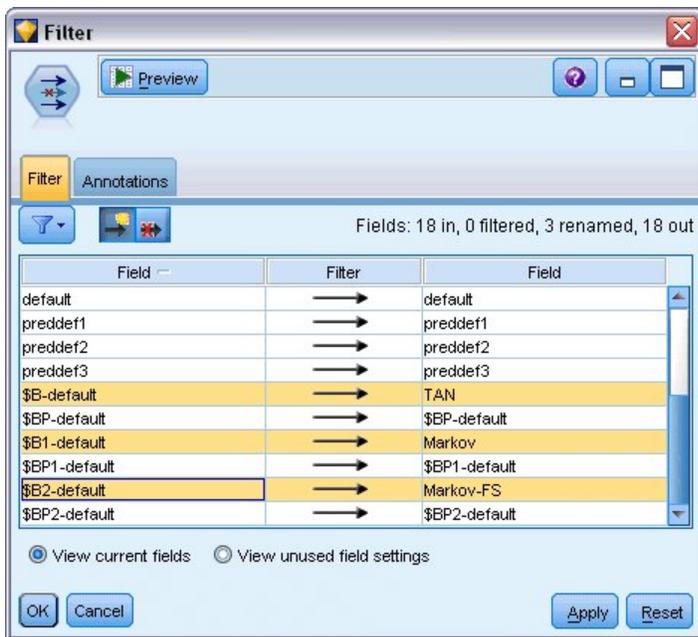


Abbildung 241. Umbenennen von Feldnamen für Modelle

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

7. Fügen Sie einen Evaluierungsdiagrammknoten 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.

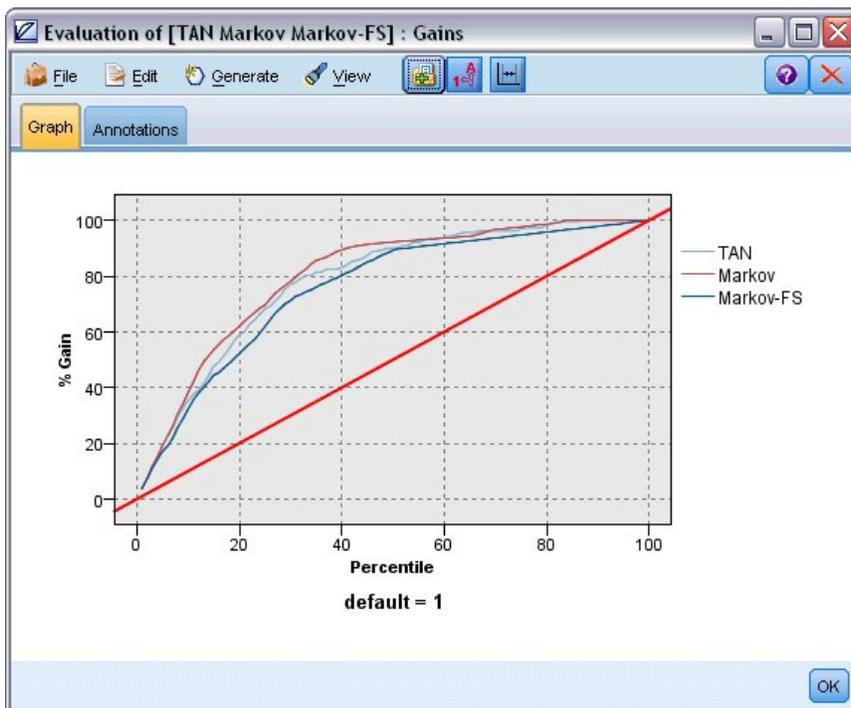


Abbildung 242. Evaluation der Modellgenauigkeit

Um die Vorhersagequalität der einzelnen Modelle zu überprüfen, könnten Sie statt des Evaluierungsdiagramms auch einen Analyseknotten verwenden. Dieser zeigt die Genauigkeit als Prozentsatz an richtigen und falschen Vorhersagen an.

8. Fügen Sie einen Analyseknotten an den Filterknotten an und führen Sie den Analyseknotten mit den zugehörigen Standardeinstellungen aus.

Wie beim Evaluierungsdiagramm 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.

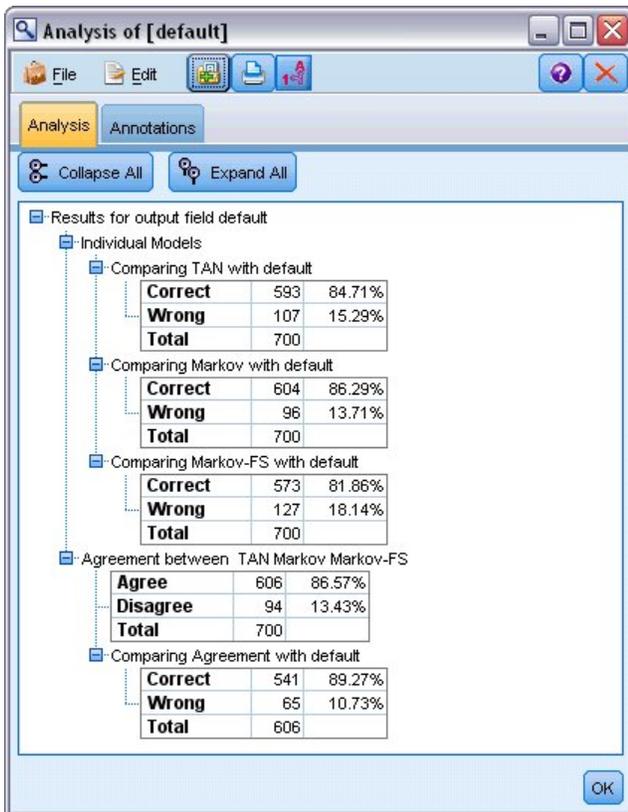


Abbildung 243. Analysieren der Modellgenauigkeit

Erläuterungen der mathematischen Grundlagen für die in IBM SPSS Modeler verwendeten Modellierungsmethoden finden Sie im *IBM 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 Partitionsknotten ein Subset der Datensätze für Test- und Validierungszwecke zurückhalten.

Kapitel 18. Erneutes Trainieren eines Modells auf monatlicher Basis (Bayes-Netz)

Mithilfe des Bayes-Netzknosens 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 auf die Datendateien *telco_Jan.sav* und *telco_Feb.sav* nimmt. 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

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *telco_Jan.sav* im Ordner *Demos* weist.

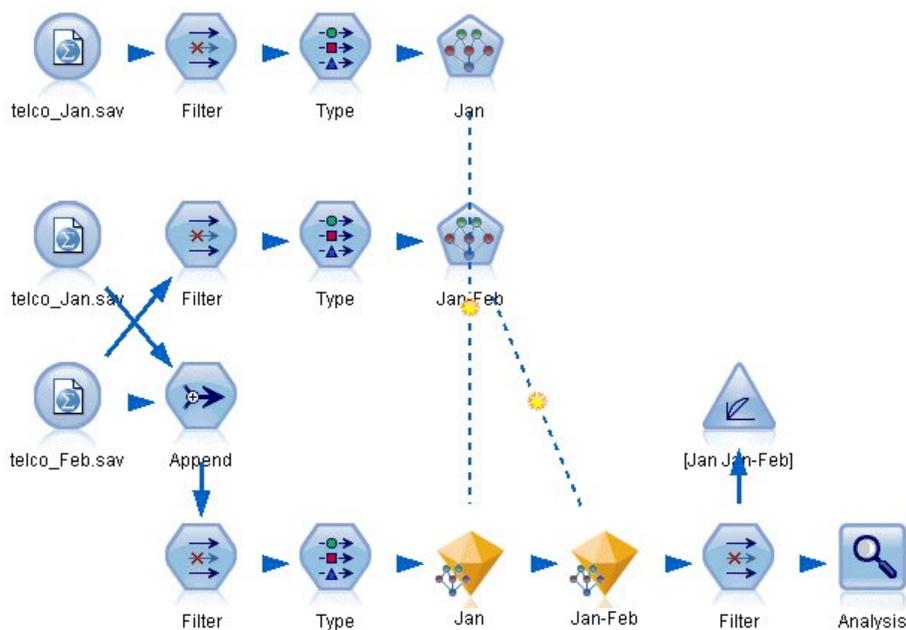


Abbildung 244. Beispielstream für Bayes-Netz

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

2. Fügen Sie dem Quellenknoten einen Filterknoten hinzu.
3. Schließen Sie alle Felder mit Ausnahme von *address*, *age*, *churn*, *custcat*, *ed*, *employ*, *gender*, *marital*, *reside*, *retire* und *tenure* aus.
4. Klicken Sie auf **OK**.

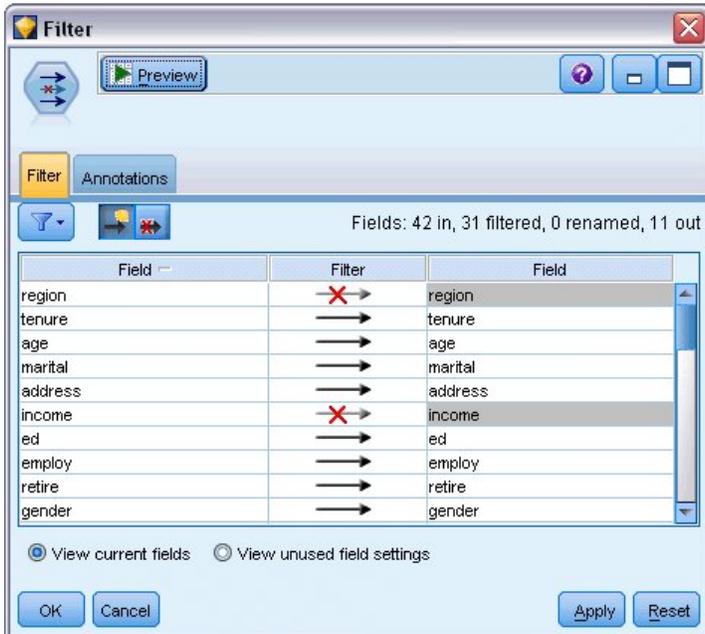


Abbildung 245. Filtern unnötiger Felder

5. Fügen Sie dem Filterknoten einen Typknoten hinzu.
6. Öffnen Sie den Typknoten und klicken Sie auf die Schaltfläche **Werte lesen**, um die Spalte *Werte* auszufüllen.
7. Damit der Evaluierungsknoten 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 246. Auswahl des Zielfelds

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

8. Fügen Sie einen Bayes-Netzknoden an den Typknoden an.
9. Wählen Sie auf der Registerkarte "Modell" als Modellnamen **Benutzerdefiniert** (Angepasst) aus und geben Sie im Textfeld den Ausdruck Jan ein.
10. Wählen Sie als Parameter-Lernmethode **Bayes-Anpassung für kleine Anzahl in den Zellen** aus.
11. Klicken Sie auf **Ausführen**. Das generierte Modellnugget wird zum Stream und zur Modellpalette in der rechten oberen Ecke hinzugefügt.

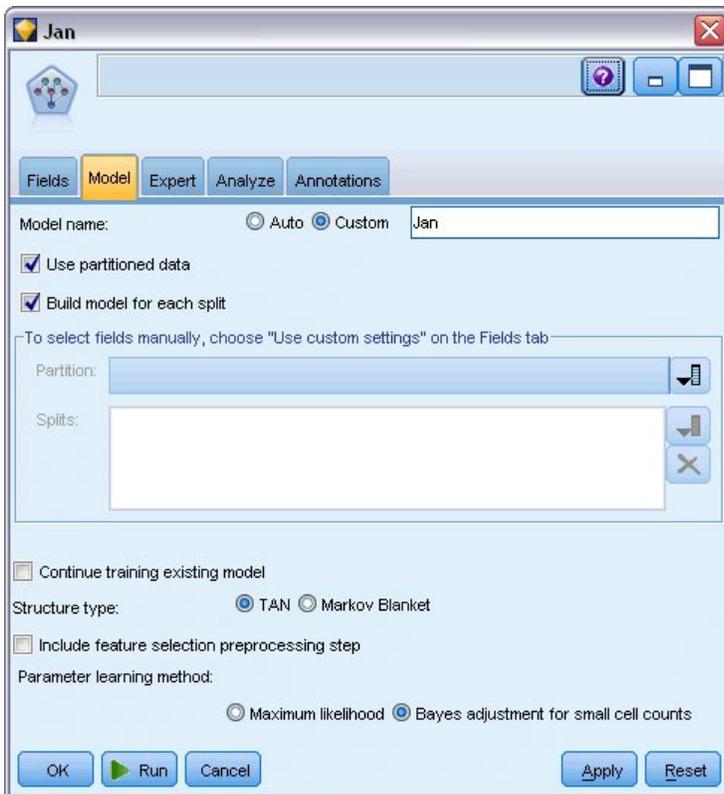


Abbildung 247. Erstellen eines Modells vom Typ "Tree Augmented Naïve Bayes"

12. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *telco_Feb.sav* im Ordner *Demos* verweist.
13. Fügen Sie diesen neuen Quellenknoten an den Filterknoten an (wählen Sie im Warnungsdialogfeld **Ersetzen**, um die Verbindung zum vorherigen Quellenknoten zu ersetzen).

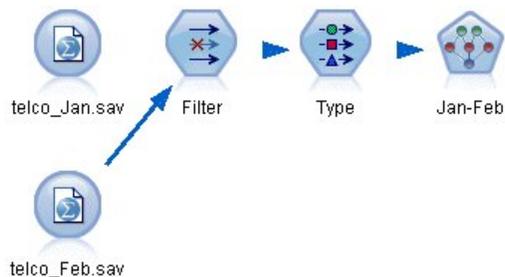


Abbildung 248. Hinzufügen der Daten des zweiten Monats

14. Wählen Sie auf der Registerkarte "Modell" des Bayes-Netzknottens als Modellnamen **Benutzerdefiniert** (Angepasst) aus und geben Sie im Textfeld den Ausdruck **Jan-Feb** ein.
15. Wählen Sie die Option **Training des bestehenden Modells fortsetzen** aus.
16. Klicken Sie auf **Ausführen**. Das generierte Modellnugget überschreibt den bestehenden im Stream, wird jedoch auch zur Modellpalette in der rechten oberen Ecke hinzugefügt.

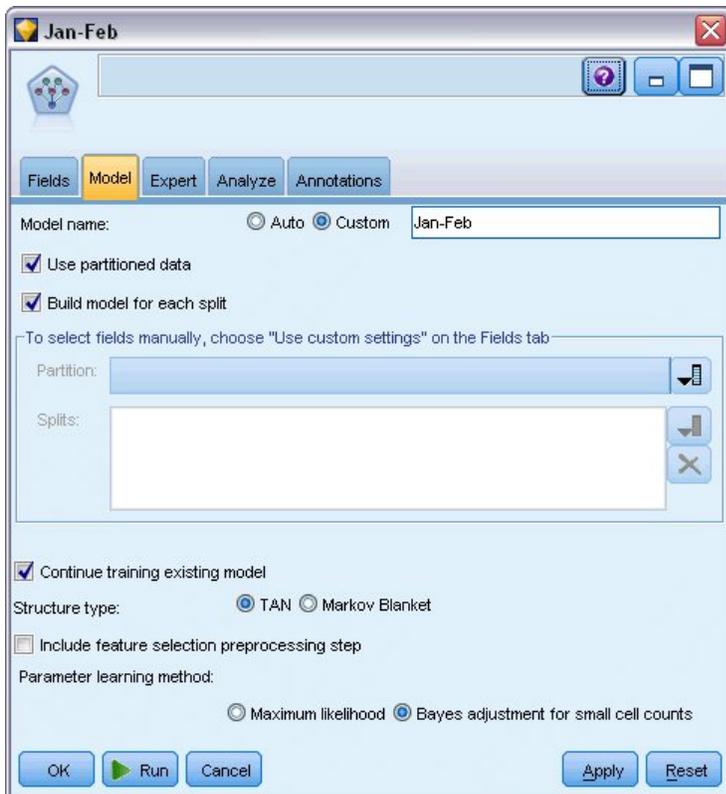


Abbildung 249. Erneutes Trainieren des Modells

Bewerten des Modells

Um die Modelle vergleichen zu können, müssen Sie die beiden Datasets kombinieren.

1. Fügen Sie einen Anhangknoten hinzu und fügen Sie die Quellenknoten *telco_Jan.sav* und *telco_Feb.sav* daran an.



Abbildung 250. Fügen Sie die beiden Datenquellen an

2. Kopieren Sie den Filter- und den Typknoten von weiter oben im Stream und fügen Sie sie in den Streamerstellungsbereich ein.
3. Fügen Sie den Anhangknoten an den neu kopierten Filterknoten an.

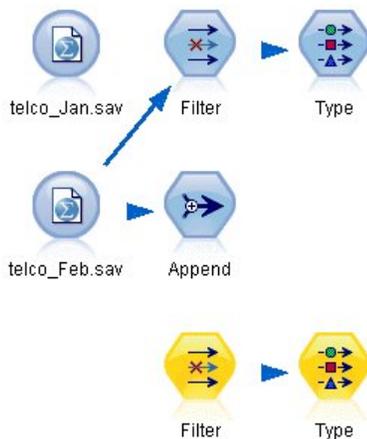


Abbildung 251. Einfügen der kopierten Knoten in den Stream

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

4. Doppelklicken Sie auf das Modellnugget "Jan", um es in den Stream zu übernehmen und an den soeben kopierten Typknoten anzufügen.
5. Fügen Sie das Modellnugget "Jan-Feb", das sich bereits im Stream befindet, an das Modellnugget "Jan" an.
6. Öffnen Sie das Modellnugget "Jan".

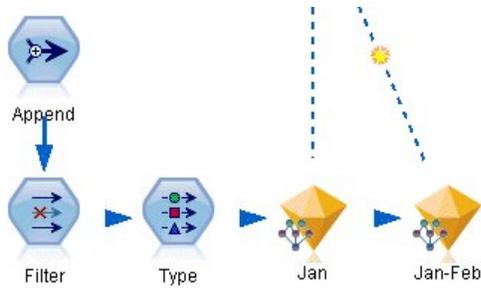


Abbildung 252. Hinzufügen der Nuggets zum Stream

Die Registerkarte "Modell" des Modellnuggets vom Typ "Bayes-Netz" gliedert sich in zwei Spalten: Die linke Spalte enthält ein Netzdiagramm 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 *Bedingten Wahrscheinlichkeiten*, also den Wert der bedingten Wahrscheinlichkeit für die einzelnen Knoten und jede Kombination von Werten in ihren übergeordneten Knoten.

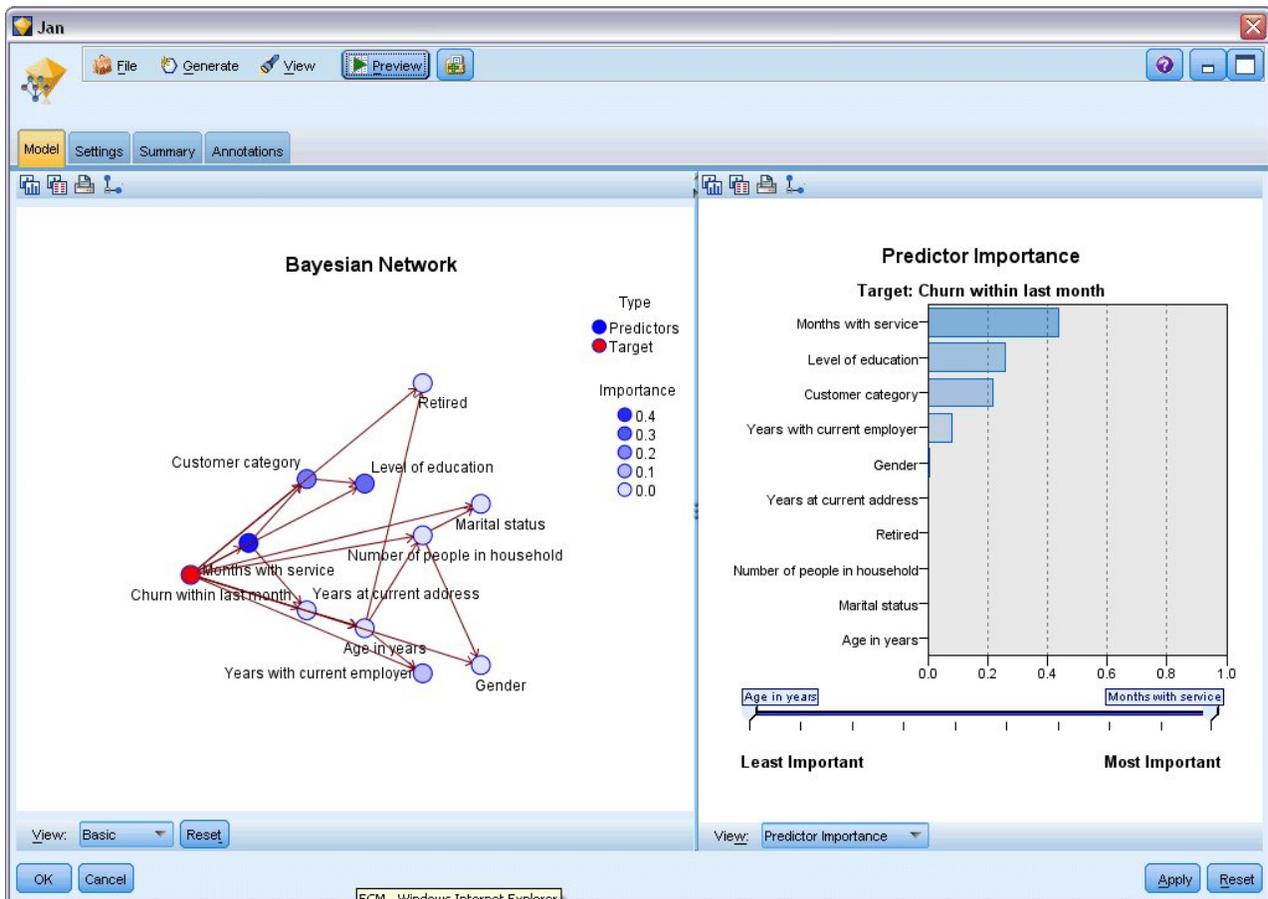


Abbildung 253. Bayes-Netzmodell mit Bedeutsamkeit der Prädiktoren

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 gleichgeordneten Knoten.

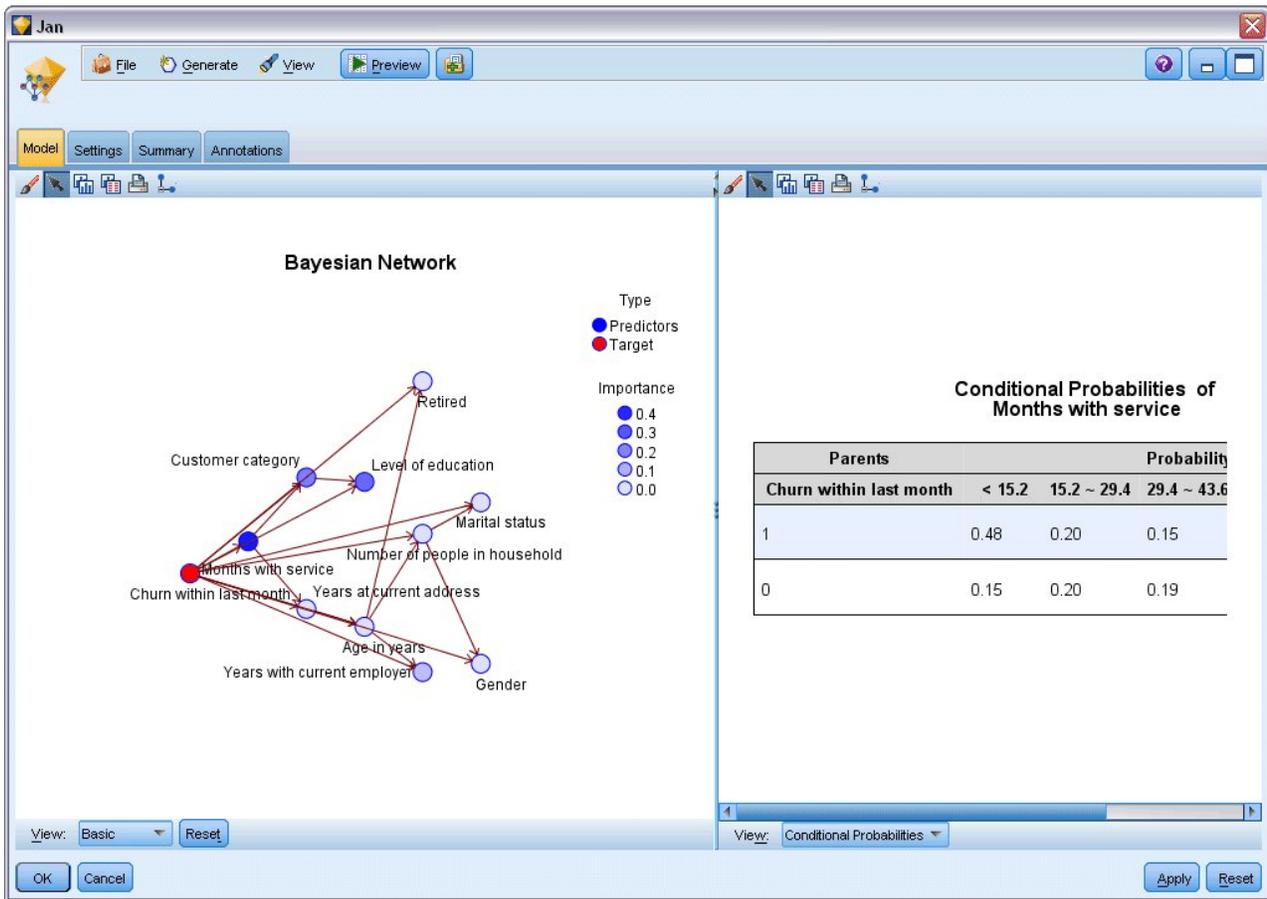


Abbildung 254. Bayes-Netzmodell mit bedingten Wahrscheinlichkeiten

7. Um die Modellausgaben zugunsten größerer Klarheit umzubenennen, müssen Sie einen Filterknoten an das Modellnugget "Jan-Feb" anfügen.
8. Benennen Sie in der rechten Spalte *Feld* "\$B-churn" in "Jan" und "\$B1-churn" in "Jan-Feb" um.

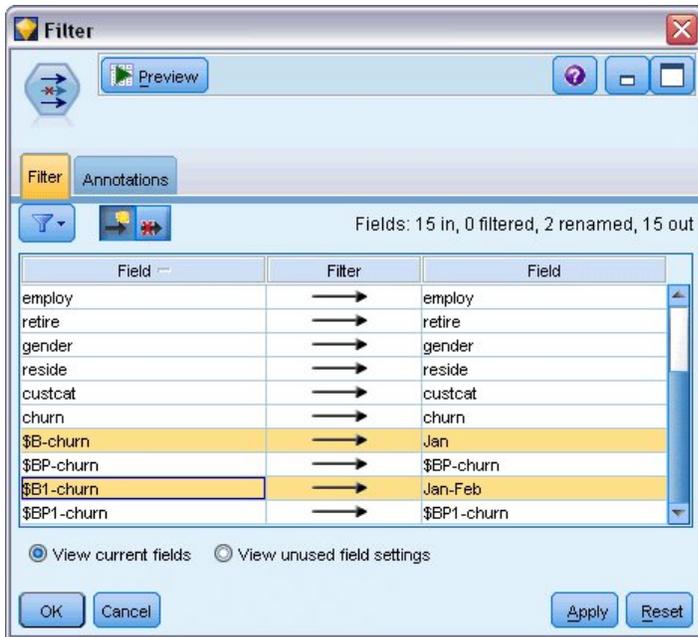


Abbildung 255. Umbenennen von Feldnamen für Modelle

Mit einem Analyseknoden können Sie überprüfen, wie gut die einzelnen Modelle die Abwanderung vorhersagen. Dieser Knoden gibt die Genauigkeit als Prozentsätze an richtigen und falschen Vorhersagen an.

9. Fügen Sie einen Analyseknoden an den Filterknoden an.
10. Öffnen Sie den Analyseknoden und klicken Sie auf **Ausführen**.

Dies zeigt, dass beide Modelle ein ähnliches Maß an Genauigkeit für die Vorhersage der Abwanderung aufweisen.

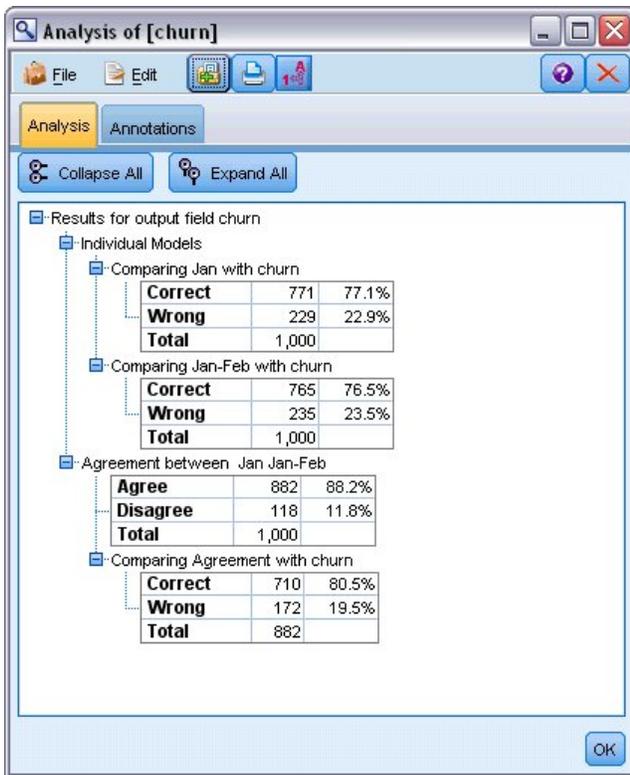


Abbildung 256. Analysieren der Modellgenauigkeit

Als Alternative zum Analyseknoden können Sie mit einem Evaluierungsdiagramm die Vorhersagegenauigkeit der Modelle durch Erstellung eines Gewinnndiagramms vergleichen.

11. Fügen Sie einen Diagrammknoden vom Typ "Evaluation" an den Filterknoden an.

Führen Sie den Diagrammknoden mit all seinen Standardeinstellungen aus.

Wie der Analyseknoden 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.

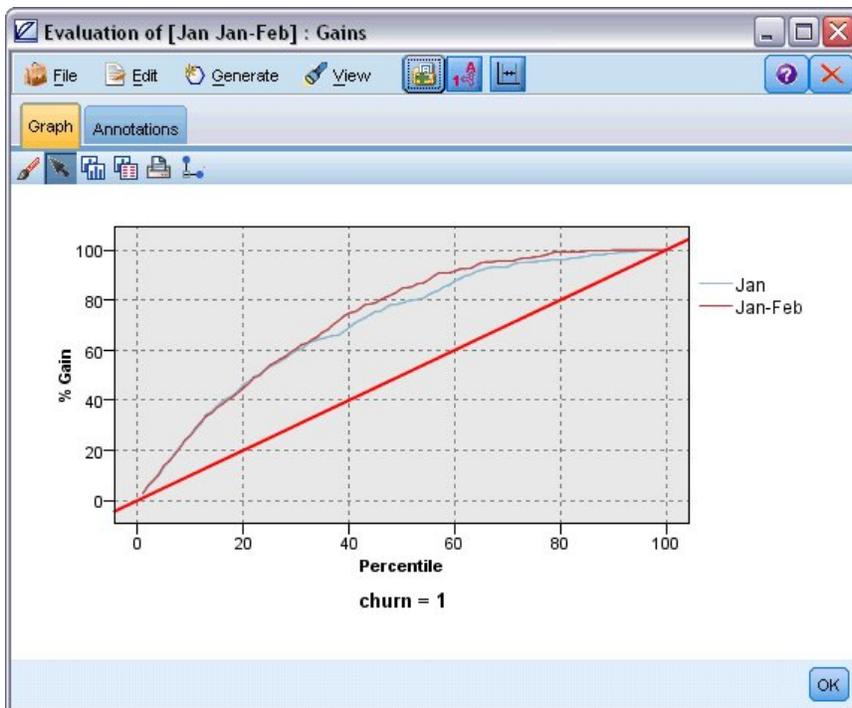


Abbildung 257. Evaluation der Modellgenauigkeit

Erläuterungen der mathematischen Grundlagen für die in IBM SPSS Modeler verwendeten Modellierungsmethoden finden Sie im *IBM 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 ein Subset der Datensätze für Test- und Validierungszwecke zurückhalten.

Kapitel 19. Werbeaktion für Einzelhandelsumsatz (Netz/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. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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.

	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

Abbildung 258. Auswirkungen von Werbeaktionen auf den Umsatz

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.

	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

Abbildung 259. Anstieg der Einkünfte nach der Werbeaktion

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.

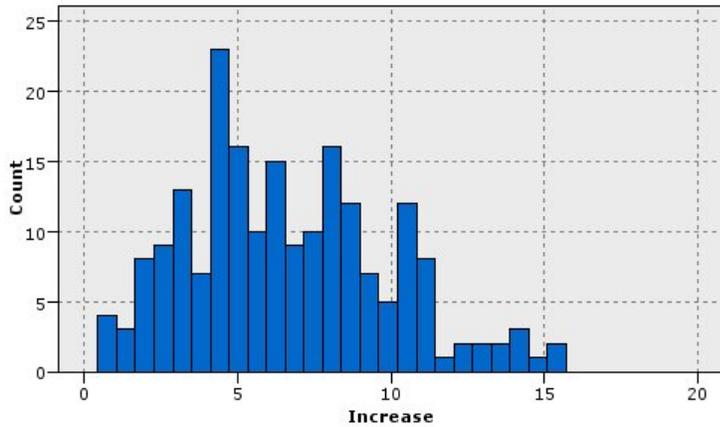


Abbildung 260. Histogramm mit dem Anstieg der Einkünfte

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.

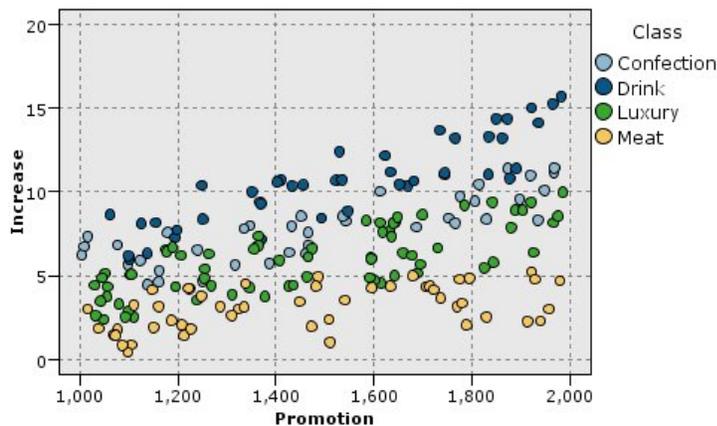


Abbildung 261. Anstieg der Einkünfte im Vergleich zu den Ausgaben für die Werbeaktion

Lernen und Testen

Der Stream `goodslearn.str` trainiert ein neuronales Netz und einen Entscheidungsbaum, diese Vorhersage bzgl. des Anstiegs an Einkünften zu treffen.

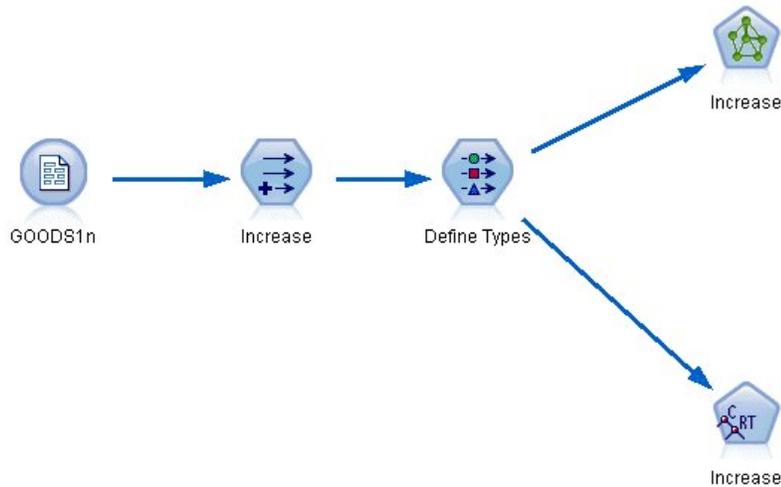


Abbildung 262. Modellierung des Streams "goodslearn.str"

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 Analyseknotten in Reihe, indem Sie die Eingabedatei (Datendatei) in *GOODS2n* ändern und den Analyseknotten ausführen. Anhand der Ausgabe dieses Knottes, 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 20. Bedingungsüberwachung (Netz/C5.0)

Dieses Beispiel betrifft die Überwachungsstatusinformationen eines Systems 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 "Momentaufnahme"-Bericht auf dem System 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 auf die Datendateien *COND1n* und *COND2n* verweisen. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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. Dies ist in der folgenden Tabelle dargestellt:

Zeit	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
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 gegebenenfalls 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 Temperatur- oder Energiezeitreihe sichtbare Muster enthält, können Sie zwischen bevorstehenden Fehlerbedingungen unterscheiden und möglicherweise ihr Auftreten vorher-sagen. Für Temperatur und Energie stellt der nachfolgende Stream die mit den drei verschiedenen Fehler-codes verknüpften Zeitreihen in voneinander getrennten Diagrammen dar, was zu insgesamt sechs Dia-grammen führt. Auswahlknoten trennen die mit den verschiedenen Fehlercodes verknüpften Daten voneinander.

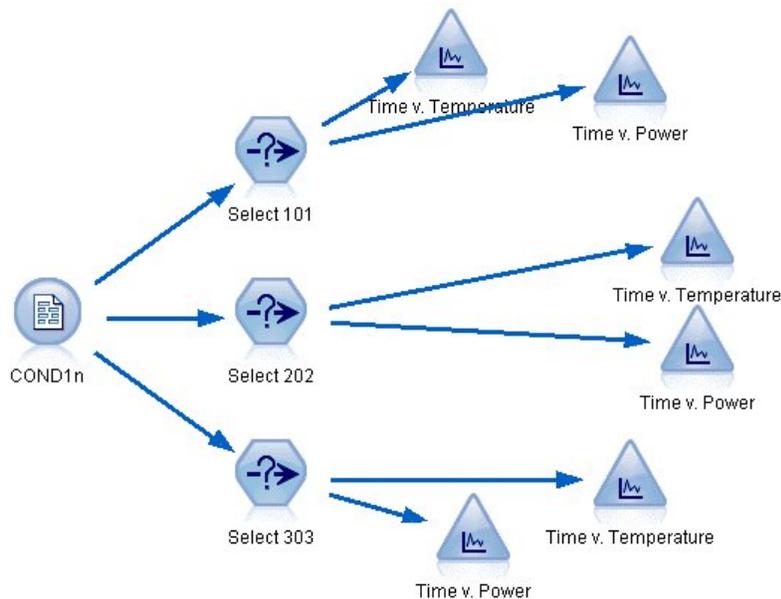


Abbildung 263. Condplot-Stream

Die Ergebnisse dieses Streams werden in dieser Abbildung dargestellt.

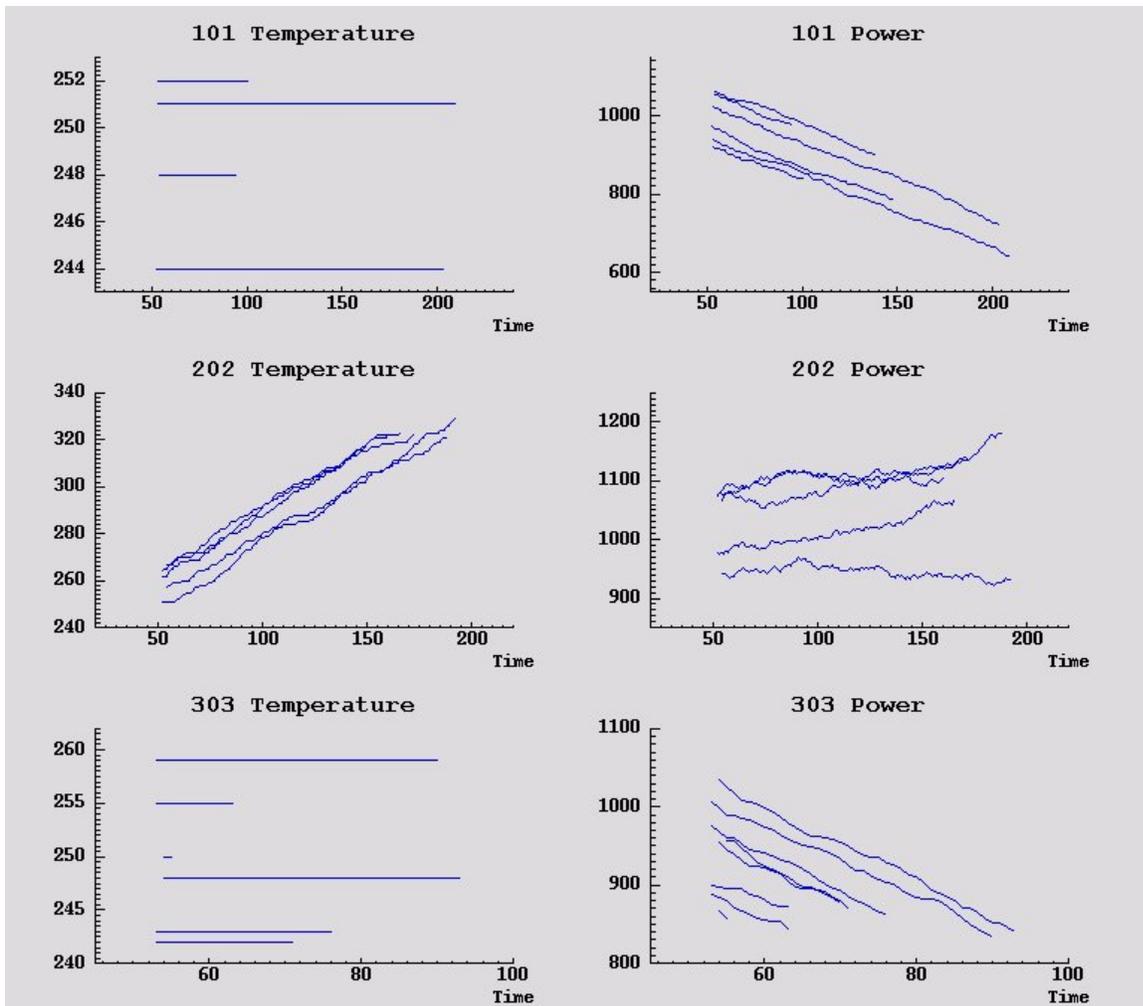


Abbildung 264. Temperatur und Energie im Laufe der Zeit

Die Diagramme zeigen deutlich Muster, die 202-Fehler von 101- und 303-Fehlern unterscheiden. Die 202-Fehler zeigen eine im Laufe der Zeit steigende Temperatur und fluktuierende Energie; 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 bei 303-Fehlern steiler.

Basierend auf diesen Diagrammen ist es offensichtlich, dass das Vorhandensein und die Geschwindigkeit einer Änderung der 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.

Datenvorbereitung

Basierend auf den Ergebnissen einer Datenexploration leitet der Stream *condlearn.str* die relevanten Daten ab und lernt, wie Fehler vorhergesagt werden.

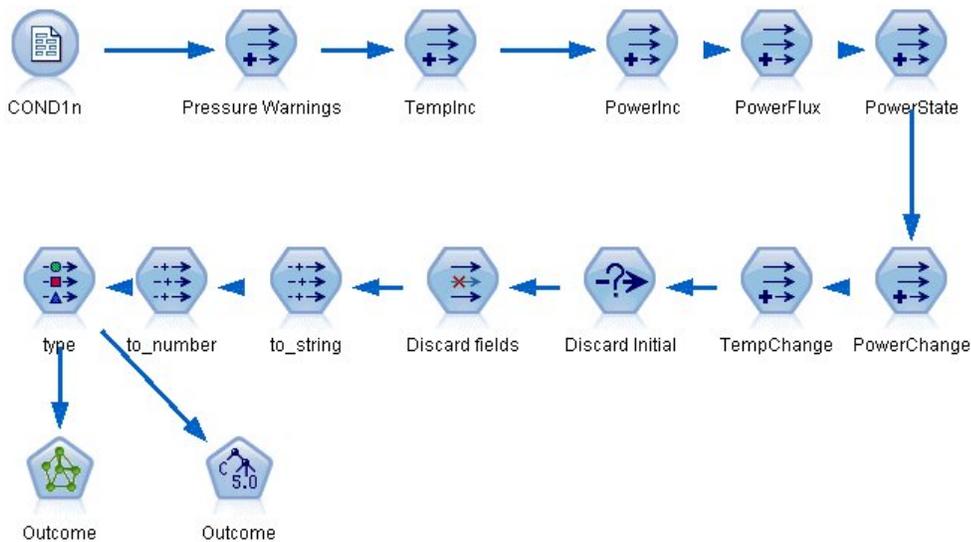


Abbildung 265. Condlearn-Stream

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 im letzten Datensatz und in diesem in entgegengesetzte Richtungen 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 zu *Stabil* zurück, 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 Netz (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: eins stellt das neuronale Netz und eins die Regel dar.

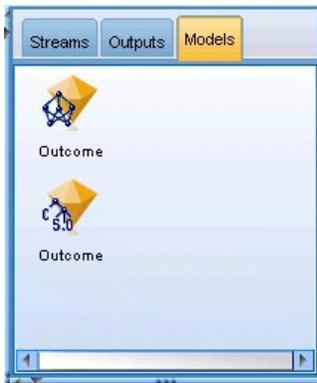


Abbildung 266. Model Manager mit Modellnuggets

Die Modellnuggets werden ebenfalls dem vorhandenen Stream hinzugefügt, sodass Sie das System testen oder die Modellergebnisse exportieren können. In diesem Beispiel testen wir die Ergebnisse des Modells.

Testen

Die Modellnuggets werden in den Stream eingefügt und sind beide mit dem Typknoten verbunden.

1. Positionieren Sie die Nuggets wie abgebildet, sodass der Typknoten an das Netznugget anschließt, das mit dem C5.0-Nugget verbunden ist.
2. Fügen Sie einen Analysekn timer an das C5.0-Nugget an.
3. Bearbeiten Sie den ursprünglichen Quellenknoten, um die Datei *COND2n* (anstelle von *COND1n*) zu lesen, da *COND2n* Testdaten enthält, die noch nicht eingesehen wurden.

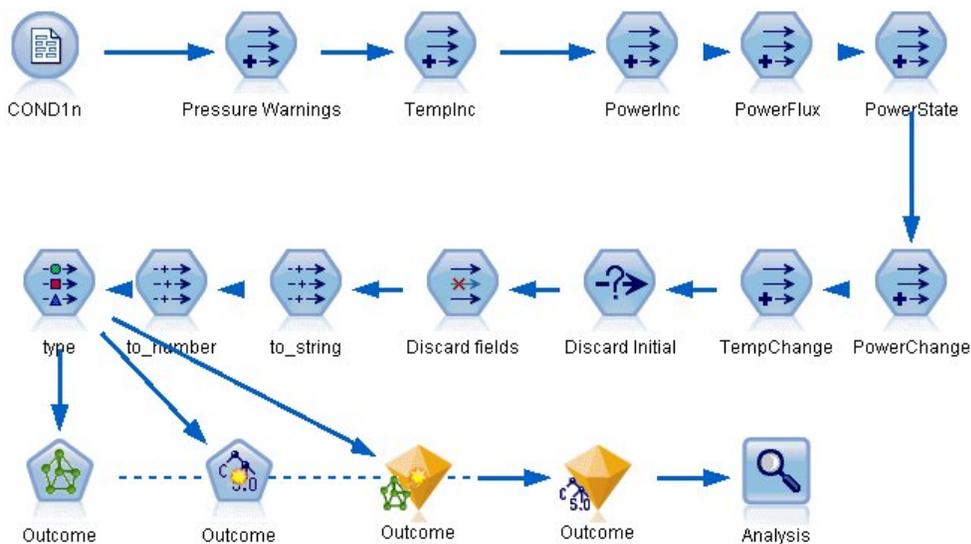


Abbildung 267. Testen des trainierten Netzes

4. Öffnen Sie den Analysekn timer und klicken Sie auf "Ausführen".

Dies führt zu Zahlen, die die Genauigkeit des trainierten Netzes und der trainierten Regel widerspiegeln.

Kapitel 21. Klassifizieren von Kunden im Telekommunikationsbereich (Diskriminanzanalyse)

Die Diskriminanzanalyse ist ein statistisches Verfahren zur Klassifizierung von Datensätzen auf der Grundlage der Werte von Eingabefeldern. Sie entspricht der linearen Regression, außer dass statt eines numerischen Zielfelds ein kategoriales verwendet wird.

Nehmen wir beispielsweise an, dass ein Telekommunikationsanbieter seinen Kundenstamm nach Service-nutzungsmustern 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 auf die Datendatei *telco.sav* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. 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	Beschriftung
1	Basic Service (Basisservice)
2	E-Service
3	Plus Service (Plus-Service)
4	Total Service (Umfassender Service)

Erstellen des Streams

1. Legen Sie zunächst die Streameigenschaften fest, um die Feld- und Wertbeschriftungen in der Ausgabe anzuzeigen. Wählen Sie die folgenden Befehle aus den Menüs aus:
Datei > Streameigenschaften... > Optionen > Allgemein
2. Vergewissern Sie sich, dass die Option **Feld- und Wertbeschriftungen in Ausgabe anzeigen** ausgewählt ist, und klicken Sie auf **OK**.

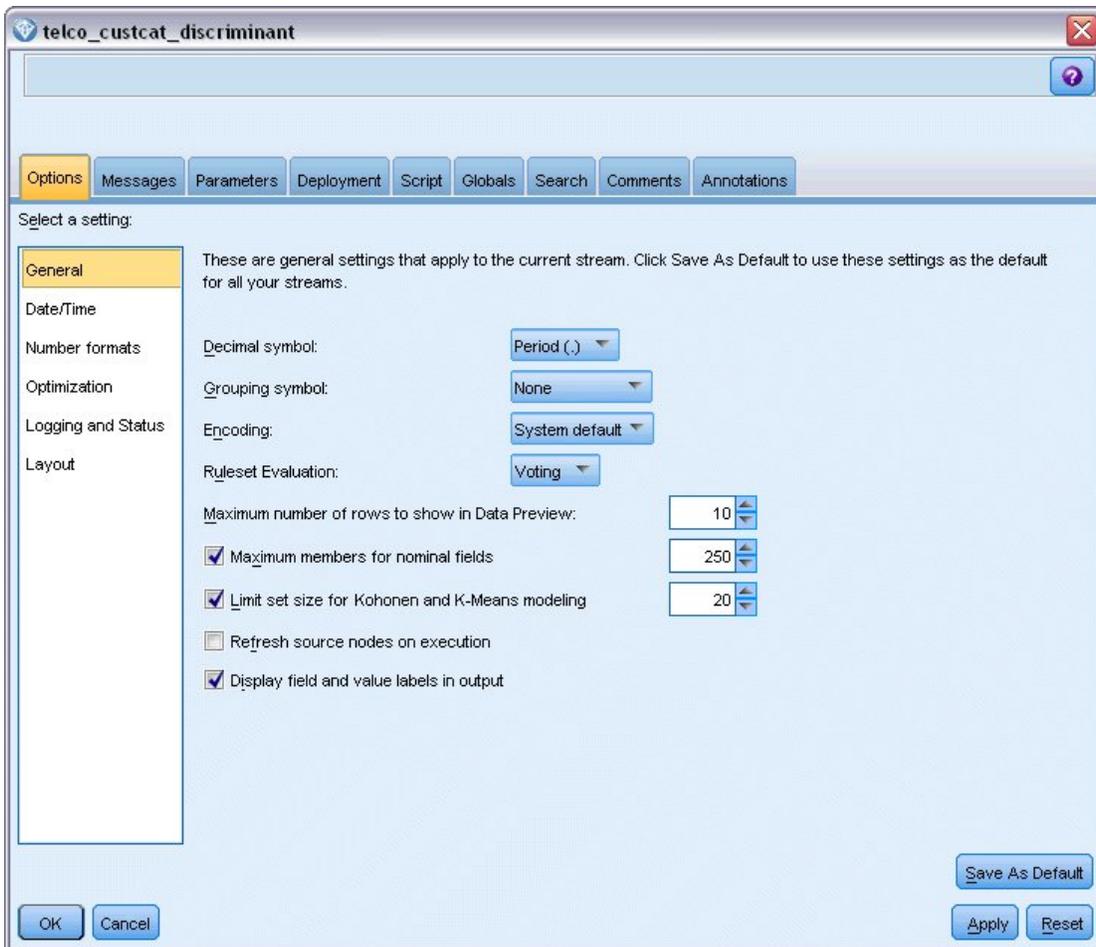


Abbildung 268. Streameigenschaften

3. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

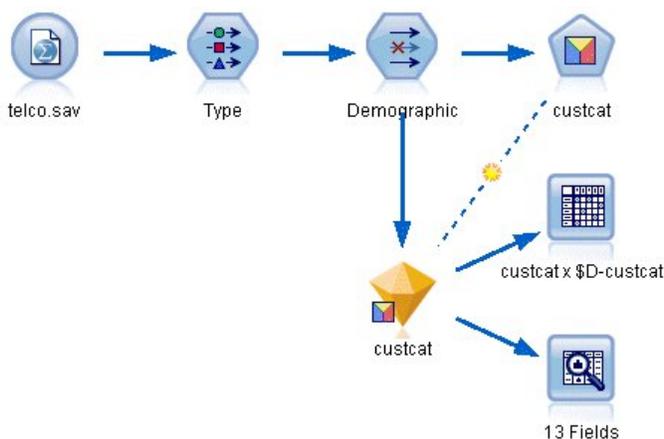


Abbildung 269. Beispielstream zur Klassifizierung von Kunden mithilfe der Diskriminanzanalyse

- a. 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.

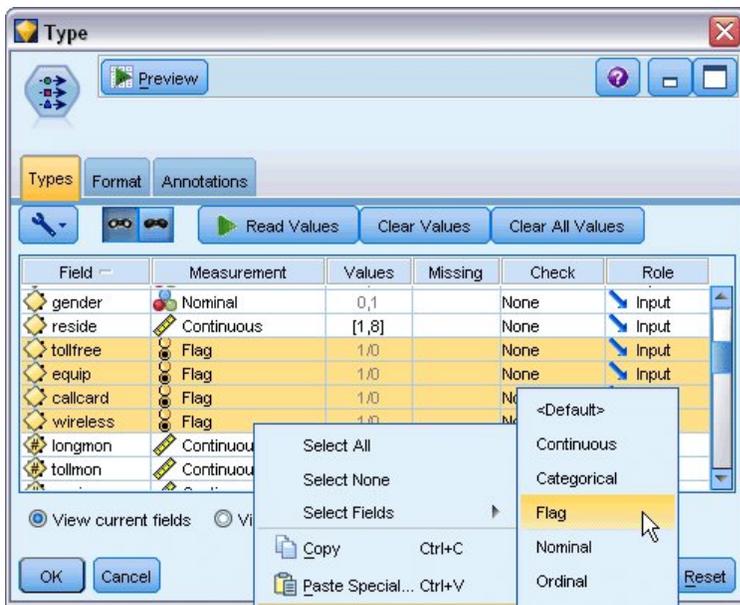


Abbildung 270. Festlegen des Messniveaus für mehrere Felder

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. Übernehmen Sie also **Nominal** für sein Messniveau.

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

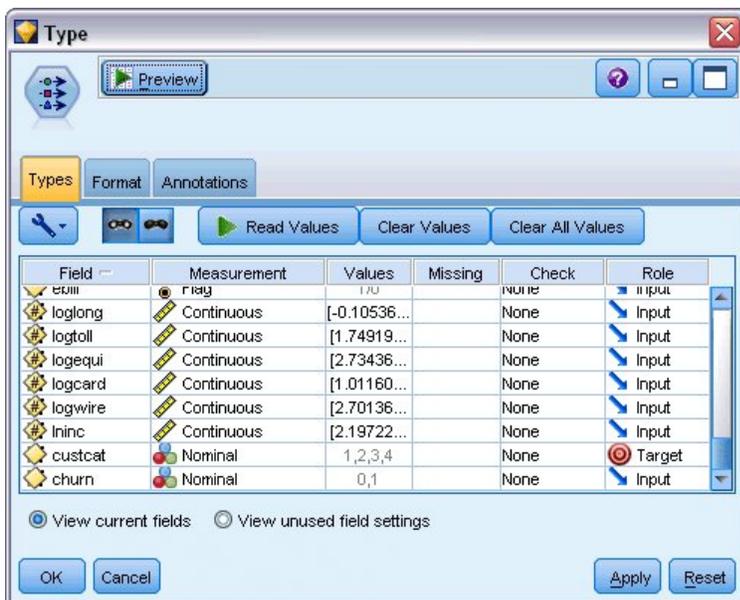


Abbildung 271. Festlegen der Feldrolle

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.

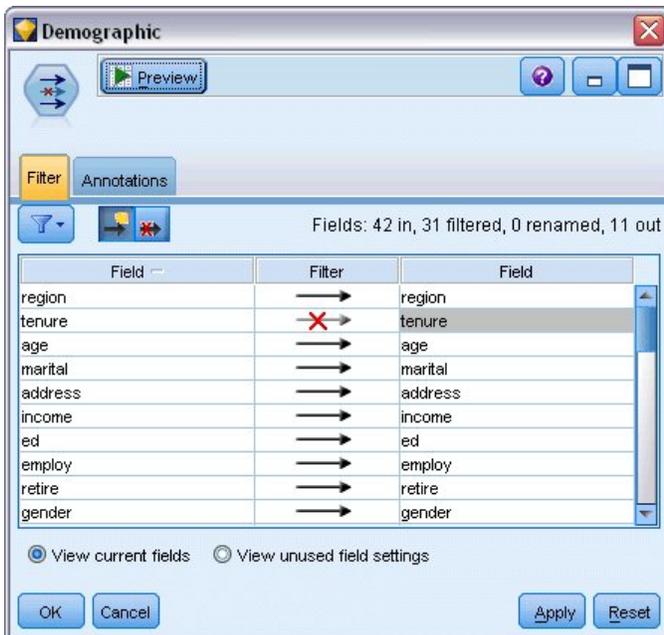


Abbildung 272. Filtern nach demografischen Feldern

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

4. Klicken Sie im Diskriminanzknoten auf die Registerkarte "Modell" und wählen Sie die Methode **Schrittweise** aus.

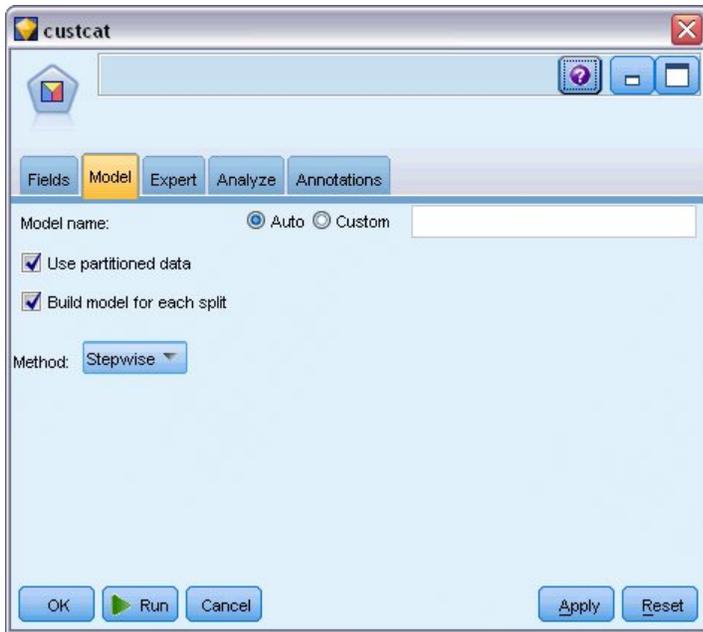


Abbildung 273. Auswählen der Modelloptionen

5. Setzen Sie auf der Registerkarte "Experten" den Modus auf **Experten** und klicken Sie auf **Ausgabe**.
6. Wählen Sie auf der Registerkarte "Erweiterte Ausgabe" die Optionen **Zusammenfassung**, **Territorien** und **Zusammenfassung der Schritte** aus und klicken Sie auf **OK**.

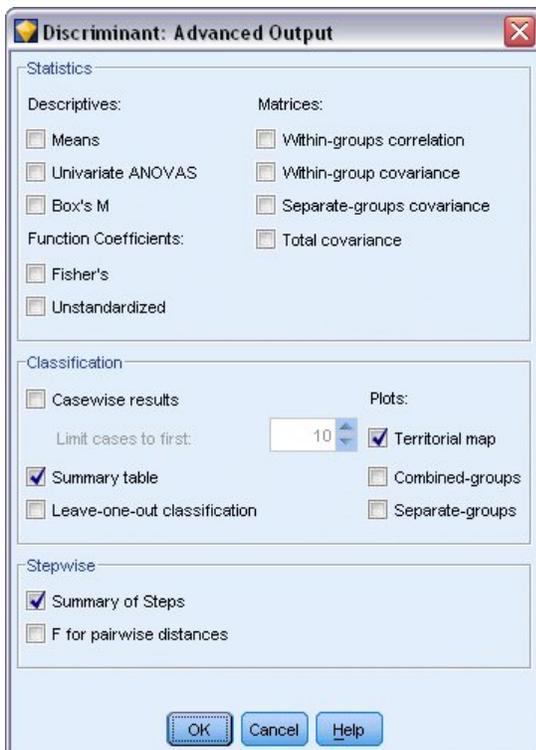


Abbildung 274. Auswahl der Ausgabeoptionen

Untersuchen des Modells

1. 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 Modellnugget im Stream.

Auf der Registerkarte "Übersicht" werden (unter anderem) die Ziele und die vollständige Liste der Eingaben (Prädiktorfelder) angezeigt, die zur Erwägung vorgelegt wurden.

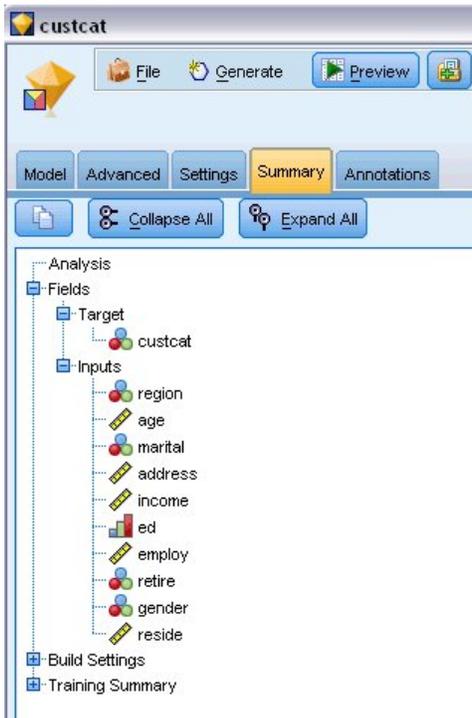


Abbildung 275. Modellübersicht mit Ziel- und Eingabefeldern

So erhalten Sie Einzelheiten zu den Ergebnissen der Diskriminanzanalyse:

2. Klicken Sie auf die Registerkarte "Erweitert".
3. Klicken Sie auf die Schaltfläche "In externem Browser starten" (direkt unter der Registerkarte "Modelle"), um die Ergebnisse in Ihrem Web-Browser anzuzeigen.

Ausgabeanalyse für die Verwendung von Diskriminanzanalysen hinsichtlich der Klassifizierung von Telekommunikationskunden

Schrittweise Diskriminanzanalyse

Step		Tolerance	Min. Tolerance	F to Enter	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
	Retired	1.000	1.000	3.005	.991
	Years with current employer	1.000	1.000	16.976	.951
	Gender	1.000	1.000	.373	.999
	Number of people in household	1.000	1.000	3.976	.988

Abbildung 276. Nicht in die Analyse eingeschlossene Variablen, Schritt 0

Wenn eine große Anzahl von Prädiktoren (Einflussvariablen) 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 Prädiktoren enthält. Bei jedem Schritt wird der Prädiktor mit dem größten *F-Wert für Aufnahme*, der die Eintragskriterien überschreitet (standardmäßig 3,84), dem Modell hinzugefügt.

Step		Tolerance	Min. Tolerance	F to Enter	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

Abbildung 277. Nicht in die Analyse eingeschlossene Variablen, Schritt 3

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.

Step		Tolerance	F to Remove	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

Abbildung 278. Variablen in der Analyse

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 an-

dere 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 Prädiktoren haben daher möglicherweise keine *praktische Bedeutung*. Wenn Sie Erfahrungen mit den Daten haben und in etwa wissen, welche Prädiktoren wichtig sind, sollten Sie diese Kenntnisse nutzen und keine schrittweisen Methoden verwenden. Wenn hingegen viele Prädiktoren 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

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.198	80.2	80.2	.407
2	.048	19.4	99.6	.214
3	.001	.4	100.0	.031

Abbildung 279. Eigenwerte

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.

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.796	227.345	9	.000
2 through 3	.953	47.486	4	.000
3	.999	.929	1	.335

Abbildung 280. Wilks-Lambda

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

	Function		
	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*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions. Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

Abbildung 281. Strukturmatrix

Wenn mehr als eine Diskriminanzfunktion vorhanden ist, markiert ein Stern (*) 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

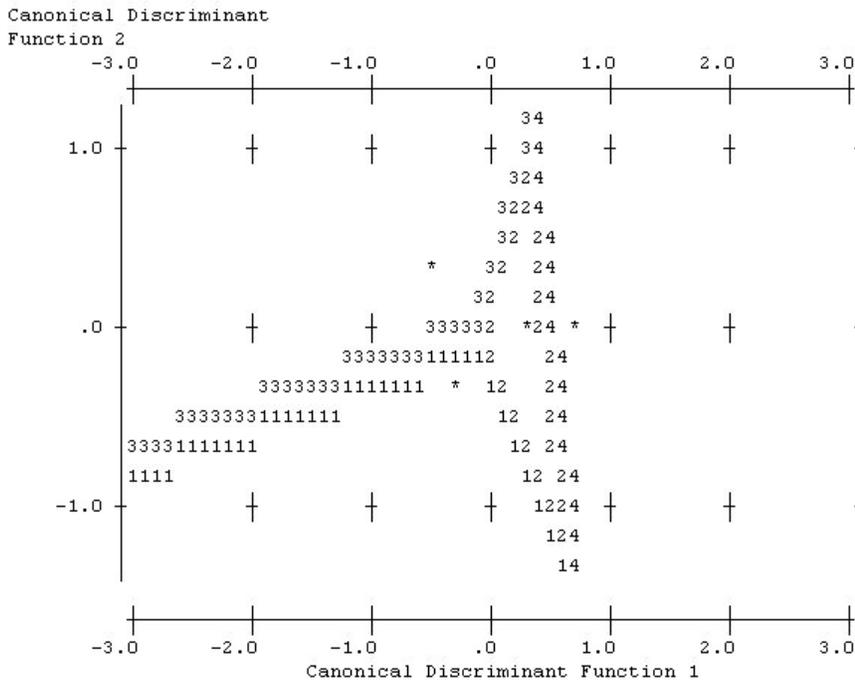


Abbildung 282. Territorien

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 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 Stern (*) 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

		Customer category	Predicted Group Membership				Total
			Basic service	E-service	Plus service	Total service	
Original	Count	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% of original grouped cases correctly classified.

Abbildung 283. Klassifikationsergebnisse

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 Prädiktoren) alle Kunden in die Gruppe, die dem Modalwert entspricht, 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 einen anderen Prädiktor 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 Kundenstamms 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 Ihres Return-on-Investment durch *Plus-Service*- oder *Gesamt-service*-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 ein Subset der Datensätze für Test- und Validierungszwecke zurückhalten.

Erläuterungen der mathematischen Grundlagen der in IBM SPSS Modeler verwendeten Modellierungsverfahren sind im IBM SPSS Modeler-Algorithmushandbuch aufgeführt. Es steht im Verzeichnis *Documentation* des Installationsdatenträgers zur Verfügung.

Kapitel 22. Analysieren von intervallzensierten Überlebensdaten (verallgemeinerte lineare Modelle)

Wenn die Analyse von Überlebensdaten mit Intervallzensierung vorgenommen wird, das heißt, 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 Ereignis-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 Dataset wurde an anderer Stelle vorgestellt und analysiert¹. 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 Streamdatei im Unterordner *streams*.

Erstellen des Modells

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *ulcer_recurrence.sav* im Ordner *Demos* verweist.

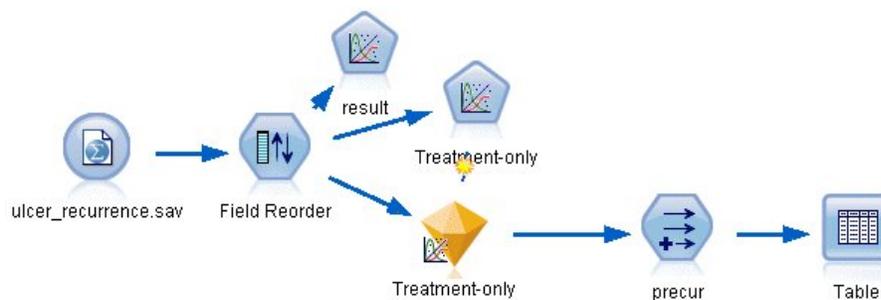


Abbildung 284. Beispielstream zur Vorhersage des erneuten Auftretens von Geschwüren

2. Filtern Sie auf der Registerkarte "Filter" des Quellenknotens jeweils *id* und *time* heraus.

1. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

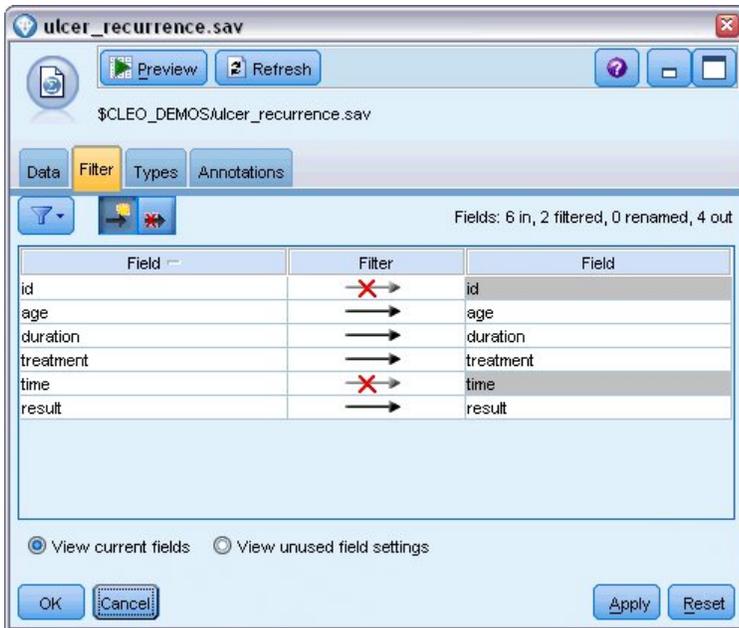


Abbildung 285. Filtern von unerwünschten Feldern

3. 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.
4. Klicken Sie auf **Werte lesen**, um die Daten zu instanziierten.



Abbildung 286. Festlegen der Feldrolle

5. 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 Collett-Ergebnisse leichter reproduzieren.



Abbildung 287. Neuordnung von Feldern für die gewünschte Eingabe in das Modell

6. Fügen Sie dem Quellenknoten einen GenLin-Knoten hinzu. Klicken Sie im GenLin-Knoten auf die Registerkarte **Modell**.
7. Wählen Sie **Erste (niedrigster Wert)** als Referenzkategorie für das Ziel aus. Dies gibt an, dass die zweite Kategorie das relevante Ereignis ist, und ihr Effekt auf das Modell liegt in der Interpretation der Parameterschätzungen. 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.

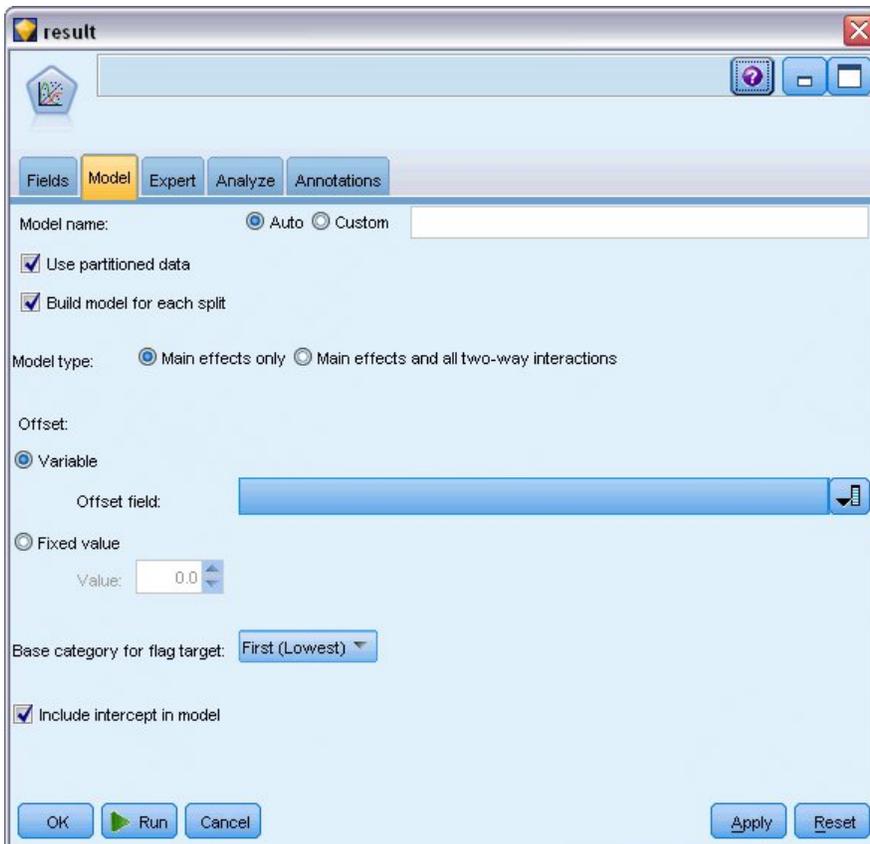


Abbildung 288. Auswählen der Modelloptionen

8. Klicken Sie auf die Registerkarte **Experten** und wählen Sie **Experten** aus, um die Expertenmodellierungsoptionen zu aktivieren.
9. Wählen Sie **Binomial** als Verteilung und **Log-Log komplementär** als Verknüpfungsfunktion (Linkfunktion) aus.
10. Wählen Sie **Fester Wert** als Methode zur Schätzung des Skalenparameters aus und behalten Sie den Standardwert 1,0 bei.
11. Wählen Sie **Absteigend** als Reihenfolge der Kategorien für Faktoren aus. 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 Parameterschätzungen.

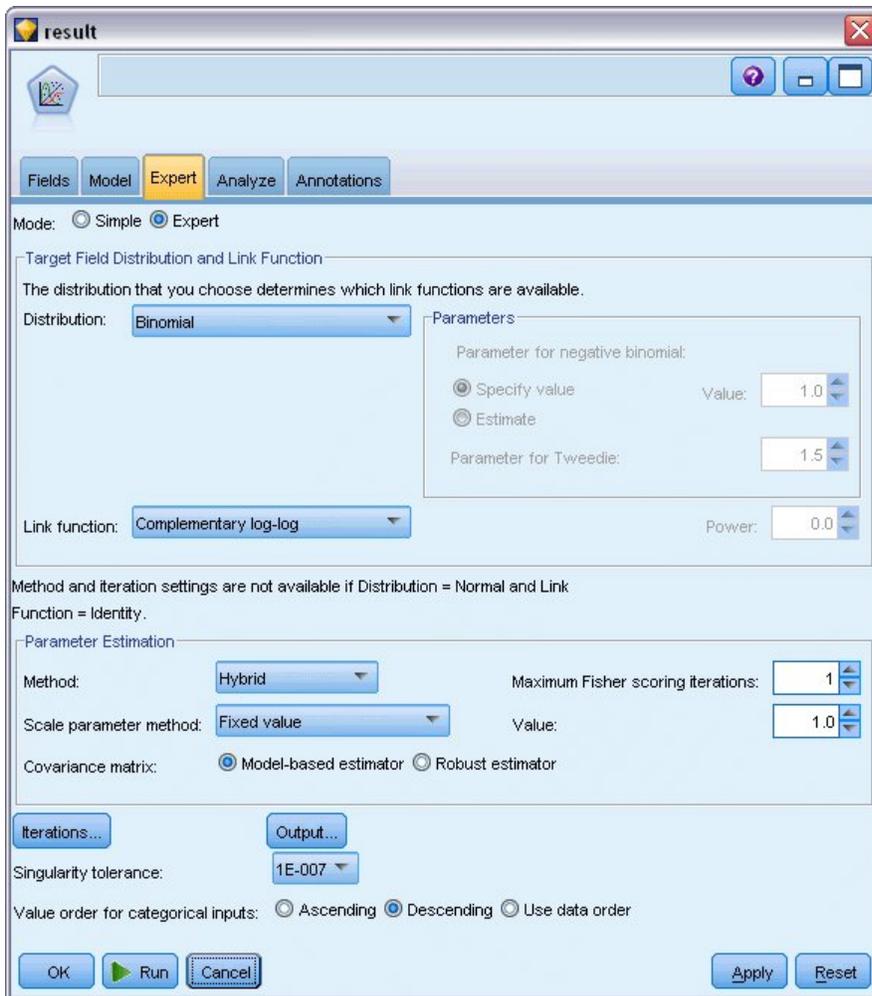


Abbildung 289. Auswählen von Expertenoptionen

- Führen Sie den Stream aus, um das Modellnugget zu generieren; dieses wird dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die Modelldetails anzuzeigen, können Sie mit der rechten Maustaste auf das Nugget klicken und **Bearbeiten** oder **Durchsuchen** auswählen.

Tests der Modelleffekte

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
duration	.003	1	.958
treatment	.382	1	.537
age	.358	1	.550

Dependent Variable: Result
Model: (Intercept), duration, treatment, age

Abbildung 290. Tests der Modelleffekte für das Haupteffektmodell

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"

1. Klicken Sie im GenLin-Knoten auf der Registerkarte "Felder" auf **Benutzerdefinierte Einstellungen verwenden**.
2. Wählen Sie *result* als Ziel aus.
3. Wählen Sie *treatment* als einzige Eingabe aus.

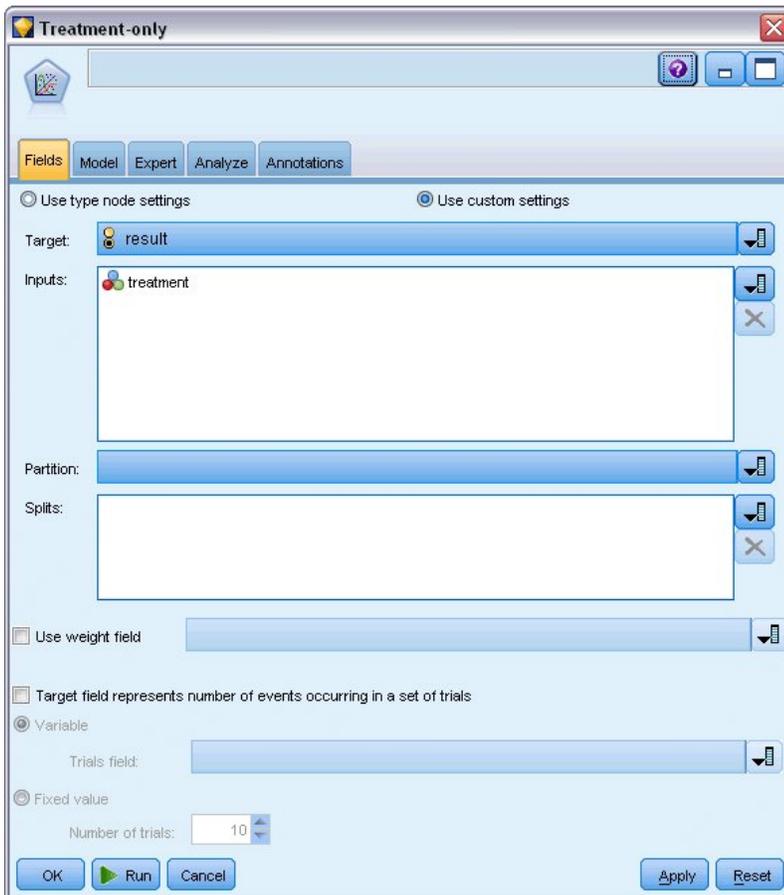


Abbildung 291. Auswählen von Feldoptionen

4. Führen Sie den Stream aus und öffnen Sie das resultierende Modellnugget.

Wählen Sie auf dem Modellnugget die Registerkarte **Erweitert** aus und scrollen Sie zum Ende.

Parameterschätzungen

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.442	.5012	-2.425	-.460	8.282	1	.004
[treatment=1]	.378	.6288	-.855	1.610	.361	1	.548
[treatment=0]	0 ^a
(Scale)	1 ^b

Dependent Variable: Result
Model: (Intercept), treatment

- a. Set to zero because this parameter is redundant.
- b. Fixed at the displayed value.

Abbildung 292. Parameterschätzungen für das Modell "Nur Behandlung"

Der Behandlungseffekt (der Unterschied der linearen Prädiktoren 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 die Parameterschätzung 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. Der lineare Prädiktor (Anfangs- und Behandlungseffekt) ist eine Schätzung von $\text{Log}(-\log(1-P(\text{recur}_{12,t})))$, wobei $P(\text{recur}_{12,t})$ 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 293. Einstellungsoptionen für Ableitungsknoten

1. 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.
2. Geben Sie auf der Registerkarte "Einstellungen" precur als Ableitungsfeld ein.
3. Wählen Sie für die Ableitung des Felds die Option **Bedingt** aus.
4. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder für die Bedingung **Wenn** zu öffnen.

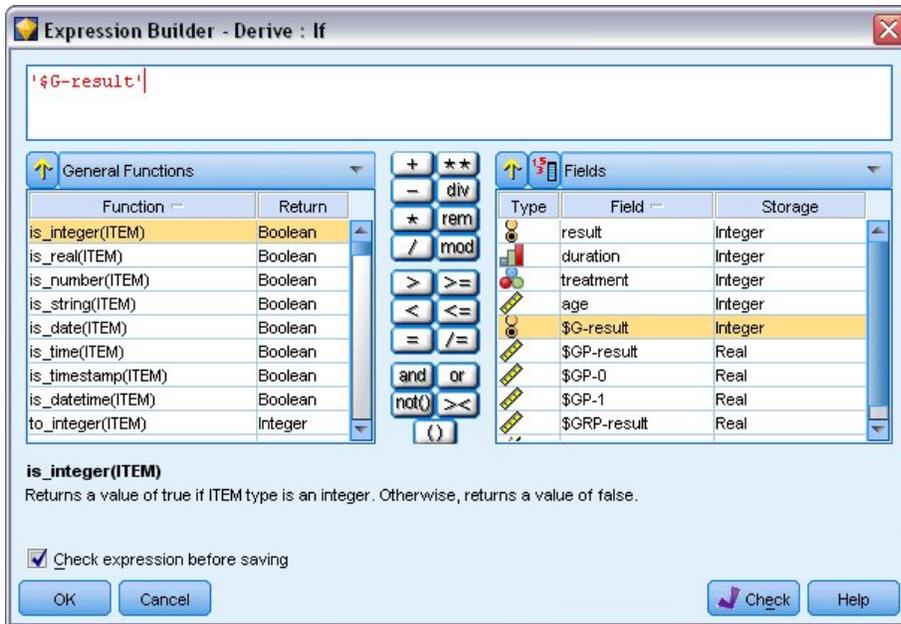


Abbildung 294. Ableitungsknoten: Expression Builder für Bedingung "Wenn"

5. Fügen Sie das Feld *\$G-result* in den Ausdruck ein.
6. 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.

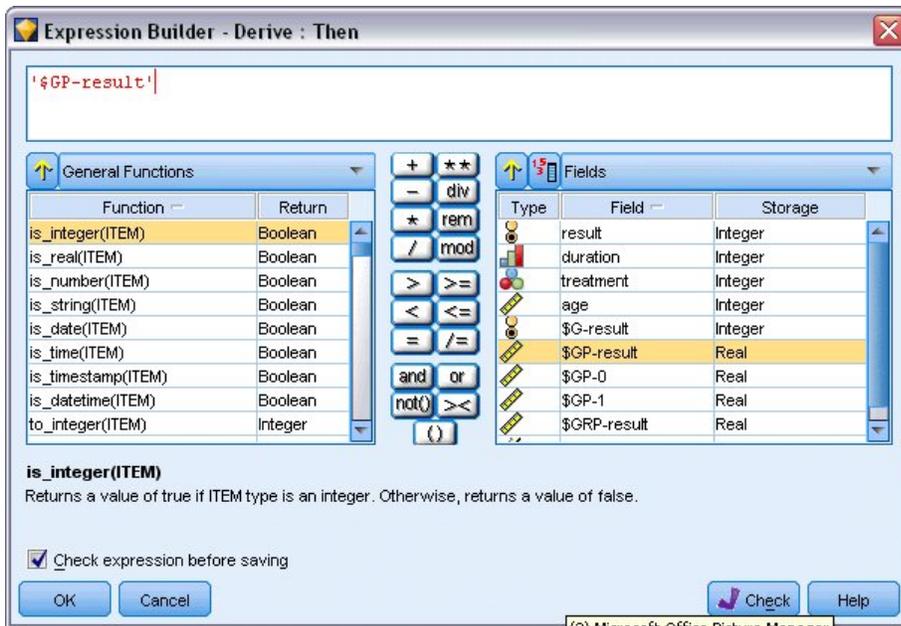


Abbildung 295. Ableitungsknoten: Expression Builder für "Dann"-Ausdruck

7. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder für den **Dann**-Ausdruck zu öffnen.
8. Fügen Sie das Feld *\$GP-result* in den Ausdruck ein.

9. Klicken Sie auf **OK**.

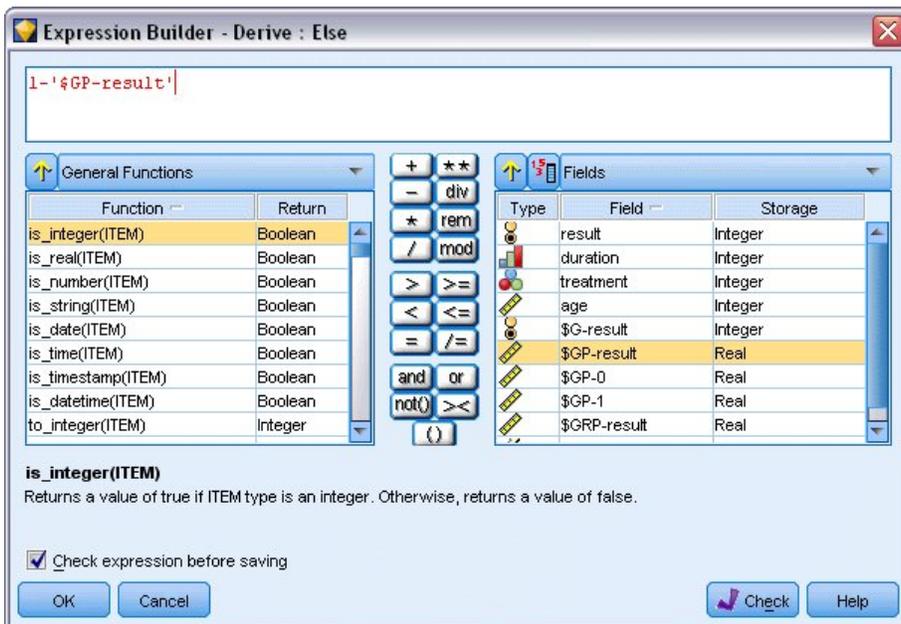


Abbildung 296. Ableitungsknoten: Expression Builder für "Sonst"-Ausdruck

10. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder für den **Sonst**-Ausdruck zu öffnen.
11. Geben Sie 1- in den Ausdruck ein und fügen Sie anschließend das Feld *\$GP-result* in den Ausdruck ein.
12. Klicken Sie auf **OK**.



Abbildung 297. Einstellungsoptionen für Ableitungsknoten

13. Fügen Sie einen Tabellenknoten an den Ableitungsknoten an und führen Sie ihn aus.

	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

Abbildung 298. Geschätzte Wahrscheinlichkeiten

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 $1 - P(\text{recur}_{12, i})$ 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 Datensets erforderlich. Dieses Dataset finden Sie in der Datei *ulcer_recurrence_recoded.sav*. 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 Dataset bei.

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *ulcer_recurrence_recoded.sav* im Ordner *Demos* verweist.

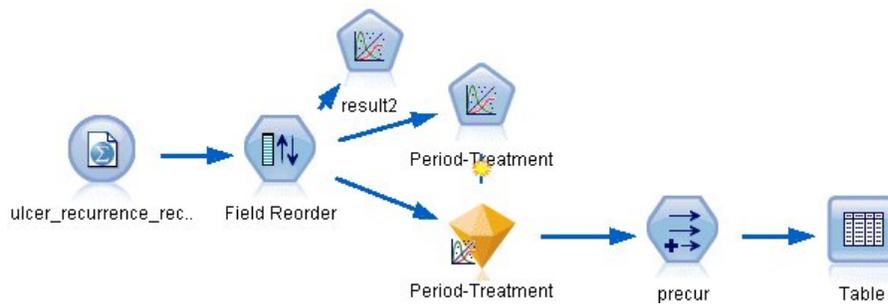


Abbildung 299. Beispielstream zur Vorhersage des erneuten Auftretens von Geschwüren

2. Filtern Sie auf der Registerkarte "Filter" des Quellenknotens jeweils *id*, *time* und *result* heraus.

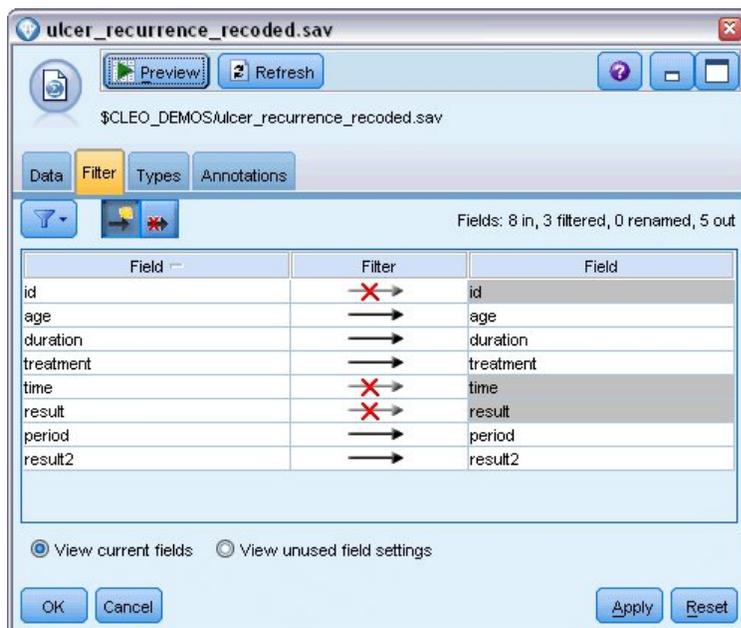


Abbildung 300. Filtern von unerwünschten Feldern

3. 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 301. Festlegen der Feldrolle

- Fügen 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 302. Neuordnung von Feldern für die gewünschte Eingabe in das Modell

- Klicken Sie im GenLin-Knoten auf die Registerkarte **Modell**.

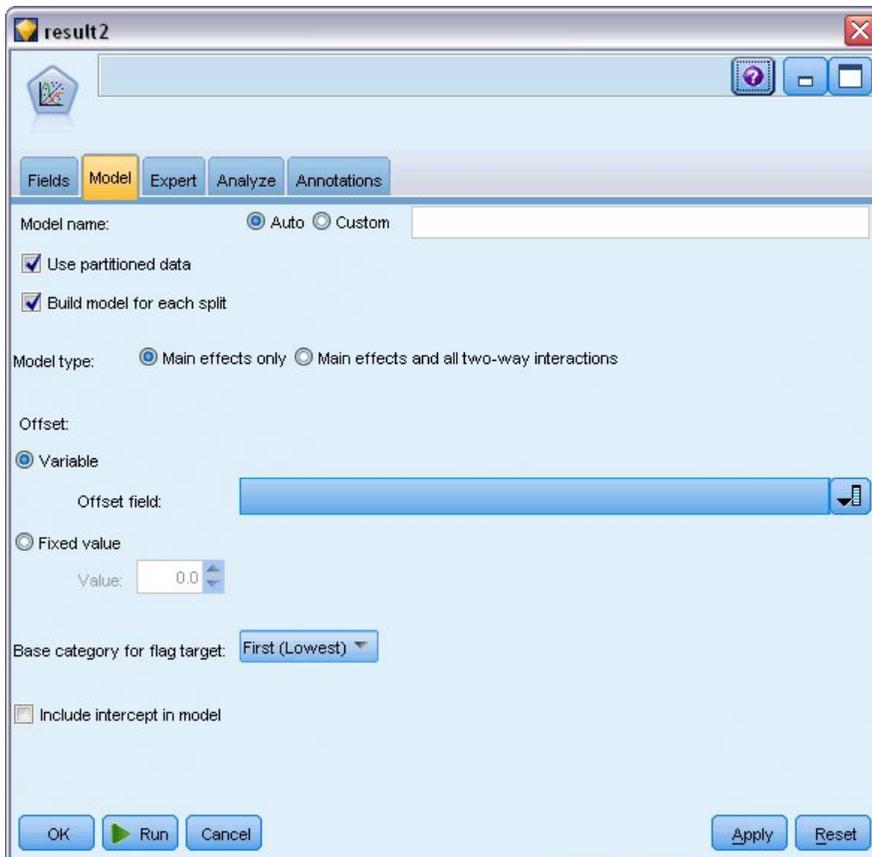


Abbildung 303. Auswählen der Modelloptionen

6. Wählen Sie **Erste (niedrigster Wert)** als Referenzkategorie für das Ziel aus. Dies gibt an, dass die zweite Kategorie das relevante Ereignis ist, und ihr Effekt auf das Modell liegt in der Interpretation der Parameterschätzungen.
7. Inaktivieren Sie die Option **Konstanten Term in Modell einschließen**.
8. Klicken Sie auf die Registerkarte **Experten** und wählen Sie **Experten** aus, um die Expertenmodellierungsoptionen zu aktivieren.

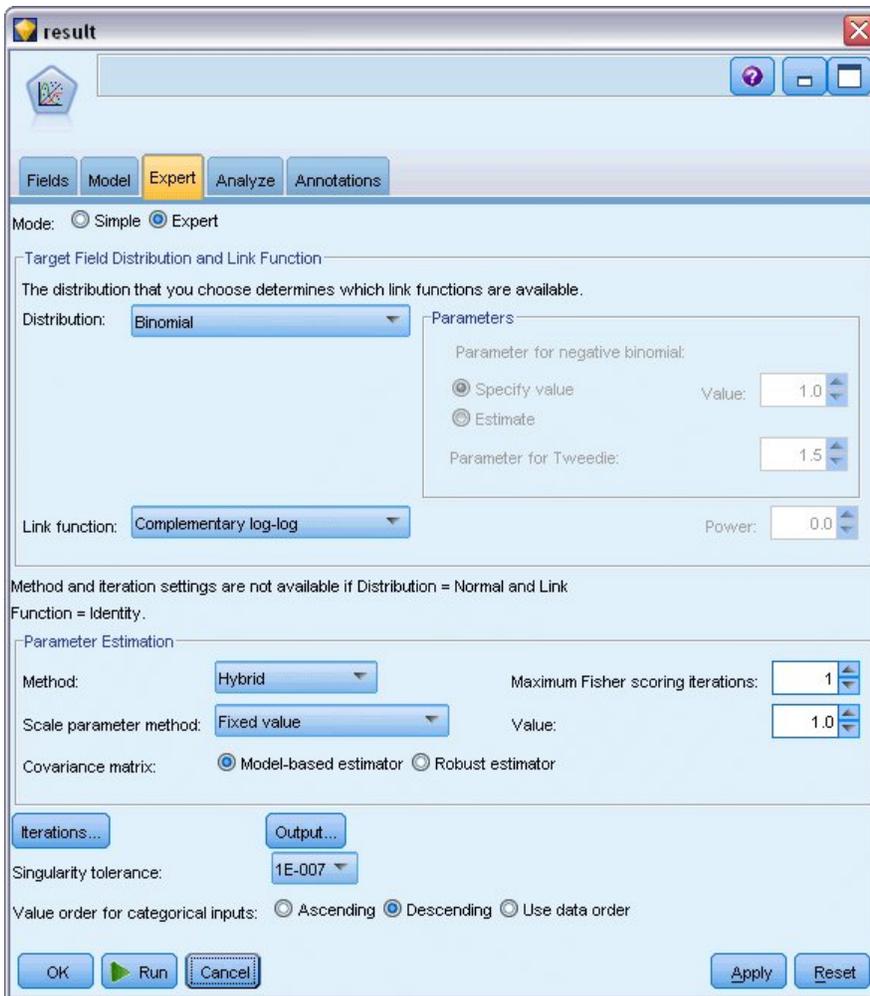


Abbildung 304. Auswählen von Expertenoptionen

9. Wählen Sie **Binomial** als Verteilung und **Log-Log komplementär** als Verknüpfungsfunktion (Linkfunktion) aus.
10. Wählen Sie **Fester Wert** als Methode zur Schätzung des Skalenparameters aus und behalten Sie den Standardwert 1,0 bei.
11. Wählen Sie **Absteigend** als Reihenfolge der Kategorien für Faktoren aus. 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 Parameterschätzungen.
12. Führen Sie den Stream aus, um das Modellnugget zu generieren; dieses wird dem Stream und der Modellpalette in der rechten oberen Ecke hinzugefügt. Um die Modelldetails anzuzeigen, können Sie mit der rechten Maustaste auf das Nugget klicken und **Bearbeiten** oder **Durchsuchen** auswählen.

Tests der Modelleffekte

Source	Type III		
	Wald Chi-Square	df	Sig.
period	.464	1	.496
duration	.000	1	.988
treatment	.117	1	.732
age	.314	1	.575

Dependent Variable: Result by period
Model: period, duration, treatment, age

Abbildung 305. Tests der Modelleffekte für das Haupteffektmodell

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

1. Klicken Sie im GenLin-Knoten auf der Registerkarte "Felder" auf **Benutzerdefinierte Einstellungen verwenden**.
2. Wählen Sie *result2* als Ziel aus.
3. Wählen Sie *period* und *treatment* als Eingaben aus.

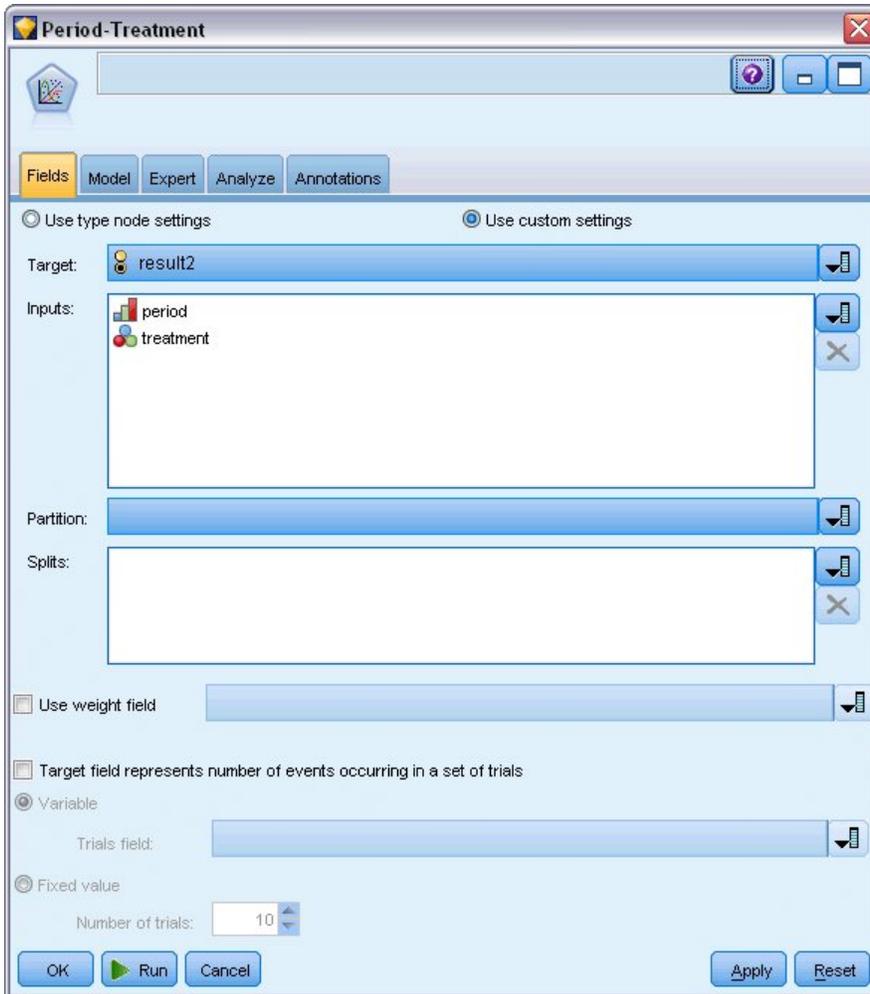


Abbildung 306. Auswählen von Feldoptionen

- 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.

Parameterschätzungen

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	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
(Scale)	1 ^b

Dependent Variable: Result by period

Model: period, treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Abbildung 307. Parameterschätzungen für das Modell "Nur Behandlung"

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 verschieden von 0. Dies ist jedoch der Fall, weil ein konstanter Term nicht angepasst ist. Der Zeiraumeffekt (der Unterschied zwischen den Werten der linearen Prädiktoren für $[period=1]$ und $[period=2]$) ist nicht statistisch signifikant, wie in den Tests der Modelleffekte deutlich wird. Der lineare Prädiktor (Zeitraum- und Behandlungseffekt) ist eine Schätzung von $\text{Log}(-\log(1-P(\text{recur}_p, t)))$, 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 308. Einstellungsoptionen für Ableitungsknoten

1. 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.
2. Geben Sie auf der Registerkarte "Einstellungen" *precur* als Ableitungsfeld ein.
3. Wählen Sie für die Ableitung des Felds die Option **Bedingt** aus.
4. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder für die Bedingung **Wenn** zu öffnen.

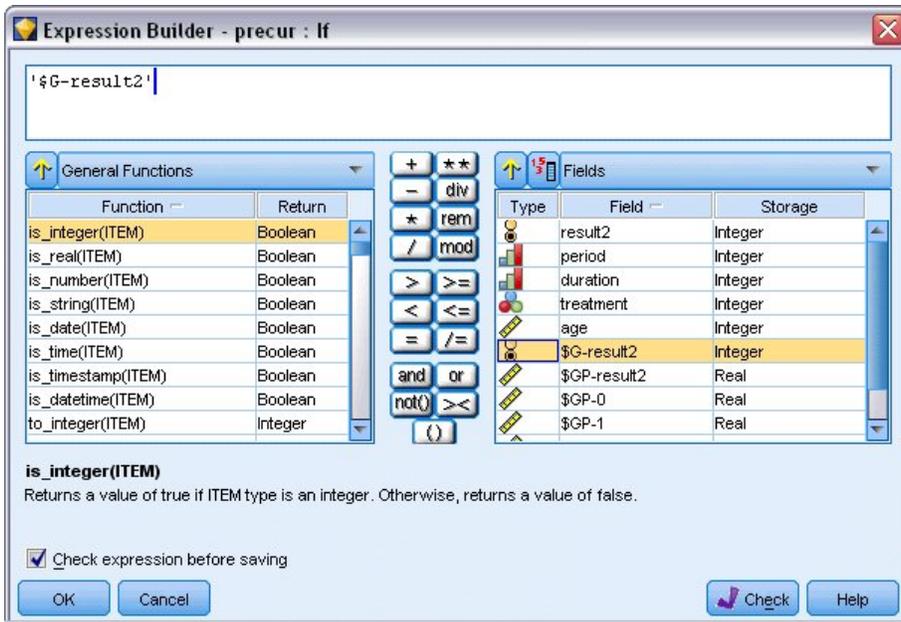


Abbildung 309. Ableitungsknoten: Expression Builder für Bedingung "Wenn"

5. Fügen Sie das Feld \$G-result2 in den Ausdruck ein.
6. 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.

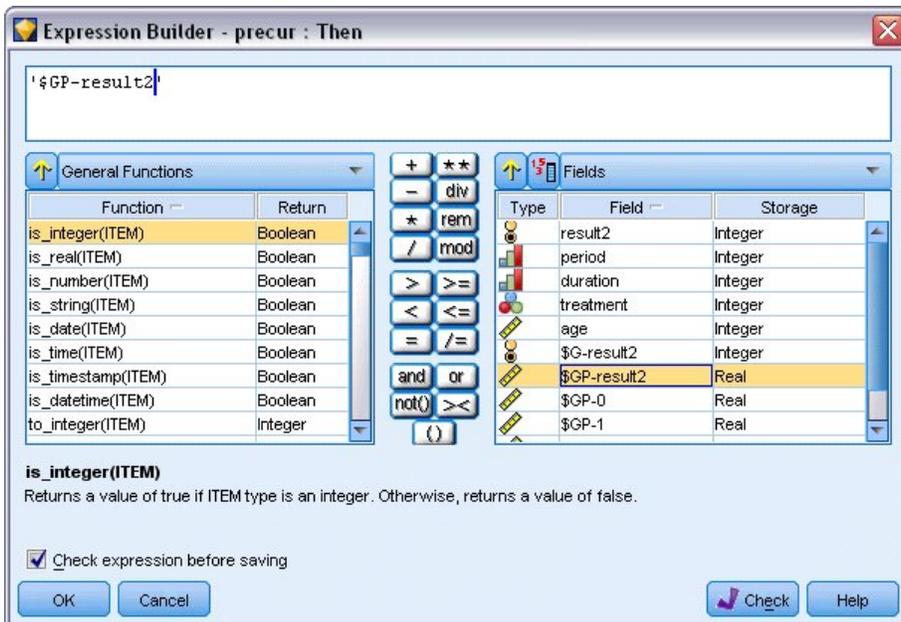


Abbildung 310. Ableitungsknoten: Expression Builder für "Dann"-Ausdruck

7. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder für den **Dann**-Ausdruck zu öffnen.
8. Fügen Sie das Feld \$GP-result2 in den Ausdruck ein.

9. Klicken Sie auf **OK**.

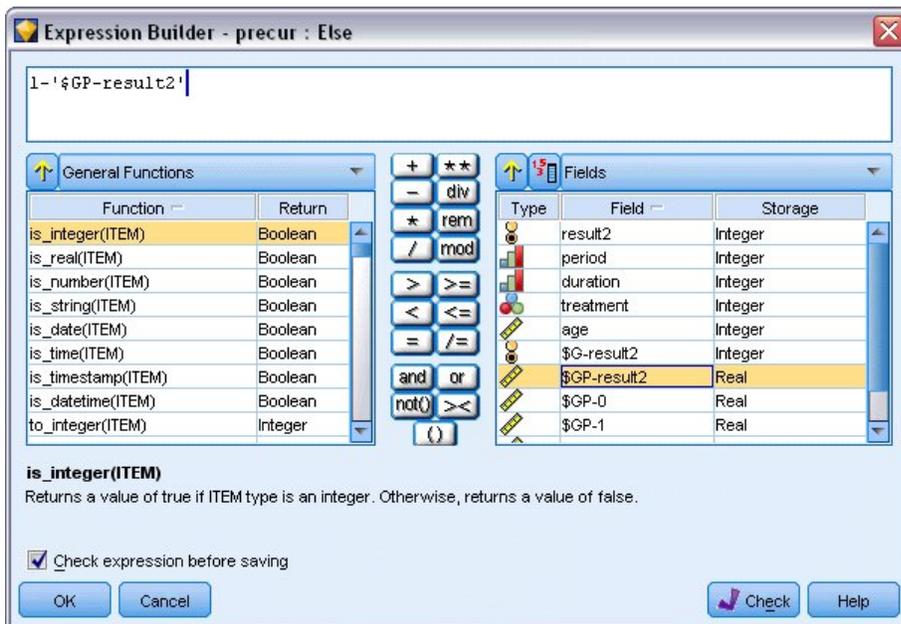


Abbildung 311. Ableitungsknoten: Expression Builder für "Sonst"-Ausdruck

10. Klicken Sie auf die Schaltfläche für den Taschenrechner, um Expression Builder für den **Sonst**-Ausdruck zu öffnen.
11. Geben Sie 1- in den Ausdruck ein und fügen Sie anschließend das Feld *\$GP-result2* in den Ausdruck ein.
12. Klicken Sie auf **OK**.



Abbildung 312. Einstellungsoptionen für Ableitungsknoten

13. Fügen Sie einen Tabellenknoten an den Ableitungsknoten an und führen Sie ihn aus.

	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	0.817	0.183
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

Abbildung 313. Geschätzte Wahrscheinlichkeiten

Tabelle 3. Geschätzte Wahrscheinlichkeiten für ein erneutes Auftreten

Behandlung	6 Monate	12 Monate
A	0,104	0,153
B	0,125	0,183

Aus den geschätzten Wahrscheinlichkeiten für ein erneutes Auftreten kann die Überlebenswahrscheinlichkeit für 12 Monate geschätzt werden als $1 - (P(\text{recur}_{1, t}) + P(\text{recur}_{2, t}) \times (1 - P(\text{recur}_{1, t})))$; d. h. für jede Behandlung:

$$A: 1 - (0,104 + 0,153 \times 0,896) = 0,759$$

$$B: 1 - (0,125 + 0,183 \times 0,875) = 0,715$$

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 *IBM SPSS Modeler-Algorithmushandbuch* aufgeführt.

Verwandte Prozeduren

Die Prozedur "Verallgemeinerte lineare Modelle" ist ein leistungsfähiges Tool zur Anpassung vieler verschiedener Modelle.

- Die Prozedur "Verallgemeinerte Schätzungsgleichungen" erweitert das verallgemeinerte lineare Modell um Messwiederholungen.
- Mithilfe der Prozedur "Lineare gemischte Modelle" können Sie Modelle für metrische abhängige Variablen mit einer zufälligen Komponente und/oder Messwiederholungen anpassen.

Empfohlene Texte

Weitere Informationen zu verallgemeinerten linearen Modellen finden Sie in den folgenden Texten:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Kapitel 23. Verwenden der Poisson-Regression für die Analyse von Schiffsschadensraten (verallgemeinerte lineare Modelle)

Verallgemeinerte lineare Modelle können zur Anpassung einer Poisson-Regression für die Analyse von Häufigkeitsdaten verwendet werden. So bezieht sich beispielsweise ein an anderer Stelle ⁽²⁾ vorgestelltes und analysiertes Dataset auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfalhäufigkeiten können unter Angabe der Werte der Prädiktoren 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 Streamdatei im Unterordner *streams*.

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 Prädiktoren linear ist. Wenn eine Poisson-Regression mithilfe von verallgemeinerten linearen Modellen an die Unfallraten angepasst werden soll, müssen Sie daher *Logarithm of aggregate months of service* (Logarithmus des Aggregats der Betriebsmonate) verwenden.

Anpassen einer Poisson-Regression mit Überdispersion

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *ships.sav* im Ordner *Demos* verweist.

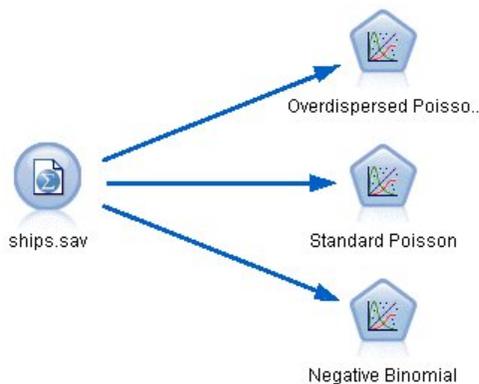


Abbildung 314. Beispielstream für die Analyse von Schadensraten

2. Schließen Sie auf der Registerkarte "Filter" des Quellenknotens das Feld *months_service* aus. Die logarithmustransformierten Werte dieser Variablen befinden sich im Feld *log_months_service*, das für die Analyse verwendet wird.

2. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

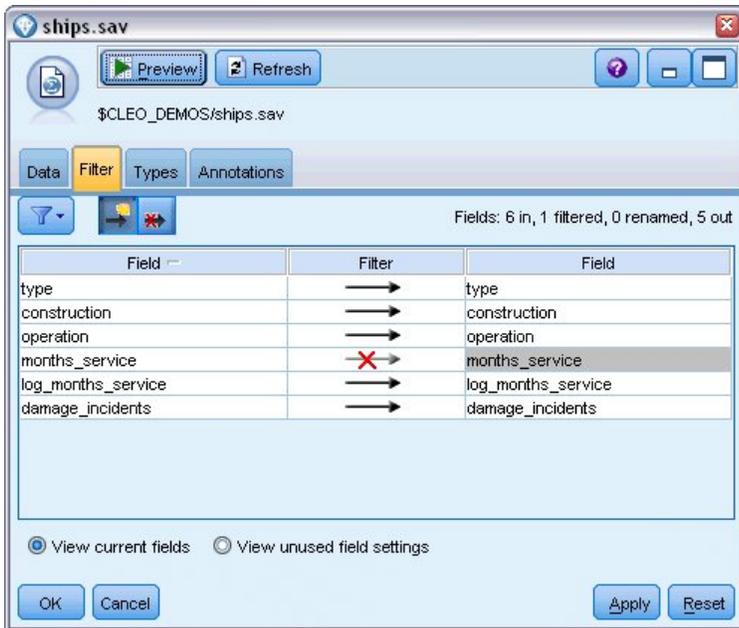


Abbildung 315. Filtern eines nicht benötigten Felds

(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.)

3. 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.
4. Klicken Sie auf **Werte lesen**, um die Daten zu instanziierten.



Abbildung 316. Festlegen der Feldrolle

5. Fügen Sie dem Quellenknoten einen Genlin-Knoten hinzu. Klicken Sie im Genlin-Knoten auf die Registerkarte **Modell**.

6. Wählen Sie *log_months_service* als Offset-Variablen aus.

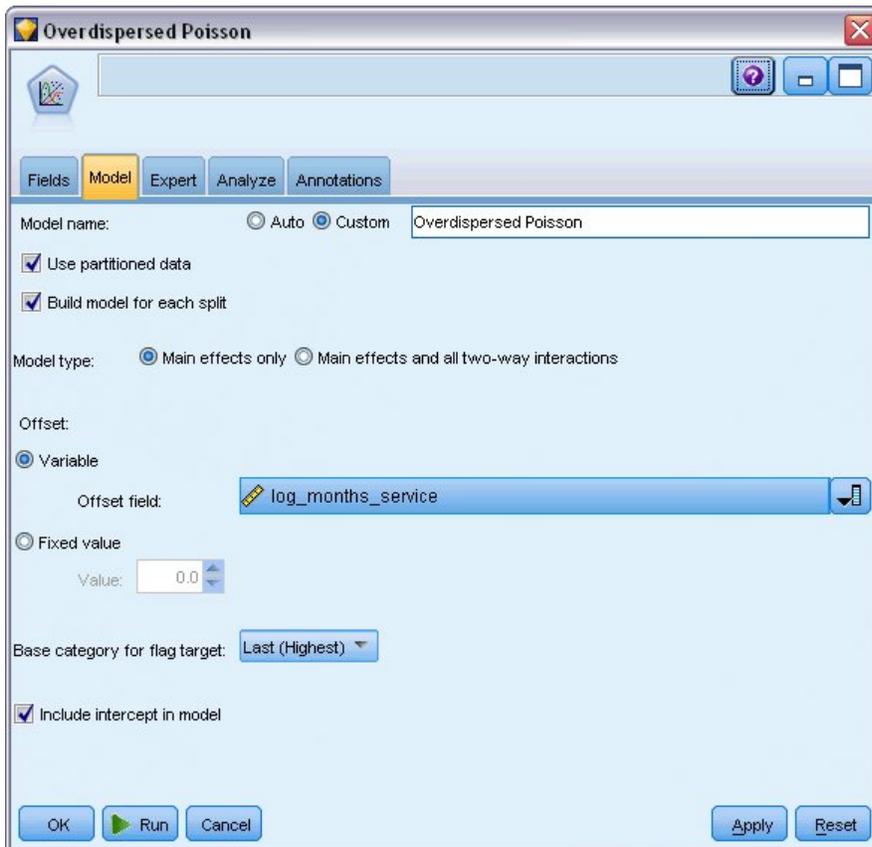


Abbildung 317. Auswählen der Modelloptionen

7. Klicken Sie auf die Registerkarte **Experten** und wählen Sie **Experten** aus, um die Expertenmodellierungsoptionen zu aktivieren.

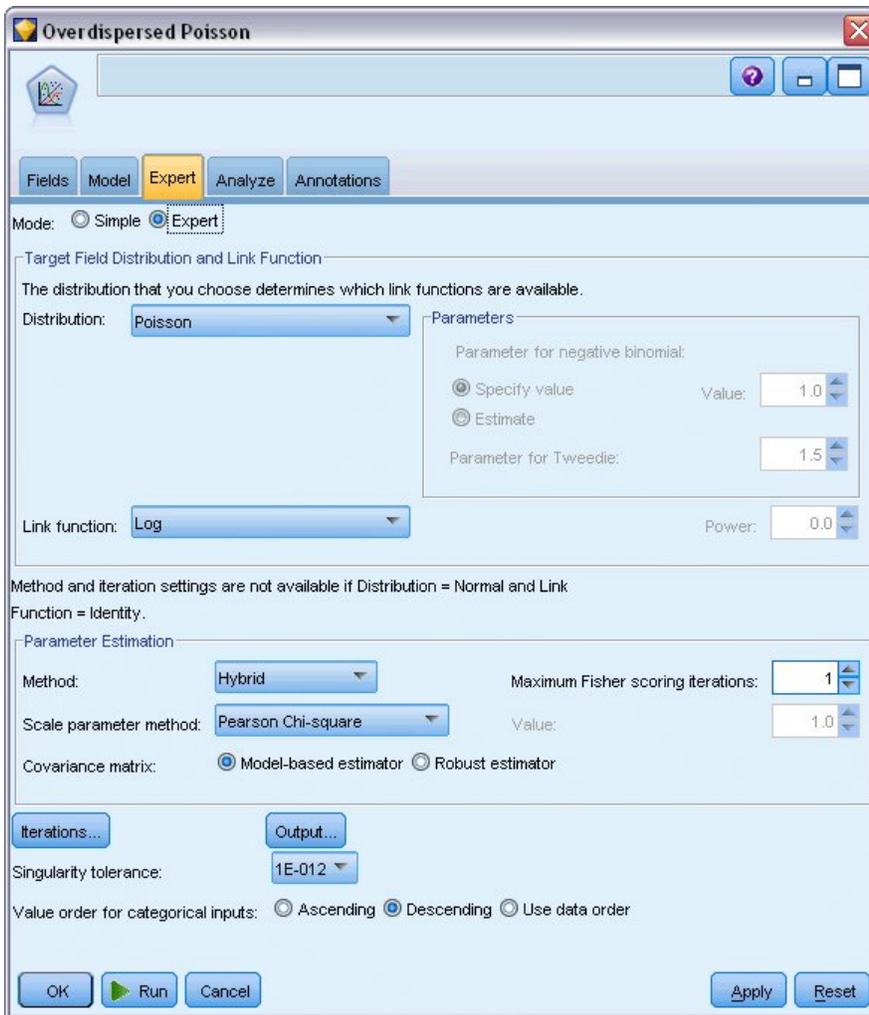


Abbildung 318. Auswählen von Expertenoptionen

8. Wählen Sie **Poisson** als Verteilung für die Antwort und **Log** als Verknüpfungsfunktion aus.
9. Wählen Sie **Pearson-Chi-Quadrat** als Methode zur Schätzung des Skalenparameters aus. Der Skalenparameter wird in einer Poisson-Regression üblicherweise mit 1 angegeben, doch McCullagh und Nelder nutzen die Pearson-Chi-Quadratschätzung, um konservativere Varianzschätzungen und Signifikanzniveaus zu erhalten.
10. Wählen Sie **Absteigend** als Reihenfolge der Kategorien für Faktoren aus. 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 Parameterschätzungen.
11. Klicken Sie auf **Ausführen**, um das Modellnugget zu erstellen; dieses wird dem Streamerstellungsbe-
reich und der Modellapalette 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

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	22.883	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	25.000	25	
Log Likelihood ^a	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

- a. The full log likelihood function is displayed and used in computing information criteria.
b. Information criteria are in small-is-better form.

Abbildung 319. Statistik für Anpassungsgüte

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 Schätzungen 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

Likelihood Ratio Chi-Square	df	Sig.
107.633	8	.000

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

- a. Compares the fitted model against the intercept-only model.

Abbildung 320. Omnibus-Test

Der Omnibus-Test ist ein Likelihood-Quotient-Chi-Quadrat-Test des aktuellen Modells im Vergleich zum Nullmodell (hier: Intercept-Modell (Modell mit konstantem Term)). Der Signifikanzwert von weniger als 0,05 zeigt an, dass das aktuelle Modell besser geeignet ist als das Nullmodell.

Tests der Modelleffekte

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	2138.657	1	.000
type	15.415	4	.004
construction	17.242	3	.001
operation	6.249	1	.012

Dependent Variable: Number of damage incidents
 Model: (Intercept), type, construction, operation, offset = log_months_service

Abbildung 321. Tests der Modelleffekte

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 Haupteffektterme tragen einen Teil zum Modell bei.

Parameterschätzungen

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-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
(Scale)	1.691 ^b

Dependent Variable: Number of damage incidents
 Model: (Intercept), type, construction, operation, offset = log_months_service

- a. Set to zero because this parameter is redundant.
 b. Computed based on the Pearson chi-square.

Abbildung 322. Parameterschätzungen

In der Tabelle der Parameterschätzungen ist der Effekt der einzelnen Prädiktoren zusammengefasst. 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 Prädiktoren im Modell bieten.

- Bei Kovariaten weisen positive (negative) Koeffizienten auf positive (inverse) Beziehungen zwischen Prädiktoren 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 vom Effekt der betreffenden Faktorstufe in Bezug zur Referenzkategorie ab.

Auf der Grundlage der Parameterschätzungen 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 geschätzte Parameter von Typ *C* [*type=3*] ist niedriger als der von Typ *B*, aber die Variabilität in der Schätzung für *C* verwischt 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 Poisson-Standardregression zu testen. Ein Vorschlag für einen formellen Test für die Feststellung von Überdispersion ist jedoch ein Likelihood-Quotienten-Test zwischen einer Poisson-Standardregression und einer negativen binomialen Regression, wobei alle sonstigen Einstellungen gleich bleiben. Wenn keine Überdispersion in der Poisson-Regression vorliegt, sollte die Statistik $-2 \times (\text{Log-Likelihood für Poisson-Modell} - \text{Log-Likelihood für negatives binomiales Modell})$ eine gemischte Verteilung mit der Hälfte der Wahrscheinlichkeitsmasse bei 0 und dem Rest in einer Chi-Quadrat-Verteilung mit 1 Freiheitsgrad aufweisen.

1. Wählen Sie **Fester Wert** als Methode zur Schätzung des Skalenparameters aus. Standardmäßig entspricht dieser dem Wert 1.

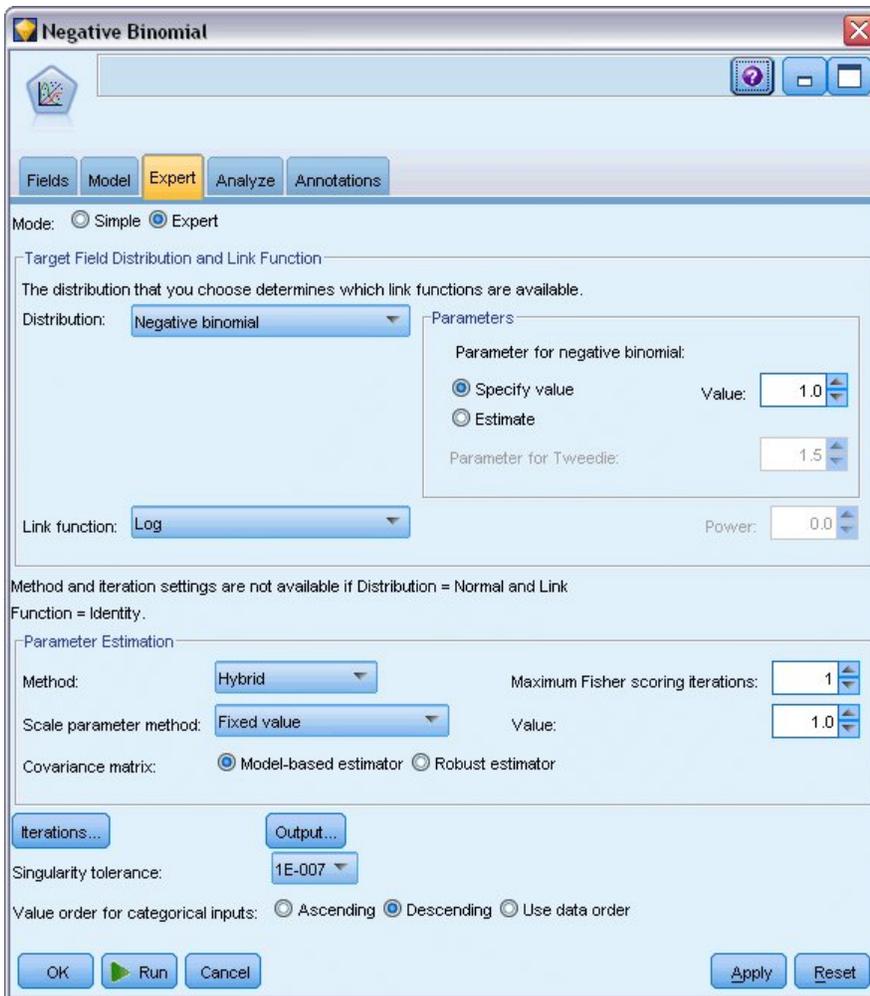


Abbildung 323. Registerkarte "Experten"

2. Um die negative binomiale Regression anzupassen, kopieren Sie den geöffneten Genlin-Knoten, fügen ihn an den Quellenknoten an, öffnen den neuen Knoten und klicken auf die Registerkarte **Experten**.
3. Wählen Sie **Negativ binomial** als Verteilung aus. Behalten Sie den Standardwert 1 für den Hilfsparameter bei.
4. Führen Sie den Stream aus und durchsuchen Sie die Registerkarte "Erweitert" für die neu erstellten Modellnuggets.

Statistik für Anpassungsgüte

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	38.695	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	42.275	25	
Log Likelihood ^a	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents

Model: (Intercept), type, construction, operation, offset = log_months_service

- a. The full log likelihood function is displayed and used in computing information criteria.
- b. Information criteria are in small-is-better form.

Abbildung 324. Statistiken der Anpassungsgüte für die Poisson-Standardregression

Die ausgegebene Log-Likelihood für die Poisson-Standardregression ist -68,281. Vergleichen Sie dies mit dem negativen binomialen Modell.

	Value	df	Value/df
Deviance	11.145	25	.446
Scaled Deviance	11.145	25	
Pearson Chi-Square	8.815	25	.353
Scaled Pearson Chi-Square	8.815	25	
Log Likelihood ^a	-83.725		
Akaike's Information Criterion (AIC)	185.450		
Finite Sample Corrected AIC (AICC)	192.950		
Bayesian Information Criterion (BIC)	199.187		
Consistent AIC (CAIC)	208.187		

Dependent Variable: Number of damage incidents

Model: (Intercept), type, construction, operation, offset = log_months_service

- a. The full log likelihood function is displayed and used in computing information criteria.
- b. Information criteria are in small-is-better form.

Abbildung 325. Statistiken der Anpassungsgüte für die negative binomiale Regression

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 Dataset 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 Dataset kein Problem sein.

Zusammenfassung

Mit verallgemeinerten linearen Modellen haben Sie drei verschiedene Modelle für Häufigkeitsdaten angepasst. Die negative binomiale Regression konnte keine Verbesserung gegenüber der 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 *IBM SPSS Modeler-Algorithmushandbuch* aufgeführt.

Verwandte Prozeduren

Die Prozedur "Verallgemeinerte lineare Modelle" ist ein leistungsfähiges Tool zur Anpassung vieler verschiedener Modelle.

- Die Prozedur "Verallgemeinerte Schätzungsgleichungen" erweitert das verallgemeinerte lineare Modell um Messwiederholungen.
- Mithilfe der Prozedur "Lineare gemischte Modelle" können Sie Modelle für metrische abhängige Variablen mit einer zufälligen Komponente und/oder Messwiederholungen anpassen.

Empfohlene Texte

Weitere Informationen zu verallgemeinerten linearen Modellen finden Sie in den folgenden Texten:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Kapitel 24. Anpassen einer Gammaregression an Versicherungsforderungen an Kfz-Versicherungen (verallgemeinerte lineare Modelle)

Verallgemeinerte lineare Modelle können zur Anpassung einer Gammaregression für die Analyse eines positiven Datenbereichs verwendet werden. So befasst sich beispielsweise ein an anderer Stelle ⁽³⁾ vorgestelltes und analysiertes Dataset Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit einer Gammaverteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination der Prädiktoren 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 Skalierungsgewichtung 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 Streamdatei im Unterordner *streams*.

Erstellen des Modells

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *car_insurance_claims.sav* im Ordner *Demos* verweist.



Abbildung 326. Beispielstream für die Vorhersage von Schadensansprüchen für Autos

2. 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.
3. Klicken Sie auf **Werte lesen**, um die Daten zu instanziiieren.

3. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.



Abbildung 327. Festlegen der Feldrolle

4. Fügen Sie dem Quellenknoten einen Genlin-Knoten hinzu. Klicken Sie im Genlin-Knoten auf die Registerkarte "Felder".
5. Wählen Sie *nclaims* als Skalengewichtungsfeld aus.

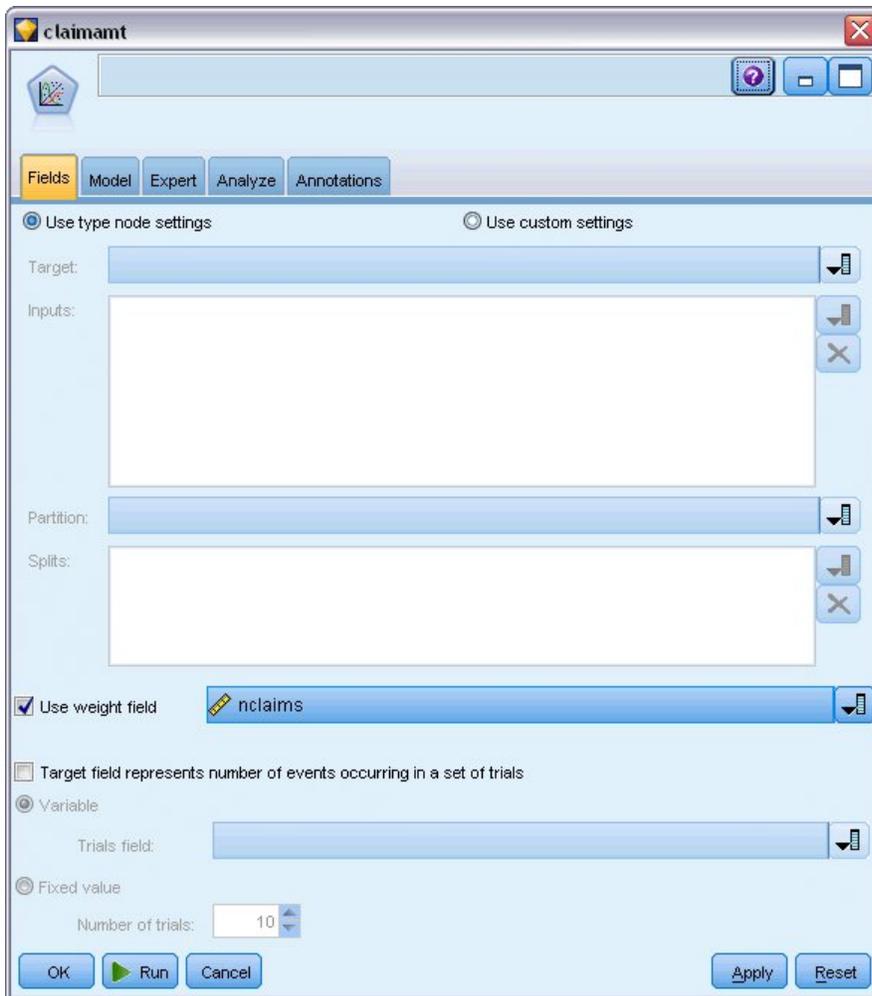


Abbildung 328. Auswählen von Feldoptionen

6. Klicken Sie auf die Registerkarte "Experten" und wählen Sie **Experten** aus, um die Expertenmodellierungsoptionen zu aktivieren.

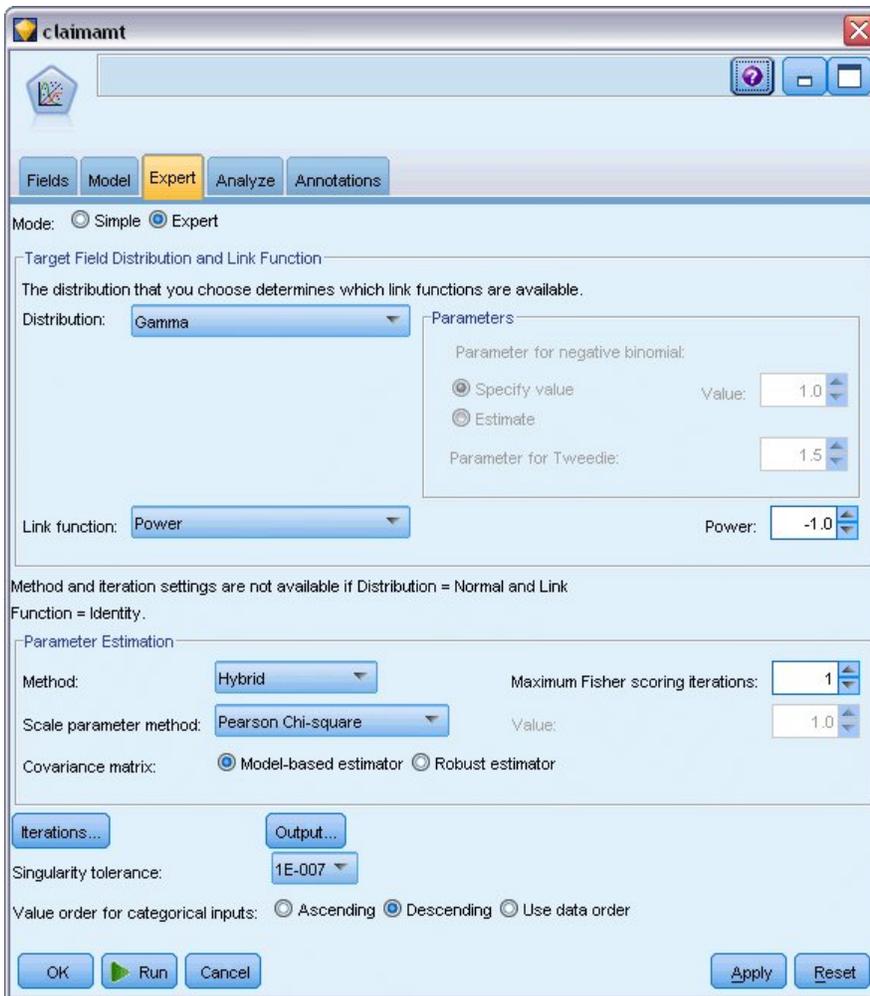


Abbildung 329. Auswählen von Expertenoptionen

7. Wählen Sie **Gamma** als Antwortverteilung aus.
8. Wählen Sie **Potenz** als Verknüpfungsfunktion aus und geben Sie -1,0 als Exponenten der Potenzfunktion ein. Dies ist eine inverse Verknüpfung.
9. Wählen Sie **Pearson-Chi-Quadrat** als Methode zur Schätzung des Skalenparameters aus. Diese Methode wird von McCullagh und Nelder verwendet. Wir folgen ihrer Methode, um ihre Ergebnisse zu reproduzieren.
10. Wählen Sie **Absteigend** als Reihenfolge der Kategorien für Faktoren aus. 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 Parameterschätzungen.
11. Klicken Sie auf **Ausführen**, um das Modellnugget zu erstellen; dieses wird dem Streamerstellungsbereich und der Modellapalette in der rechten oberen Ecke hinzugefügt. Zum Anzeigen der Modelldetails klicken Sie mit der rechten Maustaste auf das Modellnugget und wählen **Bearbeiten** oder **Durchsuchen**. Anschließend wählen Sie die Registerkarte "Erweitert" aus.

Parameterschätzungen

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.003411	.000418	.002591	.004230	66.593	1	.000
[holderage=8]	.000920	.000416	.000105	.001735	4.898	1	.027
[holderage=7]	.000916	.000408	.000117	.001716	5.046	1	.025
[holderage=6]	.000969	.000405	.000176	.001763	5.740	1	.017
[holderage=5]	.001370	.000419	.000548	.002192	10.682	1	.001
[holderage=4]	.000462	.000411	-.000342	.001267	1.268	1	.260
[holderage=3]	.000350	.000412	-.000458	.001158	.720	1	.396
[holderage=2]	.000101	.000436	-.000754	.000956	.054	1	.816
[holderage=1]	.000000 ^a
[vehiclegroup=4]	-.001421	.000181	-.001775	-.001067	61.883	1	.000
[vehiclegroup=3]	-.000614	.000170	-.000947	-.000281	13.039	1	.000
[vehiclegroup=2]	.000038	.000169	-.000293	.000368	.050	1	.823
[vehiclegroup=1]	.000000 ^a
[vehicleage=4]	.004154	.000442	.003287	.005021	88.175	1	.000
[vehicleage=3]	.001651	.000227	.001207	.002096	53.013	1	.000
[vehicleage=2]	.000366	.000101	.000169	.000564	13.191	1	.000
[vehicleage=1]	.000000 ^a
(Scale)	1.209 ^b	.	.	.001	.0004	.000	.002

Dependent Variable: Average cost of claims

Model: (Intercept), holderage, vehiclegroup, vehicleage

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

Abbildung 330. Parameterschätzungen

Der Omnibus-Test und Tests der Modelleffekte (nicht dargestellt) zeigen, dass das Modell besser als das Nullmodell funktioniert und dass alle Haupteffektterme einen Beitrag zum Modell leisten. Die Tabelle der Parameterschätzungen enthält dieselben Werte, die auch McCullagh und Nelder für die Faktorstufen und den Skalenparameter erhalten haben.

Zusammenfassung

Mit verallgemeinerten linearen Modellen haben Sie eine Gammaregression an die Forderungsdaten angepasst. Beachten Sie: In diesem Modell wurde zwar die kanonische Linkfunktion für die Gammaverteilung 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 *IBM SPSS Modeler-Algorithmushandbuch* aufgeführt.

Verwandte Prozeduren

Die Prozedur "Verallgemeinerte lineare Modelle" ist ein leistungsfähiges Tool zur Anpassung vieler verschiedener Modelle.

- Die Prozedur "Verallgemeinerte Schätzungsgleichungen" erweitert das verallgemeinerte lineare Modell um Messwiederholungen.
- Mithilfe der Prozedur "Lineare gemischte Modelle" können Sie Modelle für metrische abhängige Variablen mit einer zufälligen Komponente und/oder Messwiederholungen anpassen.

Empfohlene Texte

Weitere Informationen zu verallgemeinerten linearen Modellen finden Sie in den folgenden Texten:

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Kapitel 25. Klassifizieren von Zellproben (SVM)

Support Vector Machine (SVM) ist ein Klassifikations- und Regressionsverfahren, das sich besonders für umfangreiche Datensets eignet. Umfangreiche Datensets sind Datensets 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 Dataset mit den Eigenschaften einer Reihe von Stichproben menschlicher Zellen erstellt, die von Patienten stammen, bei denen ein Krebsrisiko angenommen wurde. Die Analyse der ursprünglichen Daten ergab, dass bei vielen der Eigenschaften deutliche Unterschiede zwischen den gutartigen und den bösartigen Proben bestehen. Der Forscher möchte ein SVM-Modell entwickeln, das die Werte 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. Als Datendatei wird die Datei *cell_samples.data* verwendet. Weitere Informationen finden Sie unter im Thema „Ordner "Demos"“ auf Seite 5.

Das Beispiel beruht auf einem Dataset, das im UCI Machine Learning Repository öffentlich zugänglich ist. Das Dataset 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
<i>ID</i>	Patienten-ID
<i>Clump</i>	Clusterdicke
<i>UnifSize</i>	Einheitlichkeit der Zellgröße
<i>UnifShape</i>	Einheitlichkeit der Zellform
<i>MargAdh</i>	Randhaftung
<i>SingEpiSize</i>	Größe einzelner Epithelzellen
<i>BareNuc</i>	Nackte Zellkerne
<i>BlandChrom</i>	Homogenes Chromatin
<i>NormNucl</i>	Normale Kernkörperchen
<i>Mit</i>	Mitose
<i>Class</i>	Gutartig oder bösartig

Für dieses Beispiel verwenden wir ein Dataset mit einer relativ kleinen Anzahl an Prädiktoren in jedem Datensatz.

Erstellen des Streams

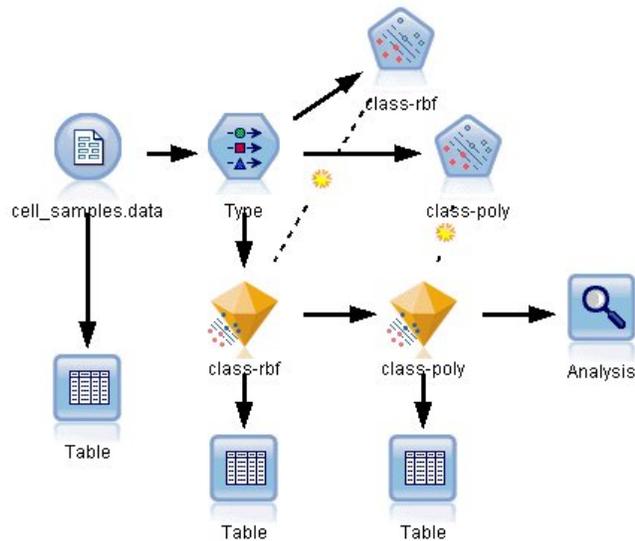


Abbildung 331. Beispielstream zur Veranschaulichung der SVM-Modellierung

1. 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 Quellendatei.
2. Fügen Sie dem Stream einen Tabellenknoten hinzu.
3. Fügen Sie den Tabellenknoten mit dem Knoten für variable Dateien an und führen Sie den Stream aus.

	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1	1	1	1	2	1	3	1	1	2
2	4	5	7	10	3	2	1	1	2
3	1	1	2	2	3	1	1	1	2
4	8	1	3	4	3	7	1	1	2
5	1	3	2	1	3	1	1	1	2
6	10	8	7	10	9	7	1	1	4
7	1	1	2	10	3	1	1	1	2
8	2	1	2	1	3	1	1	1	2
9	1	1	2	1	1	1	1	5	2
10	1	1	2	1	2	1	1	1	2
11	1	1	1	1	3	1	1	1	2
12	1	1	2	1	2	1	1	1	2
13	3	3	2	3	4	4	1	1	4
14	1	1	2	3	3	1	1	1	2
15	5	10	7	9	5	5	4	4	4
16	6	4	6	1	4	3	1	1	4
17	1	1	2	1	2	1	1	1	2
18	1	1	2	1	3	1	1	1	2
19	7	6	4	10	4	1	2	4	4
20	1	1	2	1	3	1	1	1	2

Abbildung 332. Quelldaten für SVM

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.

Field	Measurement	Values	Missing	Check	Role
UnifSize	Continuous	[1,10]		None	Input
UnifShape	Continuous	[1,10]		None	Input
MargAdh	Continuous	[1,10]		None	Input
SingEpiSize	Continuous	[1,10]		None	Input
BareNuc	Nominal	"1","10",...		None	Input
BlandChrom	Continuous	[1,10]		None	Input
NormNucl	Continuous	[1,10]		None	Input
Mit	Continuous	[1,10]		None	Input
Class	Flag	4/2		None	Target

Abbildung 333. Typknoteneinstellungen

4. Fügen Sie einen Typknoten hinzu und fügen Sie ihn an den Knoten für variable Dateien an.

5. 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.
6. 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**.
7. Klicken Sie auf **Werte lesen**.
8. 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.
9. 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.
10. Klicken Sie auf **OK**.
Der SVM-Knoten bietet eine Auswahl an Kernfunktionen zur Durchführung der Verarbeitung. Da sich nicht im Voraus sagen lässt, welche Funktion mit einem bestimmten Dataset am besten funktioniert, wählen wir abwechselnd verschiedene Funktionen aus und vergleichen die Ergebnisse. Beginnen wir mit der Standardvorgabe "RBF" (Radial Basis Function).

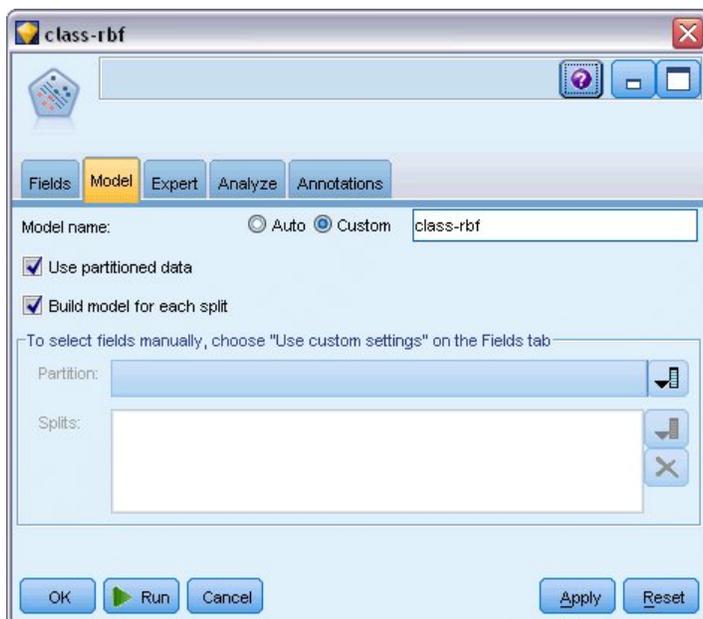


Abbildung 334. Einstellungen auf der Registerkarte "Modell"

11. Fügen Sie auf der Modellierungspalette einen SVM-Knoten an den Typknoten an.
12. 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.

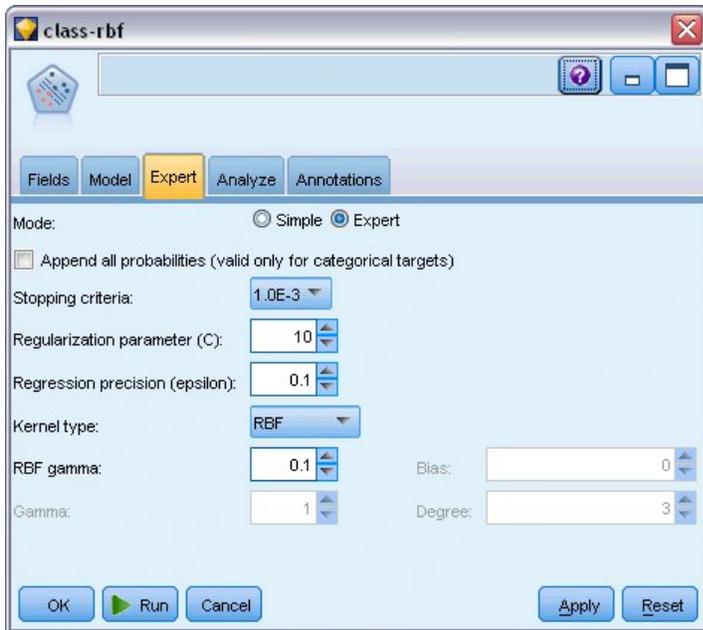


Abbildung 335. Registerkarte "Experten" - Standardeinstellungen

13. Setzen Sie auf der Registerkarte **Experten** den Wert von **Modus** auf **Experten**, um eine bessere Lesbarkeit zu erzielen. Übernehmen Sie alle Standardoptionen. Beachten Sie, dass **Kerntyp** standardmäßig auf **RBF** eingestellt ist. Im Modus "Einfach" sind alle Optionen abgeblendet.

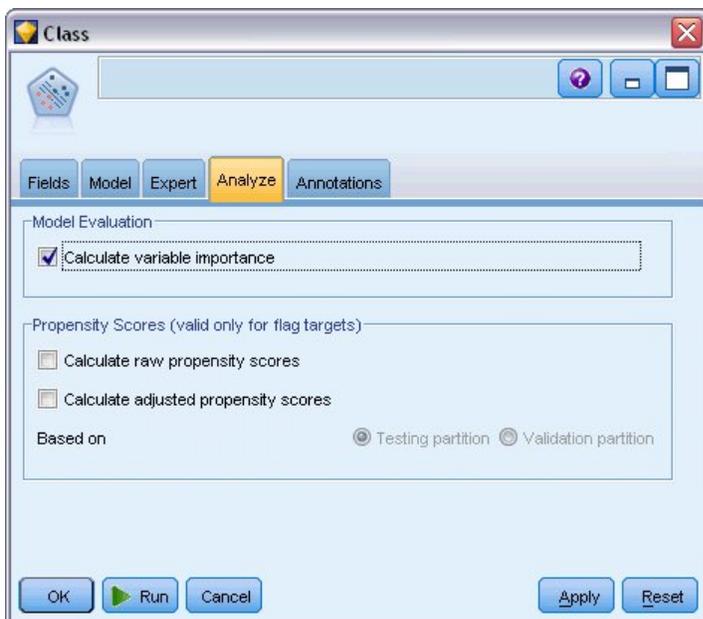


Abbildung 336. Einstellungen auf der Registerkarte "Analysieren"

14. Aktivieren Sie auf der Registerkarte **Analysieren** das Kontrollkästchen **Bedeutsamkeit der Variablen berechnen**.
15. Klicken Sie auf **Ausführen**. Das Modellnugget wird in den Stream und in der Modellpalette oben rechts im Fenster platziert.
16. Doppelklicken Sie auf das Modellnugget im Stream.

Untersuchen der Daten

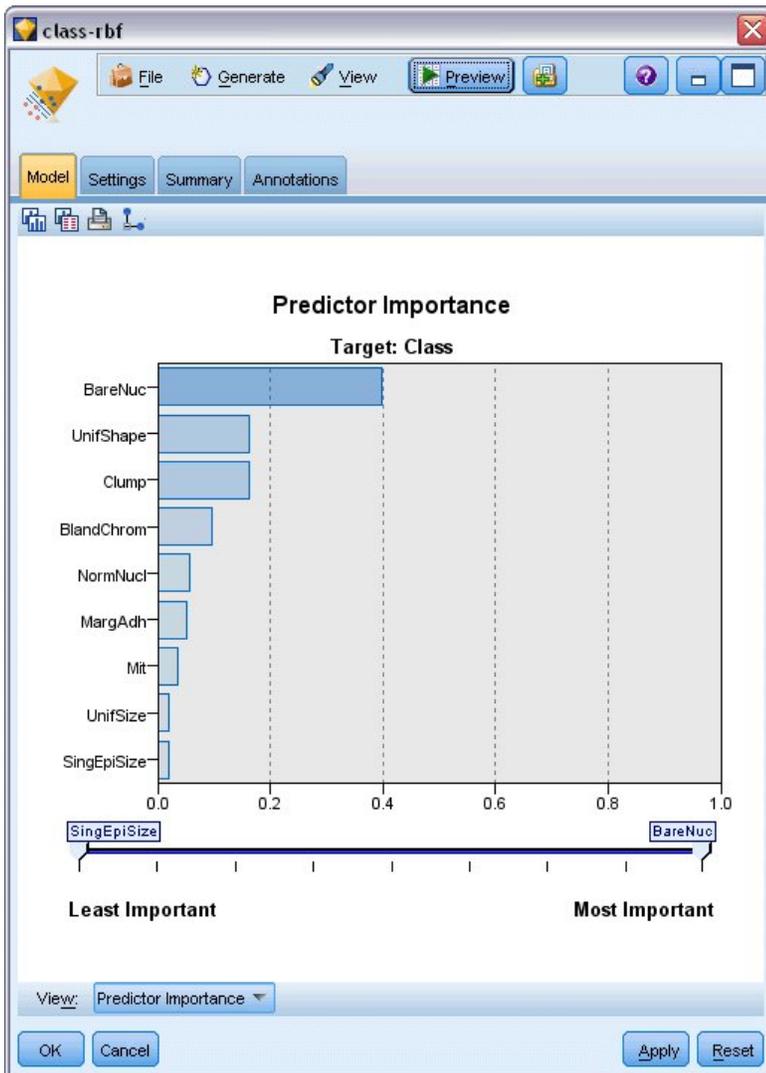


Abbildung 337. Diagramm für die Bedeutsamkeit des Prädiktors

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.

1. Klicken Sie auf **OK**.
2. Fügen Sie einen Tabellenknoten an das Modellnugget *class-rbf* an.
3. Öffnen Sie den Tabellenknoten und klicken Sie auf **Ausführen**.

	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	

Abbildung 338. Für Vorhersage und Konfidenzwert hinzugefügte Felder

4. 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 <i>Class</i> vorhergesagte Wert.
$\$SP$ -Class	Propensity-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 Propensity-Scores (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 Propensity-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

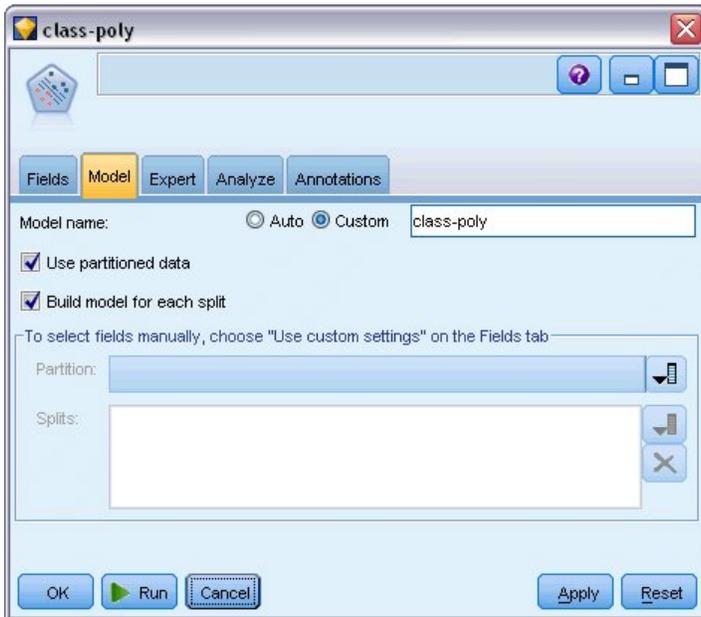


Abbildung 339. Festlegung eines neuen Namens für das Modell

1. Schließen Sie das Tabellenausgabefenster.
2. Fügen Sie einen SVM-Modellierungsknoten an den Typknoten an.
3. Öffnen Sie den neuen SVM-Knoten.
4. Wählen Sie auf der Registerkarte **Modell** die Option "Benutzerdefiniert" und geben Sie *class-poly* als Modellnamen ein.

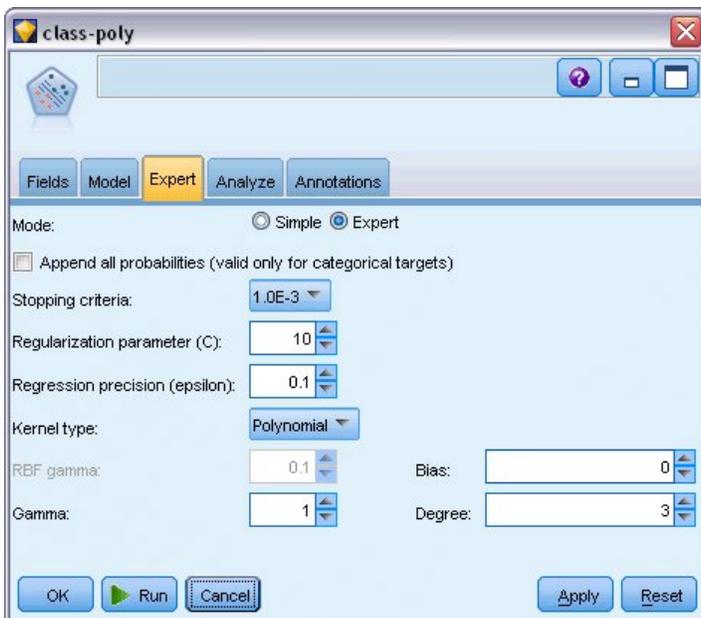


Abbildung 340. Registerkarte "Experten" - Einstellungen für den "Polynomial"

5. Setzen Sie auf der Registerkarte **Experten** den Wert von **Modus** auf **Experten**.

6. Setzen Sie **Kerntyp** auf **Polynomial** und klicken Sie auf **Ausführen**. Das Modellnugget *class-poly* wird in den Stream und in der Modellalette oben rechts im Fenster platziert.
7. Verbinden Sie das Modellnugget *class-rbf* mit dem Modellnugget *class-poly* (wählen Sie **Ersetzen** im Warnungsdialog).
8. Fügen Sie einen Tabellenknoten an das Nugget *class-poly* an.
9. Öffnen Sie den Tabellenknoten und klicken Sie auf **Ausführen**.

Vergleichen der Ergebnisse

	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	0	7	4	4	0.992	4	1.000
86	0	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

Abbildung 341. Für Polynomialfunktion hinzugefügte Felder

1. Führen Sie für die Tabellenausgabe einen Bildlauf nach rechts durch, um die neu hinzugefügten Felder anzuzeigen.
Die generierten Felder für den Funktionstyp "Polynomial" lauten *\$S1-Class* und *\$SP1-Class*.
Die Ergebnisse für "Polynomial" sehen deutlich besser aus. Viele der Propensity-Scores liegen bei 0,995 oder höher, was sehr ermutigend ist.
2. Um die Verbesserung des Modells zu bestätigen, fügen Sie einen Analyseknoden an das Modellnugget *class-poly* an.

Öffnen Sie den Analyseknoden und klicken Sie auf **Ausführen**.

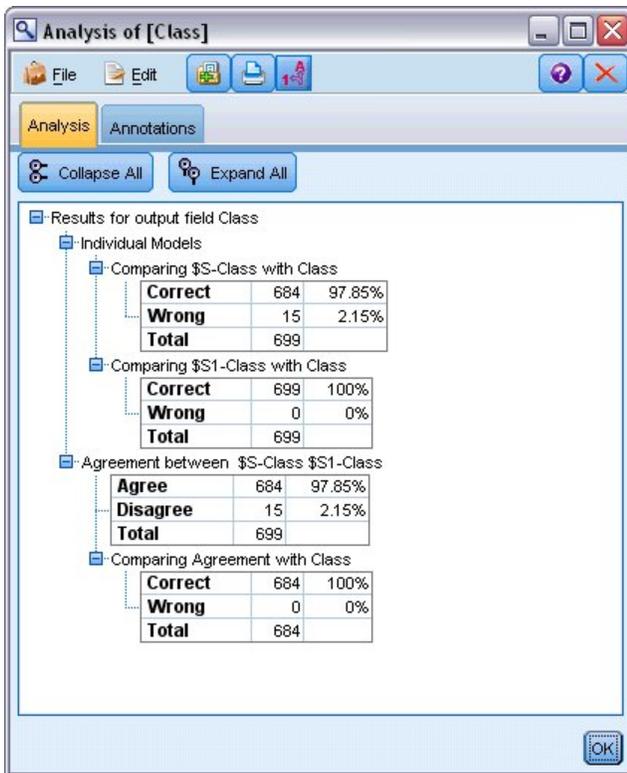


Abbildung 342. Analysenoten

Durch dieses Verfahren mit dem Analysenoten können Sie zwei oder mehr Modellnuggets desselben Typs vergleichen. Die Ausgabe aus dem Analysenoten zeigt, dass die Funktion "RBF" 97,85 % der Fälle korrekt vorhersagt, was noch recht gut ist. Die Ausgabe zeigt jedoch, dass die Polynomfunktion die Diagnose in jedem einzelnen Fall korrekt vorhersagte. In der Praxis werden Sie kaum eine 100%ige Genauigkeit erreichen, Sie können jedoch mithilfe des Analysenotens 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 Dataset eine so gute Leistung wie die Funktion "Polynomial". Bei einem anderen Dataset könnten die Ergebnisse jedoch durchaus anders sein, sodass es sich immer lohnt, das gesamte Optionsspektrum auszuschöpfen.

Zusammenfassung

Sie haben mithilfe verschiedener Typen von SVM-Kernfunktionen eine Klassifikation aus einer Reihe von Attributen vorhergesagt. Sie haben gesehen, dass unterschiedliche Kerne bei demselben Dataset zu unterschiedlichen Ergebnissen führen, und festgestellt, wie Sie die Qualitätsunterschiede zwischen den Modellen messen können.

Kapitel 26. 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 Streamdatei im Unterordner *streams*. Weitere Informationen finden Sie unter im Thema „Ordner "Demos"“ auf Seite 5.

Erstellen eines geeigneten Modells

1. Fügen Sie einen Quellenknoten für Statistikdateien hinzu, der auf *telco.sav* im Ordner *Demos* verweist.

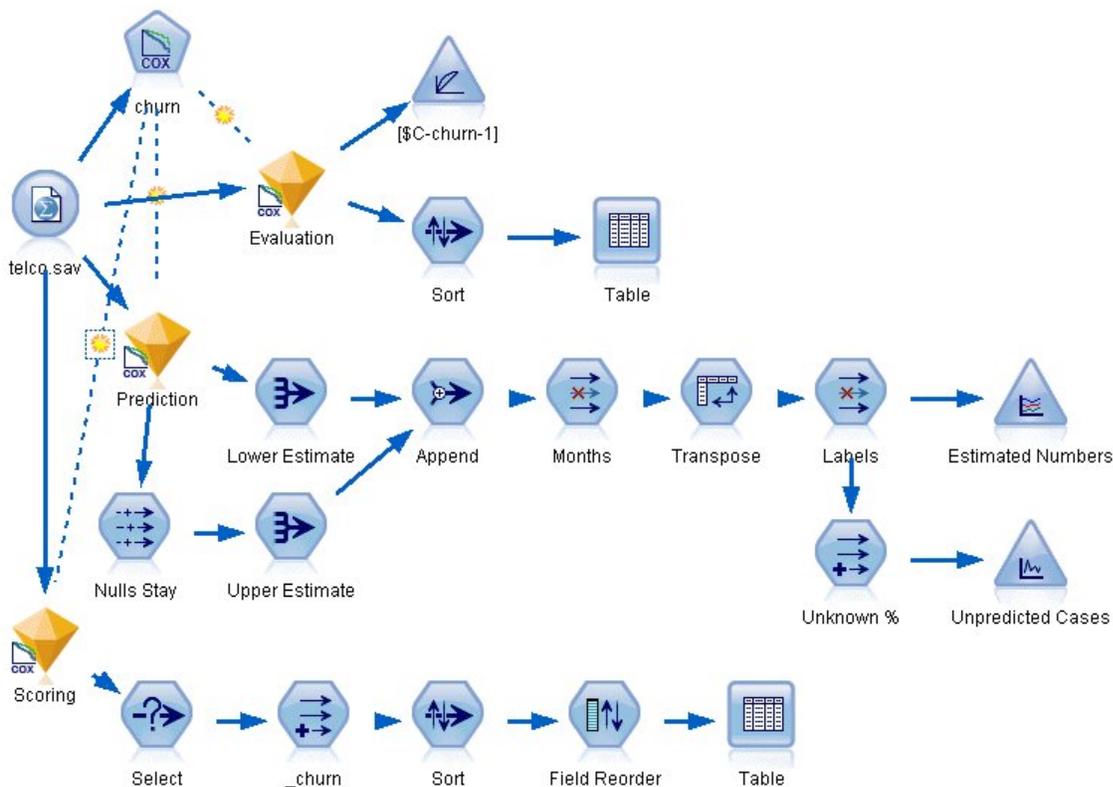


Abbildung 343. Beispielstream zur Analyse der Zeit bis zur Abwanderung

2. Schließen Sie auf der Registerkarte "Filter" des Quellenknotens die Felder *region*, *income*, *longten* bis *wireten* und *loglong* bis *logwire* aus.



Abbildung 344. 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.)

3. 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 **Eingabe** für die Rolle festgelegt sein.
4. Klicken Sie auf **Werte lesen**, um die Daten zu instanziierten.



Abbildung 345. Festlegen der Feldrolle

5. Fügen Sie einen Cox-Knoten an den Quellenknoten an; wählen Sie auf der Registerkarte **Felder** den Eintrag *tenure* als Variable für die Überlebenszeit aus.

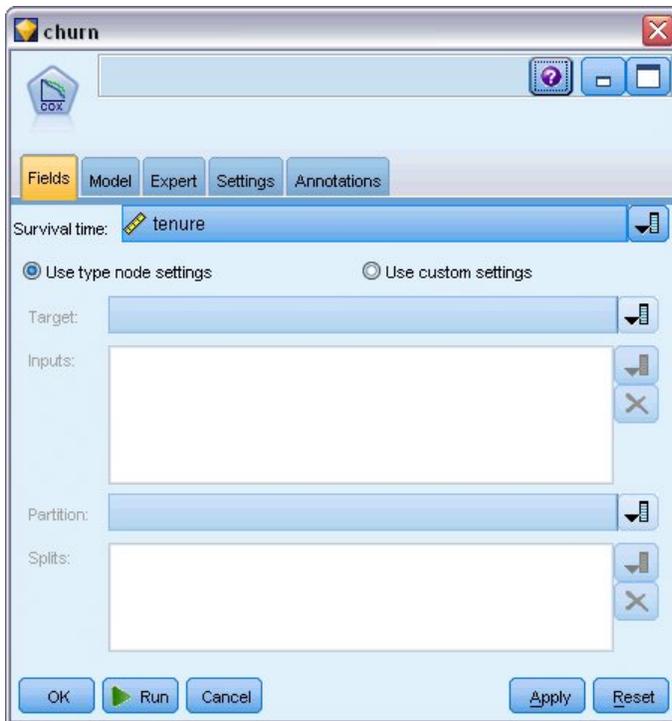


Abbildung 346. Auswählen von Feldoptionen

6. Klicken Sie auf die Registerkarte **Modell**.
7. Wählen Sie **Schrittweise** als Methode für die Auswahl der Variablen aus.

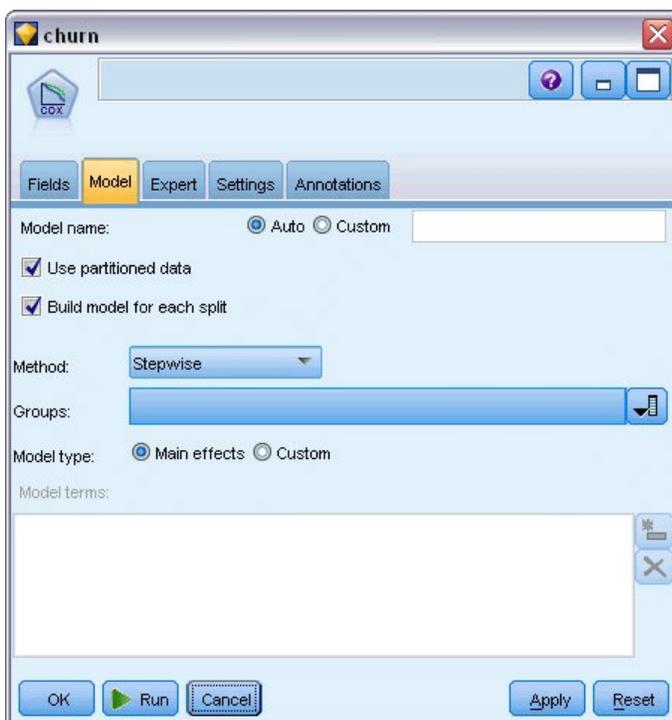


Abbildung 347. Auswählen der Modelloptionen

8. Klicken Sie auf die Registerkarte **Experten** und wählen Sie **Experten** aus, um die Expertenmodellierungsoptionen zu aktivieren.
9. Klicken Sie auf **Ausgabe**.

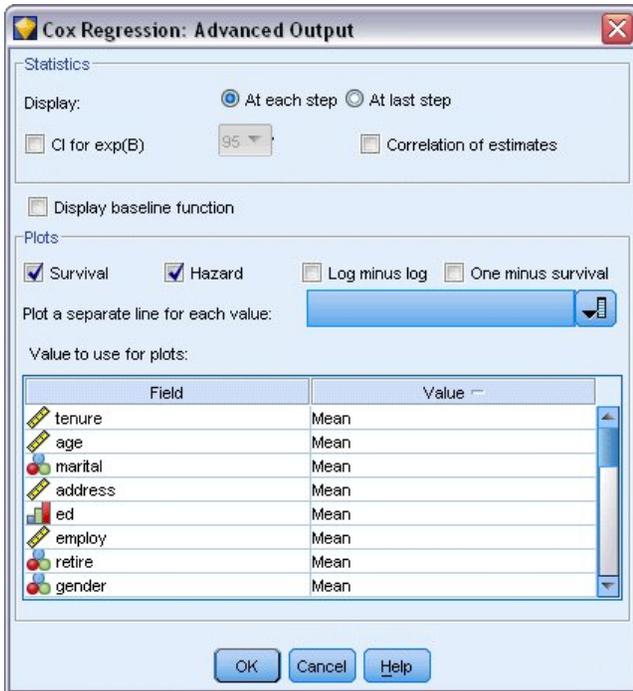


Abbildung 348. Auswahl der erweiterten Ausgabeoptionen

10. Wählen Sie die Optionen **Überleben** und **Hazard** als zu erstellende Diagramme aus und klicken Sie dann auf **OK**.
11. Klicken Sie auf **Ausführen**, um das Modellnugget 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

		N	Percent
Cases available in analysis	Event ^a	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

Abbildung 349. Zusammenfassung der Fallverarbeitung

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.

Codierungen für kategoriale Variablen

	Frequency	(1) ^b	(2)	(3)	(4)
marital ^a	0=Unmarried	505	1		
	1=Married	495	0		
ed ^a	1=Did not complete high school	204	1	0	0
	2=High school degree	287	0	1	0
	3=Some college	209	0	0	1
	4=College degree	234	0	0	0
	5=Post-undergraduate degree	66	0	0	0
retire ^a	.00=No	953	1		
	1.00=Yes	47	0		
gender ^a	0=Male	483	1		
	1=Female	517	0		
tollfree ^a	0=No	526	1		
	1=Yes	474	0		
equip ^a	0=No	614	1		
	1=Yes	386	0		
callcard ^a	0=No	322	1		
	1=Yes	678	0		
wireless ^a	0=No	704	1		
	1=Yes	296	0		
multiline ^a	0=No	525	1		
	1=Yes	475	0		
voice ^a	0=No	696	1		
	1=Yes	304	0		
pager ^a	0=No	739	1		
	1=Yes	261	0		
internet ^a	0=No	632	1		
	1=Yes	368	0		
callid ^a	0=No	519	1		
	1=Yes	481	0		
callwait ^a	0=No	515	1		
	1=Yes	485	0		
forward ^a	0=No	507	1		
	1=Yes	493	0		
confer ^a	0=No	498	1		
	1=Yes	502	0		
ebill ^a	0=No	629	1		
	1=Yes	371	0		
custcat ^a	1=Basic service	266	1	0	0
	2=E-service	217	0	1	0
	3=Plus service	281	0	0	1
	4=Total service	236	0	0	0

Abbildung 350. Codierungen für kategoriale Variablen

Die Codierungen für kategoriale Variablen sind eine nützliche Referenz für die Interpretation der Regressionskoeffizienten für kategoriale Kovariaten, insbesondere dichotome Variablen. Standardmäßig ist die Referenzkategorie die "letzte" Kategorie. So weisen Kunden mit dem Status *Married* (Verheiratet) zwar möglicherweise den Variablenwert "1" in der Datendatei auf, zum Zwecke der Regression werden sie jedoch als "0" codiert.

Variablenauswahl

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1 ^a	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2 ^b	3087.314	249.392	2	.000	305.222	1	.000	439.050	2	.000
3 ^c	3027.085	328.426	3	.000	60.229	1	.000	499.279	3	.000
4 ^d	2990.790	347.197	4	.000	36.294	1	.000	535.574	4	.000
5 ^e	2973.790	362.673	5	.000	17.000	1	.000	552.574	5	.000
6 ^f	2958.796	376.140	6	.000	14.994	1	.000	567.568	6	.000
7 ^g	2945.503	384.717	7	.000	13.293	1	.000	580.861	7	.000
8 ^h	2936.993	417.341	8	.000	8.510	1	.004	589.371	8	.000
9 ⁱ	2926.000	423.911	9	.000	10.994	1	.001	600.364	9	.000
10 ^j	2917.551	428.078	10	.000	8.449	1	.004	608.813	10	.000
11 ^k	2913.308	436.837	11	.000	4.243	1	.039	613.056	11	.000
12 ^l	2908.078	440.158	12	.000	5.230	1	.022	618.286	12	.000

- a. Variable(s) Entered at Step Number 1: callcard
b. Variable(s) Entered at Step Number 2: longmon
c. Variable(s) Entered at Step Number 3: equip
d. Variable(s) Entered at Step Number 4: employ
e. Variable(s) Entered at Step Number 5: multiline
f. Variable(s) Entered at Step Number 6: voice
g. Variable(s) Entered at Step Number 7: address
h. Variable(s) Entered at Step Number 8: equipmon
i. Variable(s) Entered at Step Number 9: ebill
j. Variable(s) Entered at Step Number 10: callid
k. Variable(s) Entered at Step Number 11: internet
l. Variable(s) Entered at Step Number 12: reside
m. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
n. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Abbildung 351. Omnibus-Tests

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.

		B	SE	Wald	df	Sig.	Exp(B)
Step 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
	multiline	.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

Abbildung 352. Variablen in der Gleichung (nur Schritt 12)

Das endgültige Modell enthält die Variablen *address*, *employ*, *reside*, *equip*, *callcard*, *longmon*, *equipmon*, *multiline*, *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 $\text{Exp}(B)$ für *address* bedeutet, dass die Abwanderungs-Hazard-Rate sich für jedes Jahr, in dem ein Kunde dieselbe Adresse hatte, um $100\% - (100\% \times 0,966) = 3,4\%$ 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 $\text{Exp}(B)$ für *callcard* bedeutet, dass die Abwanderungs-Hazard-Rate für einen Kunden, der den Telefonkarten-Service nicht abonniert hat, 2,175-mal so hoch ist wie die eines Kunden mit dem Service. Wir erinnern uns aus den Codierungen für die kategorialen Variablen, dass für die Regression gilt: *No* (Nein) = 1.
- Der Wert von $\text{Exp}(B)$ für *internet* bedeutet, dass die Abwanderungs-Hazard-Rate für einen Kunden, der den Internet-Service nicht 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.

		Score	df	Sig.
Step 12	age	.122	1	.726
	marital	.648	1	.421
	income	1.476	1	.224
	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
	custcat	.864	3	.834
custcat(1)	.466	1	.495	
custcat(2)	.450	1	.502	
custcat(3)	.019	1	.889	

Abbildung 353. Nicht im Modell verwendete Variablen (nur Schritt 12)

Die nicht im Modell verwendeten Variablen weisen jeweils Scorestatistiken 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

	Mean
age	41.684
marital	.505
address	11.551
income	77.535
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
multiline	.525
voice	.696
pager	.739
internet	.632
callid	.519
callwait	.515
forward	.507
confer	.498
ebill	.629
custcat(1)	.266
custcat(2)	.217
custcat(3)	.281

Abbildung 354. Mittelwerte von Kovariaten

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

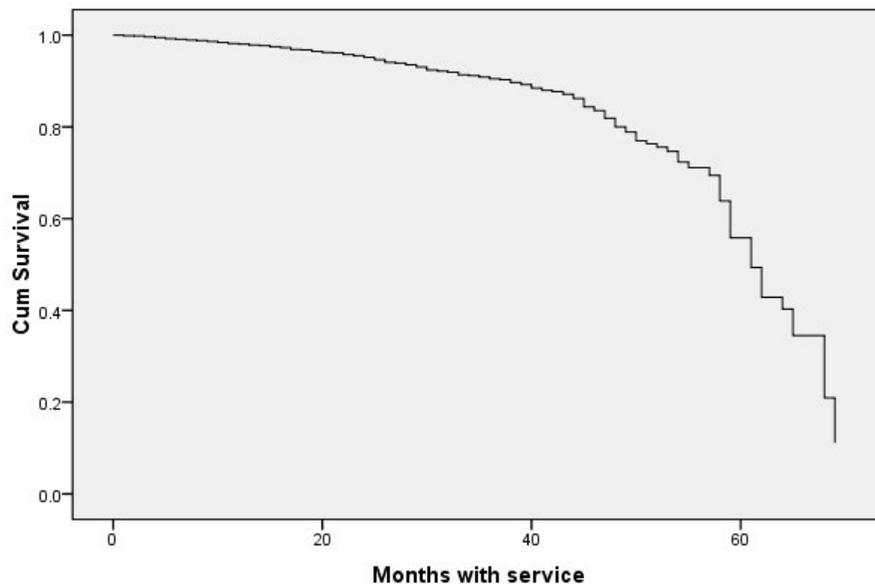


Abbildung 355. Überlebenskurve für den "durchschnittlichen" Kunden

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 stufenförmig.

Hazard-Kurve

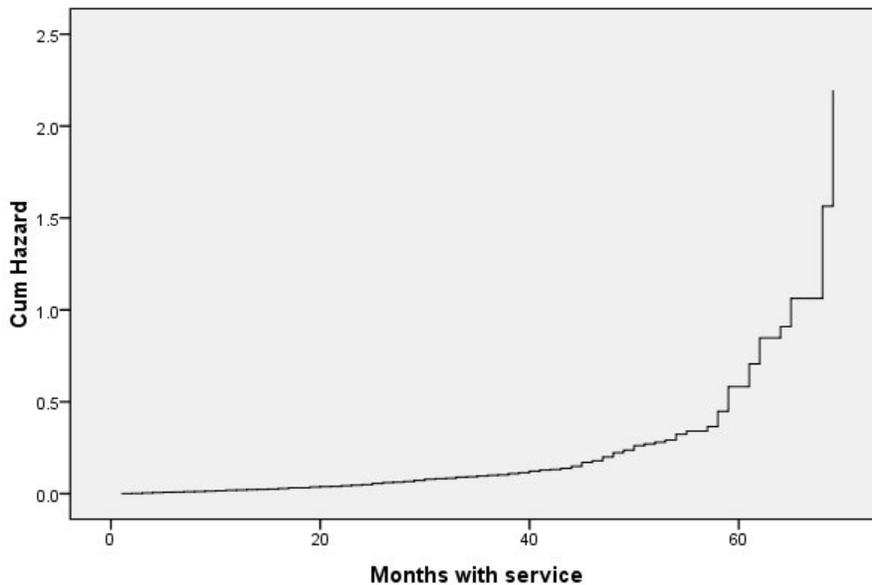


Abbildung 356. Hazard-Kurve für den "durchschnittlichen" Kunden

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.

Evaluation

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.

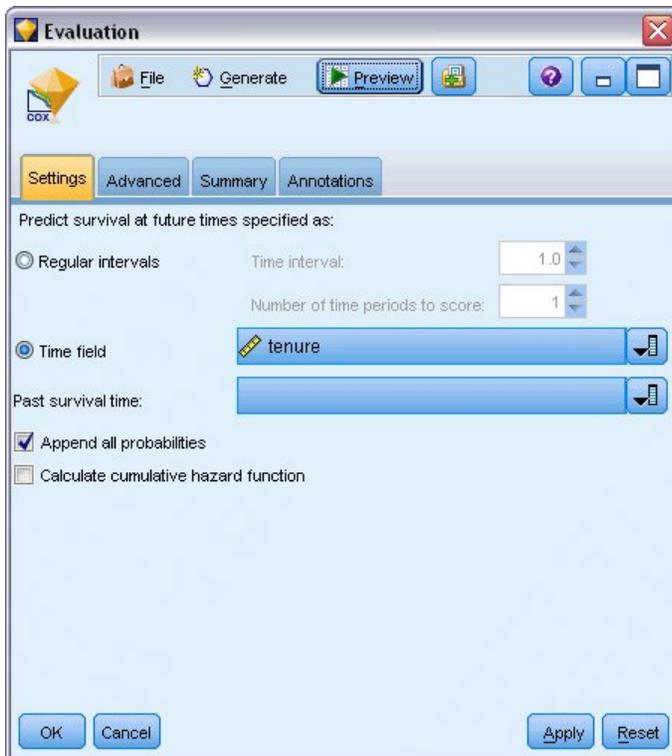


Abbildung 357. Cox-Nugget: Registerkarte "Einstellungen"

1. Platzieren Sie das Modellnugget im Erstellungsbereich und verbinden Sie es mit dem Quellenknoten, öffnen Sie das Nugget und klicken Sie auf die Registerkarte "Einstellungen".
2. Wählen Sie die Option **Zeitfeld** aus und geben Sie *tenure* an. Jeder Datensatz wird als Dauer des Kundenverhältnisses gescort.
3. Wählen Sie die Option **Alle Wahrscheinlichkeiten ausgeben** aus.

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 Evaluierungsknoten hilfreich sein.

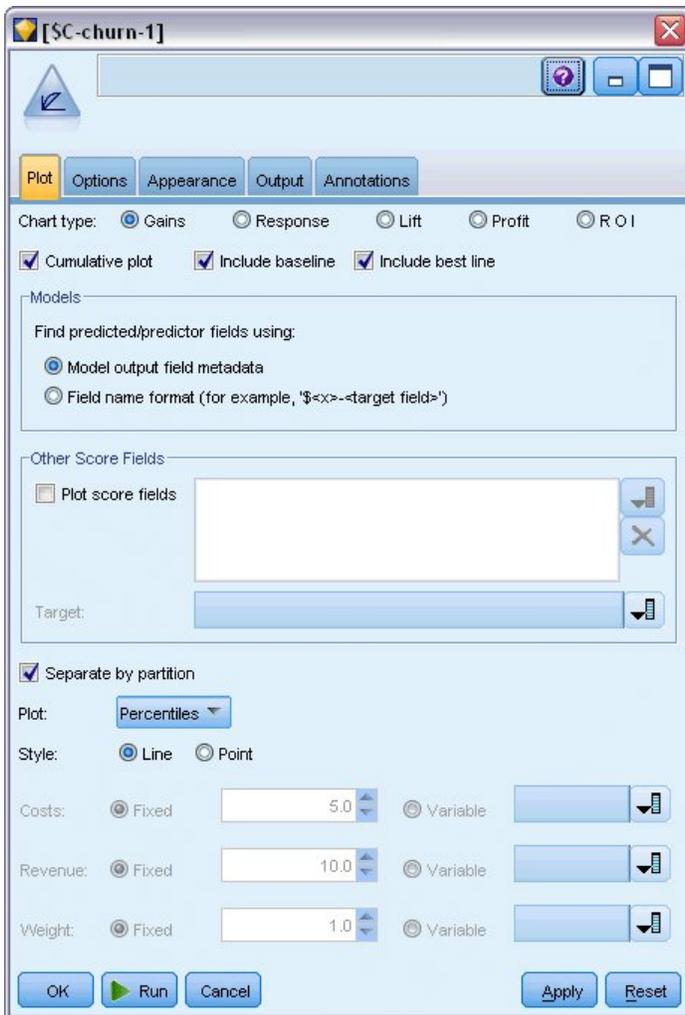


Abbildung 358. Evaluierungsknoten: Registerkarte "Plot"

4. Fügen Sie einen Evaluierungsknoten an das Modellnugget an. Wählen Sie auf der Registerkarte "Plot" (Diagramm) die Option **Beste Linie einschließen** aus.
5. Klicken Sie auf die Registerkarte **Optionen**.

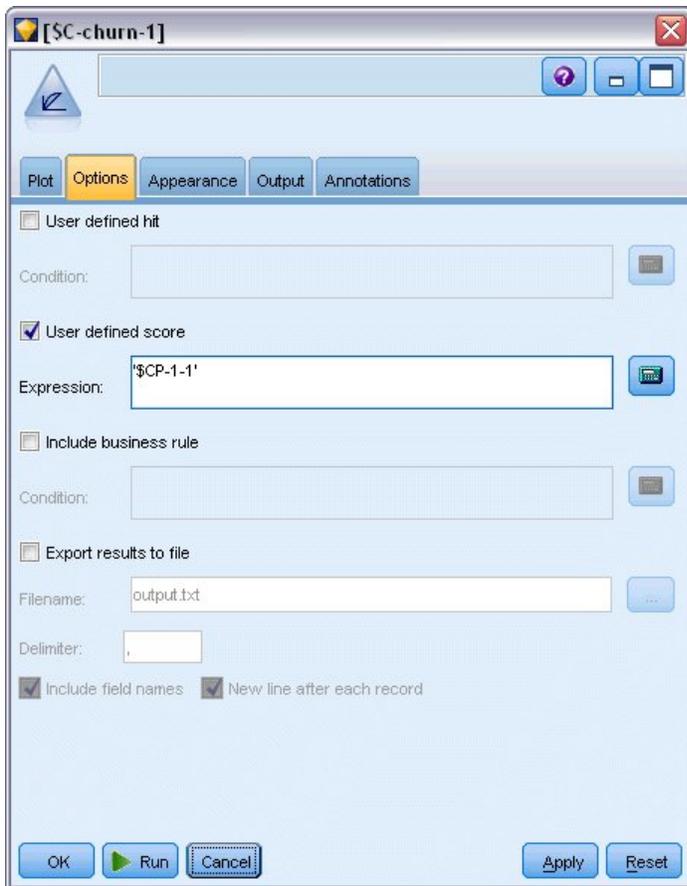


Abbildung 359. Evaluierungsknoten: Registerkarte "Optionen"

6. Wählen Sie die Option **Benutzerdefinierter Score** aus und geben Sie '\$CP-1-1' als Ausdruck ein. Dies ist ein vom Modell generiertes Feld, das der Abwanderungsneigung entspricht.
7. Klicken Sie auf **Ausführen**.

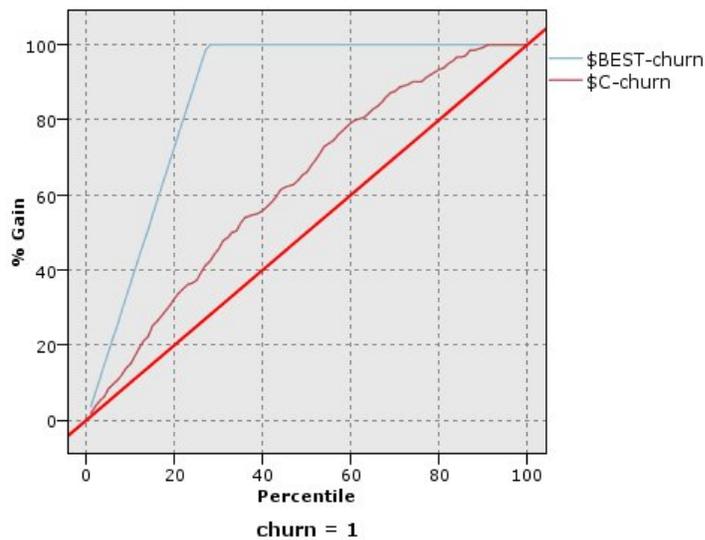


Abbildung 360. Gewinnndiagramm

Das kumulative Gewinnndiagramm 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 Dataset 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 Datasets auswählen, erhalten Sie alle abwanderungswilligen Personen im Dataset.

Die diagonale Linie ist die "Basiskurve": Wenn Sie nach dem Zufallsprinzip 20 % der Datensätze im gescorten Dataset auswählen, ist zu erwarten, dass Sie ungefähr 20 % aller Datensätze "gewinnen", die tatsächlich in Kategorie 1 fallen. Je weiter eine Kurve über der Basiskurve liegt, desto höher 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 Gewinnndiagramm 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 Gewinnndiagramm 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.

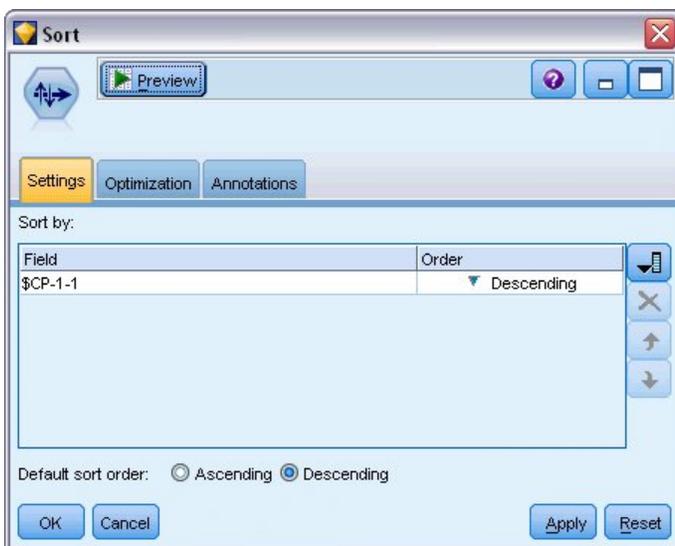


Abbildung 361. Sortierknoten: Registerkarte "Einstellungen"

- 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, fügen Sie einen Sortierknoten an das Modellnugget an.

9. 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**.

id	\$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

Abbildung 362. Tabelle

10. Verbinden Sie einen Tabellenknoten mit dem Sortierknoten.
 11. 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 Dataset 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.

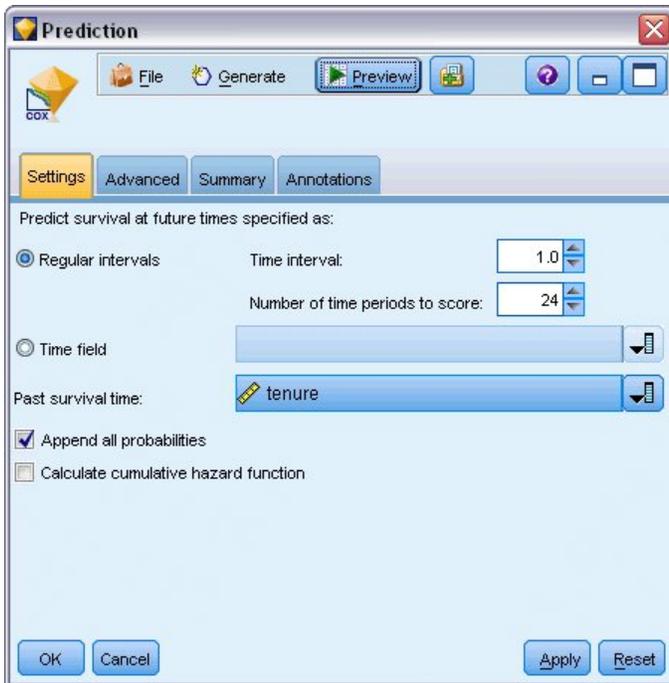


Abbildung 363. Cox-Nugget: Registerkarte "Einstellungen"

1. Doppelklicken Sie auf das Modellnugget in der Modellpalette (oder kopieren Sie das Nugget und fügen Sie es im Streamerstellungsbereich ein) und fügen Sie das neue Nugget an den Quellenknoten an.
2. Öffnen Sie die Registerkarte "Einstellungen" für das Nugget.
3. 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.
4. Wählen Sie *tenure* als Feld zur Angabe der vergangenen Überlebenszeit aus. Der Scoring-Algorithmus berücksichtigt die Zeitdauer, die jede Person Kunde des Unternehmens war.
5. Wählen Sie die Option **Alle Wahrscheinlichkeiten ausgeben** aus.

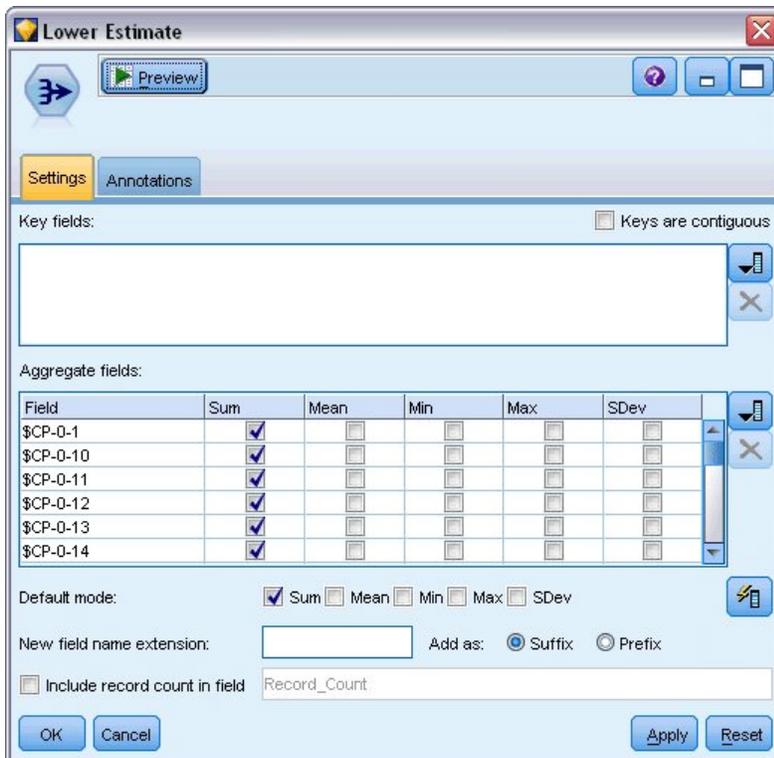


Abbildung 364. Aggregatnoten: Registerkarte "Einstellungen"

6. Fügen Sie einen Aggregatnoten an das Modellnugget an. Heben Sie auf der Registerkarte "Einstellungen" die Auswahl der Option **Mittelwert** als Standardmodus auf.
7. 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.
8. Heben Sie die Auswahl von **Datensatzanzahl einschließen in Feld** auf.
9. Klicken Sie auf **OK**. Dieser Knoten erstellt die Vorhersagen für die "Untergrenze".

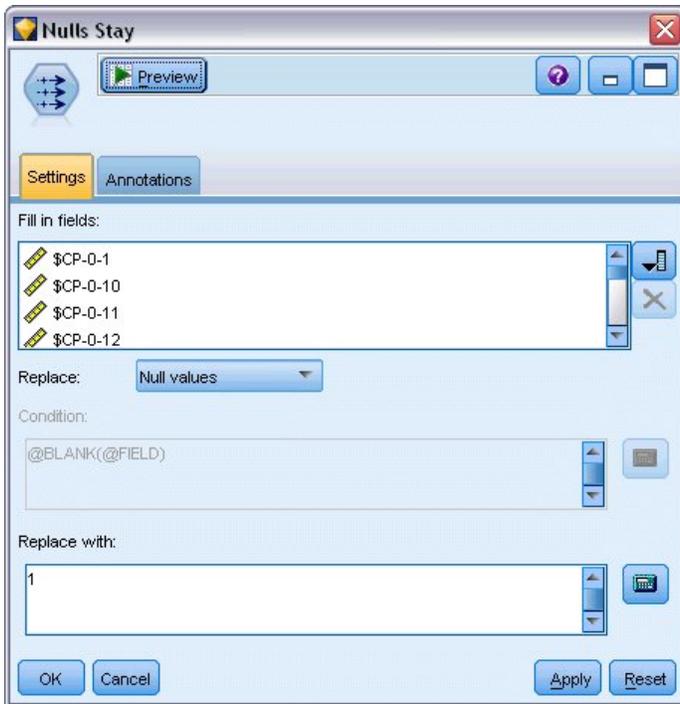


Abbildung 365. Füllerknoten: Registerkarte "Einstellungen"

10. Fügen Sie einen Füllerknoten an das Coxreg-Nugget an, das Sie soeben an den Aggregatknoten angehängt 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.
11. Legen Sie fest, dass **Nullwerte** durch den Wert 1 ersetzt werden sollen.
12. Klicken Sie auf **OK**.



Abbildung 366. Aggregatnoten: Registerkarte "Einstellungen"

13. Fügen Sie einen Aggregatnoten an den Füllerknoten an. Heben Sie auf der Registerkarte "Einstellungen" die Auswahl der Option **Mittelwert** als Standardmodus auf.
14. 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.
15. Heben Sie die Auswahl von **Datensatzanzahl einschließen in Feld** auf.
16. Klicken Sie auf **OK**. Dieser Knoten erstellt die Vorhersagen für die "Obergrenze".

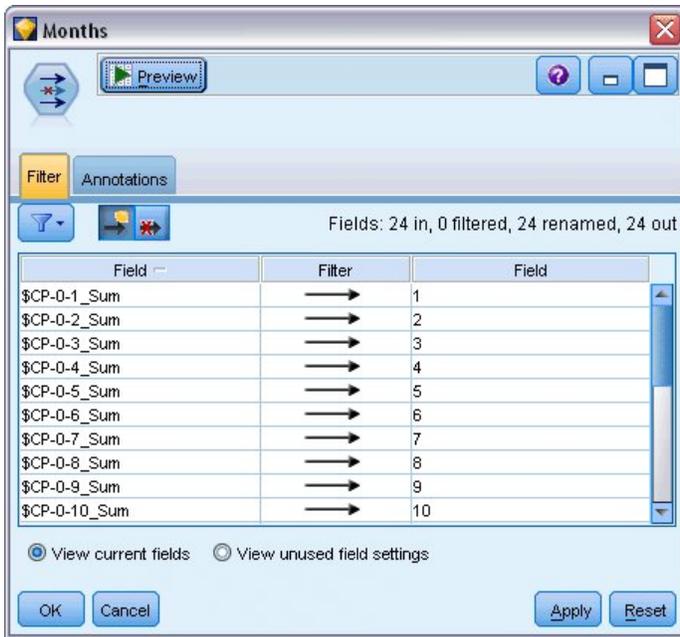


Abbildung 367. Filterknoten: Registerkarte "Einstellungen"

17. Fügen Sie einen Anhangknoten an die beiden Aggregatknöten an und fügen Sie anschließend einen Filterknoten an den Anhangknoten an.
18. Benennen Sie auf der Registerkarte "Einstellungen" des Filterknoten die Felder in 1 bis 24 um. Durch die Verwendung eines Transponierknoten werden diese Feldnamen zu Werten für die x -Achse in Diagrammen weiter unten im Stream.



Abbildung 368. Transponierknoten: Registerkarte "Einstellungen"

19. Fügen Sie einen Transponierknoten an den Filterknoten an.
20. Wählen Sie 2 als Anzahl der neuen Felder aus.

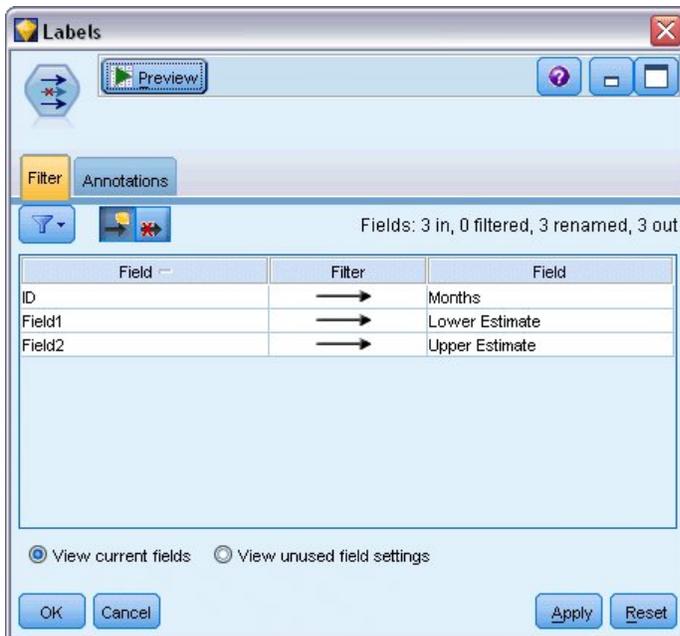


Abbildung 369. Filterknoten: Registerkarte "Filter"

21. Fügen Sie einen Filterknoten an den Transponierknoten an.

22. Benennen Sie auf der Registerkarte "Einstellungen" des Filterknotens *ID* in *Months* (Monate), *Feld1* in *Lower Estimate* (Untere Schätzung) und *Feld2* in *Upper Estimate* (Obere Schätzung) um.

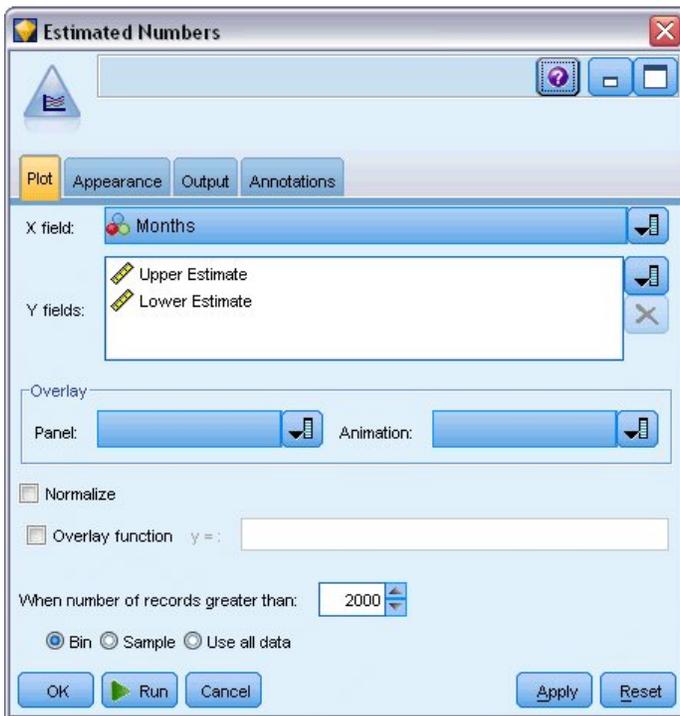


Abbildung 370. Multiplotknoten: Registerkarte "Plot"

23. Fügen Sie einen Multiplotknoten an den Filterknoten an.
24. Wählen Sie auf der Registerkarte "Plot" *Months* (Monate) als X-Feld und *Lower Estimate* (Untere Schätzung) und *Upper Estimate* (Obere Schätzung) als Y-Feld aus.

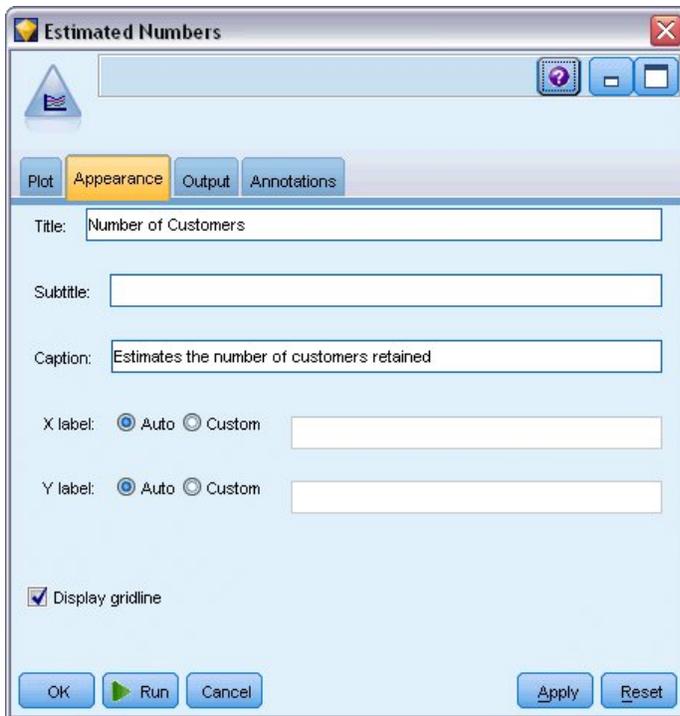
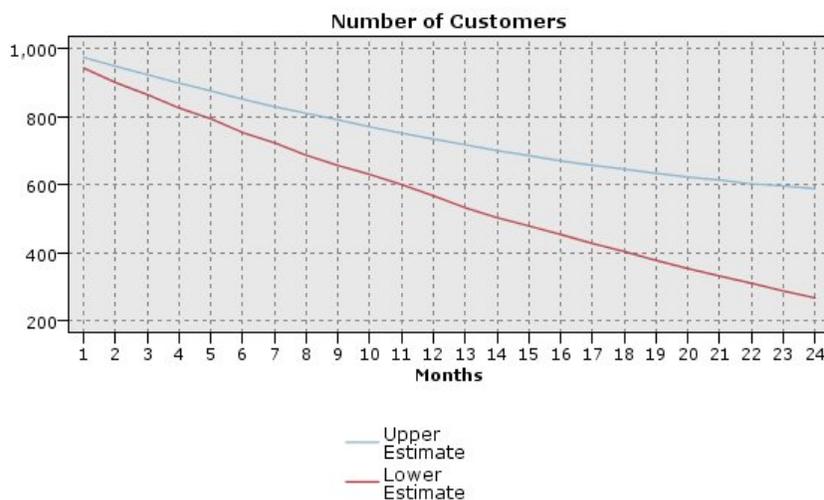


Abbildung 371. Multiplotknoten: Registerkarte "Darstellung"

25. Klicken Sie auf die Registerkarte "Darstellung".
26. Geben Sie Number of Customers (Anzahl der Kunden) als Titel ein.
27. Geben Sie Schätzt die Anzahl der gehaltenen Kunden als Titelzeile ein.
28. Klicken Sie auf **Ausführen**.



Estimates the number of customers retained

Abbildung 372. Multiplot 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 ursprüngli-

chen Kunden im Dataset noch erhalten geblieben sind; nach 24 Monaten liegt dieser Wert zwischen 288 und 597.

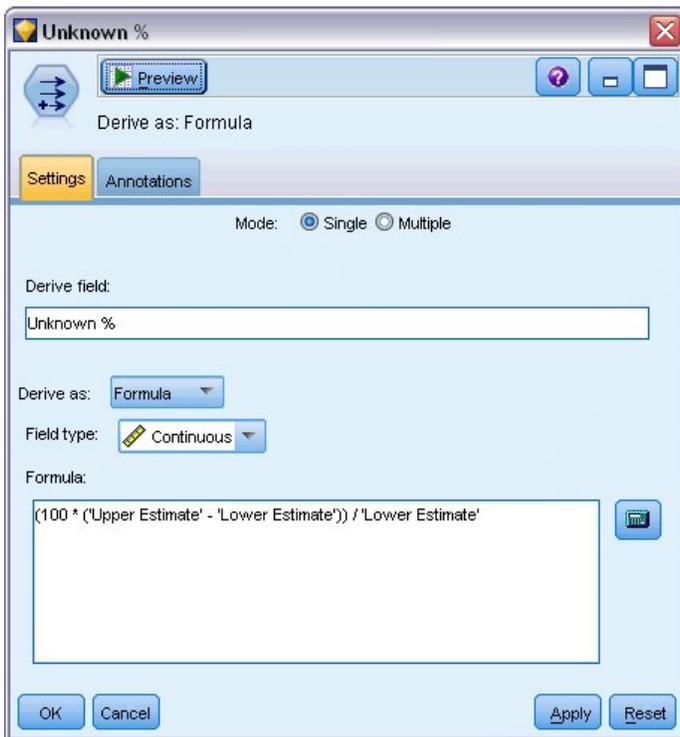


Abbildung 373. Ableitungsknoten: Registerkarte "Einstellungen"

29. Um einen weiteren Einblick darin zu erhalten, wie unsicher die Schätzungen für die Anzahl der gehaltenen Kunden sind, fügen Sie einen Ableitungsknoten an den Filterknoten an.
30. Geben Sie auf der Registerkarte "Einstellungen" des Ableitungsknotens *Unknown %* (% unbekannt) als Ableitungsfeld ein.
31. Wählen Sie **Stetig** als Feldtyp aus.
32. 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 der unteren Schätzung an.
33. Klicken Sie auf **OK**.

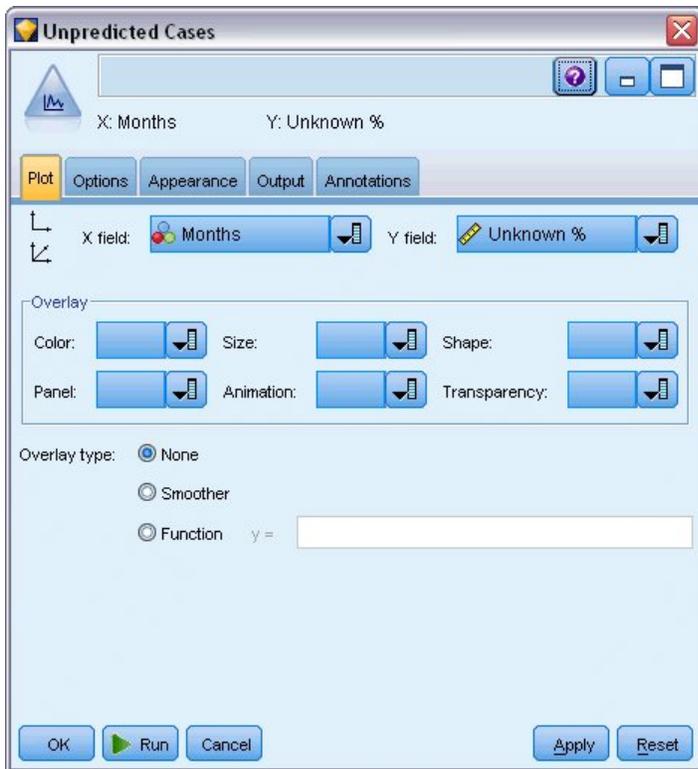


Abbildung 374. Plotknoten: Registerkarte "Plot"

34. Fügen Sie einen Plotknoten an den Ableitungsknoten an.
35. Wählen Sie auf der Registerkarte "Plot" (Diagramm) des Plotknotens *Months* (Monate) als X-Feld und *Unknown %* (% unbekannt) als Y-Feld aus.
36. Klicken Sie auf die Registerkarte **Darstellung**.

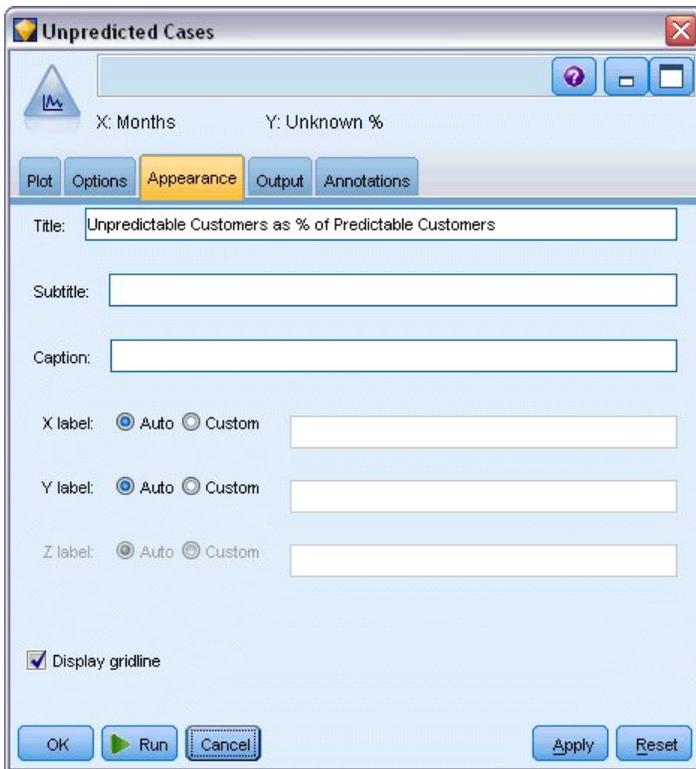


Abbildung 375. Plotknoten: Registerkarte "Darstellung"

37. Geben Sie Unpredictable Customers as % of Predictable Customers (Unvorhersagbare Kunden als % der vorhersagbaren Kunden) als Titel ein.
38. Führen Sie den Knoten aus.

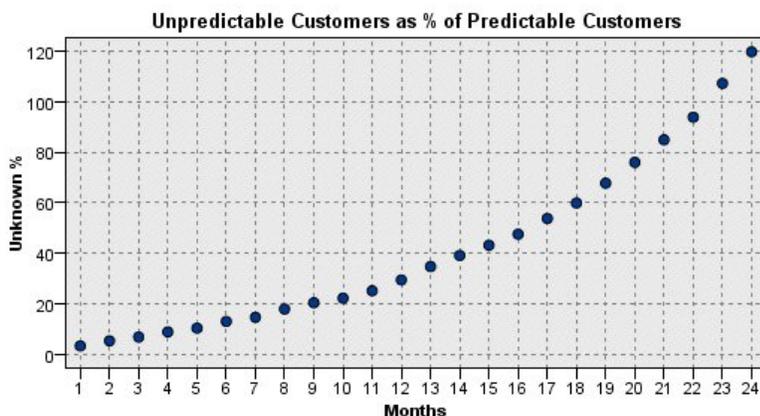


Abbildung 376. Plot der unvorhersagbaren Kunden

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.

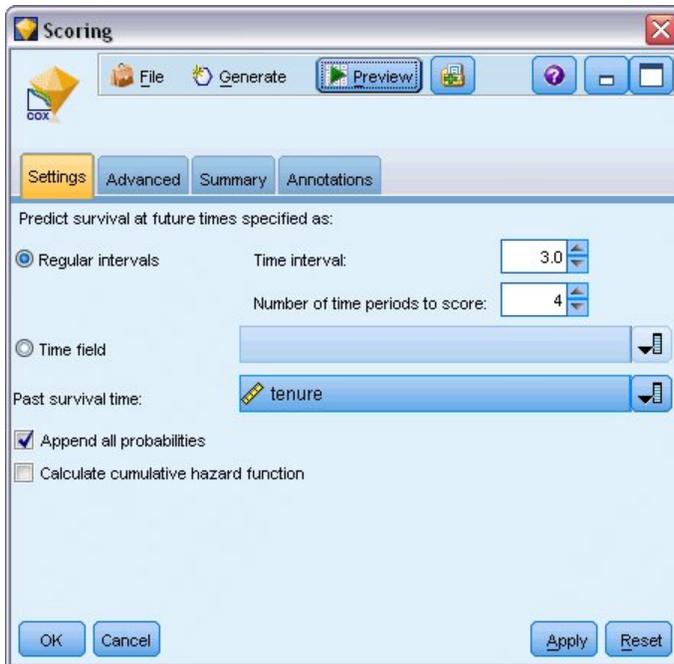


Abbildung 377. Coxreg-Nugget: Registerkarte "Einstellungen"

1. Fügen Sie ein drittes Modellnugget an Quellenknoten an und öffnen Sie das Modellnugget.
2. 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.
3. Wählen Sie *tenure* als Feld zur Angabe der vergangenen Überlebenszeit aus. Der Scoring-Algorithmus berücksichtigt die Zeitdauer, die jede Person Kunde des Unternehmens war.
4. Wählen Sie die Option **Alle Wahrscheinlichkeiten ausgeben** aus. Diese zusätzlichen Felder erleichtern die Sortierung der Datensätze zur Anzeige in einer Tabelle.

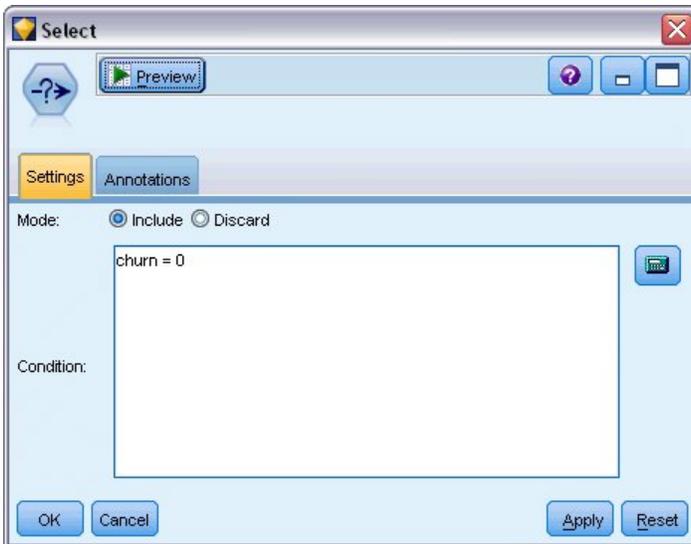


Abbildung 378. Auswahlknoten: Registerkarte "Einstellungen"

5. Fügen Sie einen Auswahlknoten an das Modellnugget an. Geben Sie auf der Registerkarte "Einstellungen" $churn=0$ als Bedingung ein. Dadurch werden Kunden, die bereits abgewandert sind, aus der Ergebnistabelle entfernt.

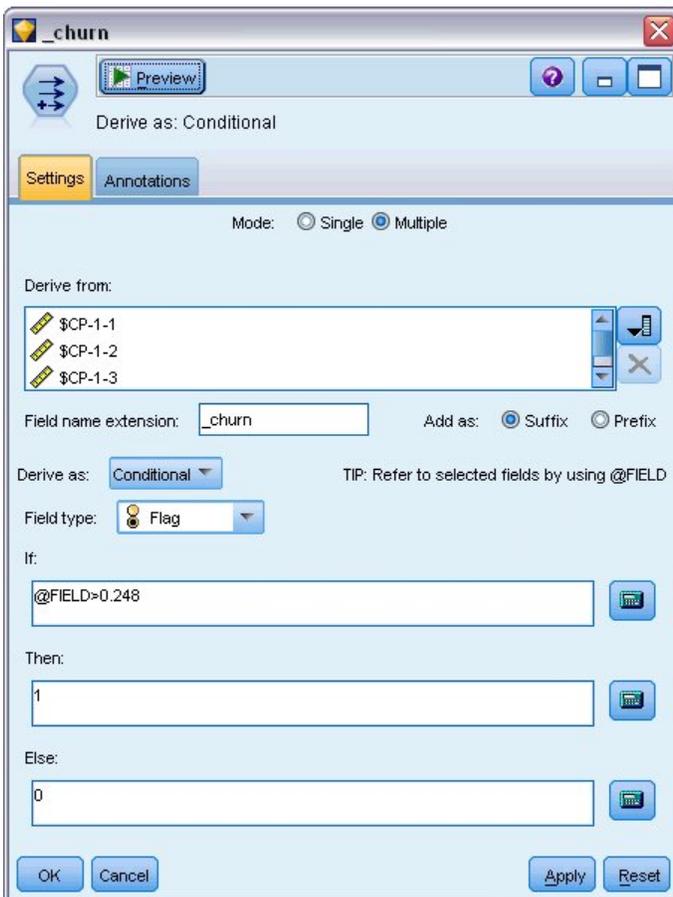


Abbildung 379. Ableitungsknoten: Registerkarte "Einstellungen"

6. Fügen Sie einen Ableitungsknoten an den Auswahlknoten an. Wählen Sie auf der Registerkarte "Einstellungen" **Mehrere** als Modus aus.
7. 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.
8. Wählen Sie für die Ableitung des Felds die Option **Bedingt** aus.
9. Wählen Sie **Flag** als Messniveau aus.
10. Geben Sie `@FIELD>0,248` als **Wenn**-Bedingung ein. Wie Sie sich erinnern, war dies der während der Evaluation ermittelte Trennwert für die Klassifizierung.
11. Geben Sie 1 als **Dann**-Ausdruck ein.
12. Geben Sie 0 als **Sonst**-Ausdruck ein.
13. Klicken Sie auf **OK**.

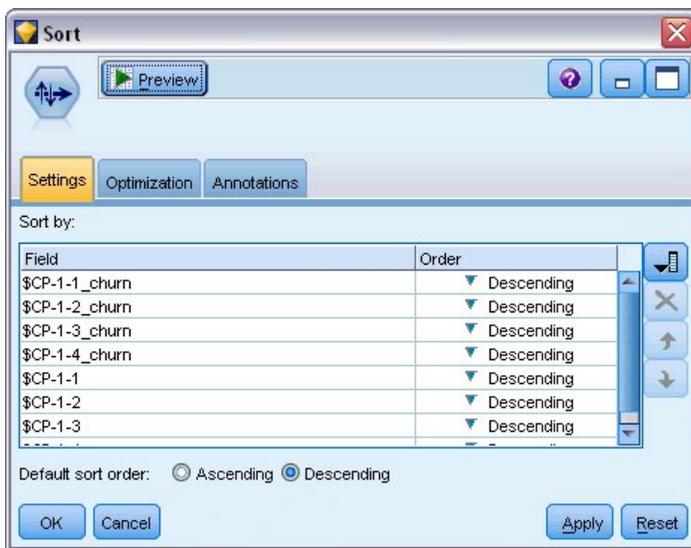


Abbildung 380. Sortierknoten: Registerkarte "Einstellungen"

14. Fügen 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.

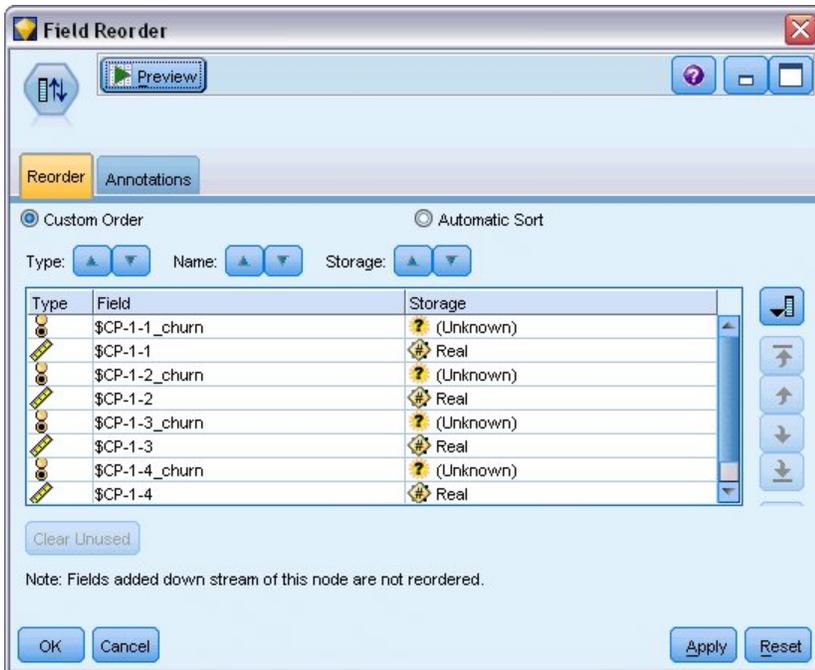


Abbildung 381. Knoten "Felder ordnen": Registerkarte "Ordnen"

- Fügen 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.

The screenshot shows a software window titled "Table (50 fields, 726 records)". The window contains a data table with the following columns: \$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, and tenur. The rows are numbered from 255 to 274. The table shows churn rates for different quarters and the corresponding tenure for each customer.

	\$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

Abbildung 382. Tabelle mit Kundenscores

16. Fügen 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.

	\$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

Abbildung 383. Tabelle mit Kunden mit Nullwerten

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 *IBM SPSS Modeler-Algorithmushandbuch* aufgeführt.

Kapitel 27. Warenkorbanalyse (Regelinduktion/C5.0)

In diesem Beispiel werden frei erfundene Daten verwendet, die den Inhalt von Supermarkt-Warenkörben (d. h. eine Sammlung von gekauften Produkten) sowie die verknüpften persönlichen Daten des Käufers beschreiben, 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 Netzdiagrammanzeige, 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 auf die Datendatei *BASKETS1n* nimmt. Die Dateien stehen im Verzeichnis *Demos* der IBM SPSS Modeler-Installation zur Verfügung. Der Zugriff über die Programmgruppe "IBM SPSS Modeler" ist im Start-Menü von Windows möglich. Die Datei *baskrule* befindet sich im Verzeichnis *streams*.

Datenzugriff

Stellen Sie unter Verwendung des Knotens "Variable Datei" eine Verbindung mit dem Dataset *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 Dataset vorkommt und deshalb für die Modellierung nicht verwendet werden kann). Wählen Sie *Nominal* als Messniveau für das Feld *Geschlecht* aus. (Damit wird sichergestellt, dass der Apriori-Modellierungsalgorithmus *Geschlecht* nicht als Flag behandelt wird.)

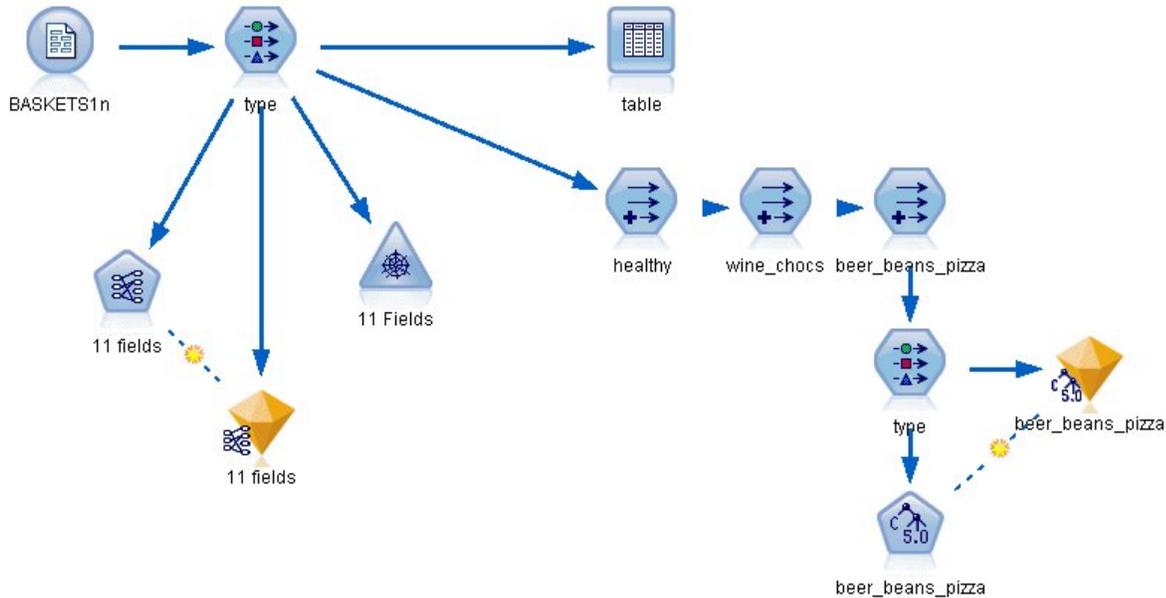


Abbildung 384. *baskrule-Stream*

Führen Sie den Stream jetzt aus, um den Typknoten zu instanziierten und die Tabelle anzuzeigen. Das Dataset enthält 18 Felder, wobei jeder Datensatz einen Korb darstellt.

Die 18 Felder werden in den folgenden Überschriften dargestellt.

Warenkorbübersicht:

- *CardID*. Treuekarten-ID 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 Apriori, um Assoziationsregeln zu erstellen. Wählen Sie die in diesem Modellierungsprozess zu verwendenden Felder aus, 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: Durch Klicken bei gedrückter Umschalttaste können Sie Optionen für mehrere Felder festlegen. So können Sie die Felder auswählen, bevor Sie eine Option aus den Spalten festlegen.



Abbildung 385. Auswählen von Feldern für die Modellierung

Sobald Sie Felder für die Modellierung festgelegt haben, verbinden Sie einen Apriori-Knoten mit dem Typknoten, bearbeiten diesen, wählen die Option **Nur wahre Werte für Flags** aus und führen den Apriori-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.

Consequent	Antecedent	Support %	Confidence %
frozenmeal	beer cannedveg	16.7	87.425
cannedveg	beer frozenmeal	17.0	85.882
beer	frozenmeal cannedveg	17.3	84.393

Abbildung 386. 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 Netzdiagrammanzeige (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.

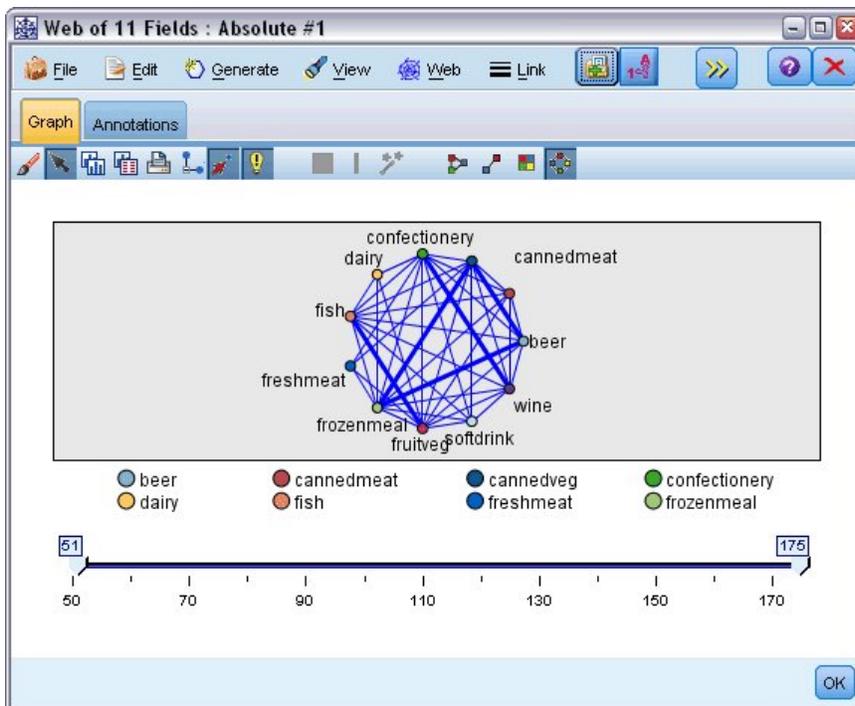


Abbildung 387. Netzdiagrammanzeige der Produktverbindungen

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.

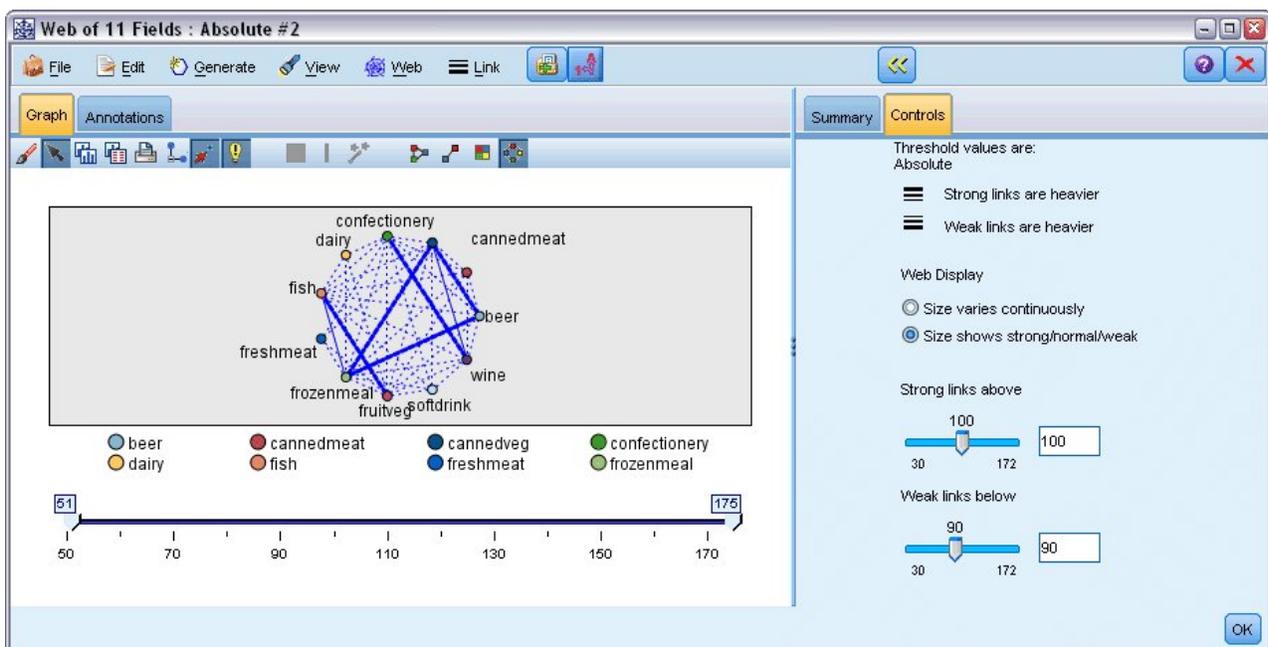


Abbildung 388. Beschränkte Netzdiagrammanzeige

1. 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.

2. Wählen Sie **Größe zeigt stark/mittel/schwach** aus.
3. Legen Sie fest: Schwache Zusammenhänge unter 90.
4. 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 Netzdiagrammanzeige 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** aus.

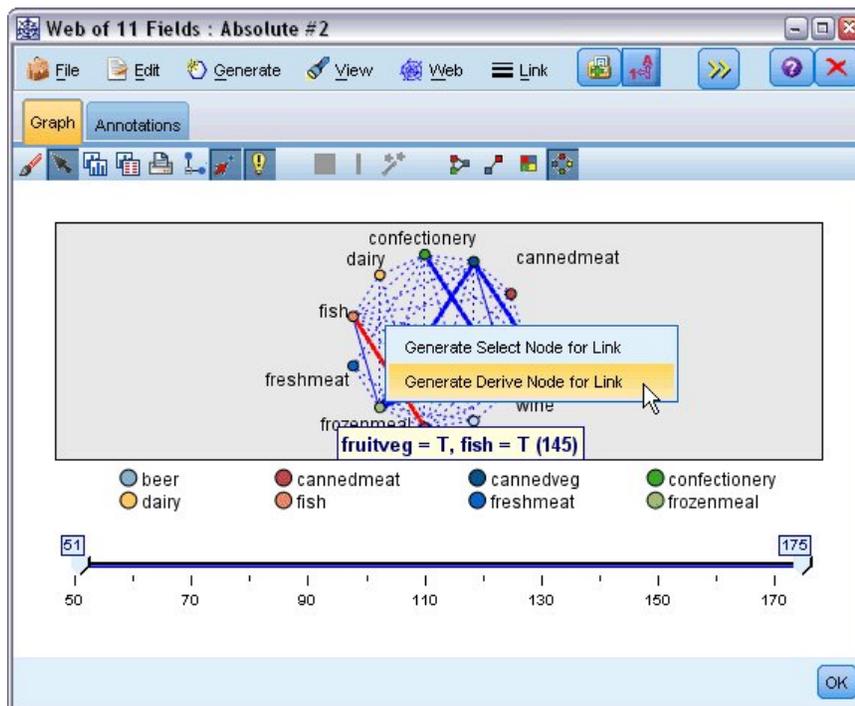


Abbildung 389. Ableiten eines Flags für eine Kundengruppe

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 Netzdiagrammanzeige Folgendes aus:

Generieren > 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 der entsprechenden Kundengruppe (z. B. *Bier_Bohnen_Pizza*), die auf *Ziel* gesetzt werden muss. Fügen Sie einen C5.0-Knoten hinzu, setzen Sie den Ausgabetyt auf **Regelset** 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:  
wenn 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 Apriori anstelle von C5.0 generiert werden; Apriori 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 Apriori) als auch die Visualisierung (mit Netzdiagrammanzeige). Diese Zusammenhänge entsprechen den Gruppierungen von Fällen in den Daten und diese Gruppen können detailliert untersucht und anhand einer Modellierung (mit C5.0-Regelsets) 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 28. Beurteilen neuer Fahrzeugangebote (KNN)

Die Nächste-Nachbarn-Analyse ist eine Methode zur Klassifizierung von Fällen anhand ihrer Ähnlichkeit zu anderen Fällen. Beim maschinellen Lernen wurde sie entwickelt, um Datenmuster zu erkennen, ohne dass eine exakte Übereinstimmung mit gespeicherten Mustern oder Fällen benötigt wird. Ähnliche Fälle liegen nah beieinander und Fälle mit geringer Ähnlichkeit sind weit voneinander entfernt. Somit gilt die Distanz zwischen zwei Fällen als Maß für ihre Unähnlichkeit.

Befinden sich Fälle nahe beieinander, werden sie als "Nachbarn" bezeichnet. Wenn ein neuer Fall (Holdout) angegeben wird, wird seine Distanz zu jedem der Fälle im Modell berechnet. Die Klassifizierungen der ähnlichsten Fälle - die nächsten Nachbarn - werden gezählt und der neue Fall wird einer Kategorie zugeordnet, die die größte Anzahl der nächsten Nachbarn enthält.

Sie können die Anzahl der zu untersuchenden nächsten Nachbarn angeben; dieser Wert wird k genannt. Die Abbildungen zeigen, wie ein neuer Fall mit zwei verschiedenen Werten für k klassifiziert würde. Wenn $k = 5$ ist, wird der neue Fall in Kategorie 1 gesetzt, da die Mehrheit der nächsten Nachbarn zur Kategorie 1 gehört. Wenn jedoch $k = 9$ ist, wird der neue Fall in Kategorie 0 gesetzt, da die Mehrheit der nächsten Nachbarn zur Kategorie 0 gehört.

Die Nächste-Nachbarn-Analyse kann auch zur Berechnung von Werten für ein stetiges Ziel verwendet werden. Dabei wird der durchschnittliche oder Median-Zielwert der nächsten Nachbarn verwendet, um den vorhergesagten Wert für den neuen Fall zu beziehen.

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 feststellen, 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 Unterverzeichnis *streams* verfügbar ist. Als Datendatei wird die Datei *car_sales_knn_mod.sav* verwendet. Weitere Informationen finden Sie unter im Thema „Ordner "Demos"“ auf Seite 5.

Erstellen des Streams

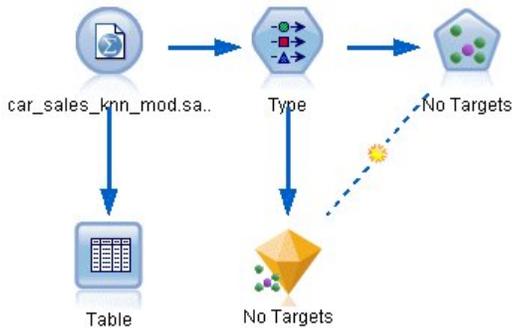


Abbildung 390. Beispielstream für KNN-Modellierung

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.

1. Fügen Sie dem Quellenknoten für Statistikdateien einen Tabellenknoten hinzu.
2. Öffnen Sie den Tabellenknoten und klicken Sie auf **Ausführen**.

	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...

Abbildung 391. Quelldaten für Autos und LKW

Die Details für die beiden Prototypen *newCar* und *newTruck* wurden an das Ende der Datei gefügt. Den Quelldaten 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 Holdouts 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 Festlegen von 1 für den *partition*-Wert der beiden Holdout-Datensätze, während alle anderen Datensätze in diesem Feld 0 enthalten, können wir dieses Feld später beim Festlegen 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.

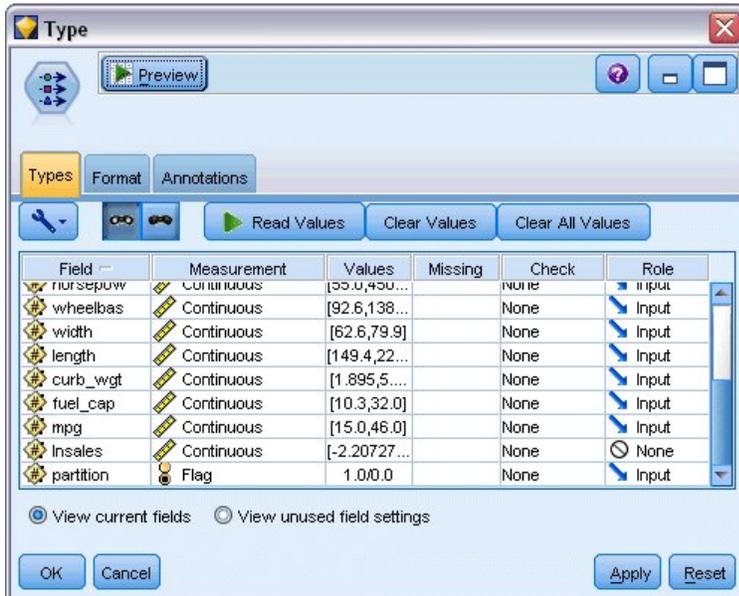


Abbildung 392. Typknoteneinstellungen

3. Fügen Sie dem Stream einen Typknoten hinzu.
4. Fügen Sie dem Quellenknoten für Statistikdateien einen Typknoten hinzu.
5. 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**.

6. Setzen Sie die Rolle für alle anderen Felder (*manufact* bis *type* plus *Insales*) auf **Keine**.
7. Setzen Sie das Messniveau für das letzte Feld (*partition*) auf **Flag**. Stellen Sie sicher, dass seine Rolle auf **Eingabe** eingestellt ist.
8. Klicken Sie auf **Werte lesen**, um die Datenwerte in den Stream einzulesen.
9. Klicken Sie auf **OK**.

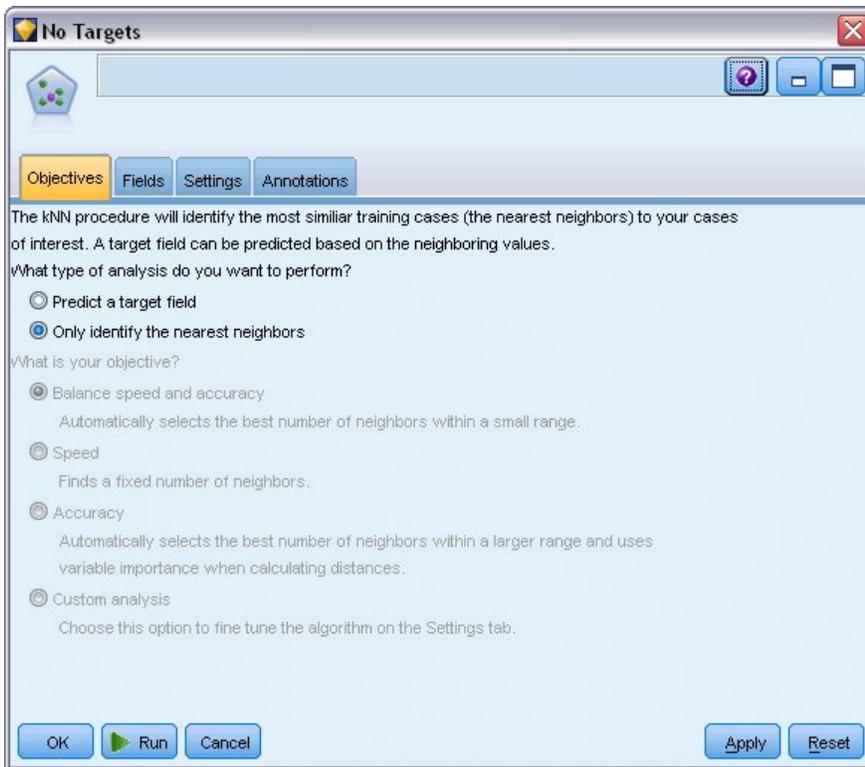


Abbildung 393. Auswahl zur Identifizierung der nächsten Nachbarn

10. Verbinden Sie einen KNN-Knoten mit dem Typknoten.
11. Öffnen Sie den KNN-Knoten.
Diesmal werden wir kein Zielfeld vorhersagen, da wir nur die nächsten Nachbarn für unsere beiden Prototypen finden möchten.
12. Wählen Sie auf der Registerkarte **Ziele** die Option **Nur nächste Nachbarn identifizieren**.
13. Klicken Sie auf die Registerkarte **Einstellungen**.

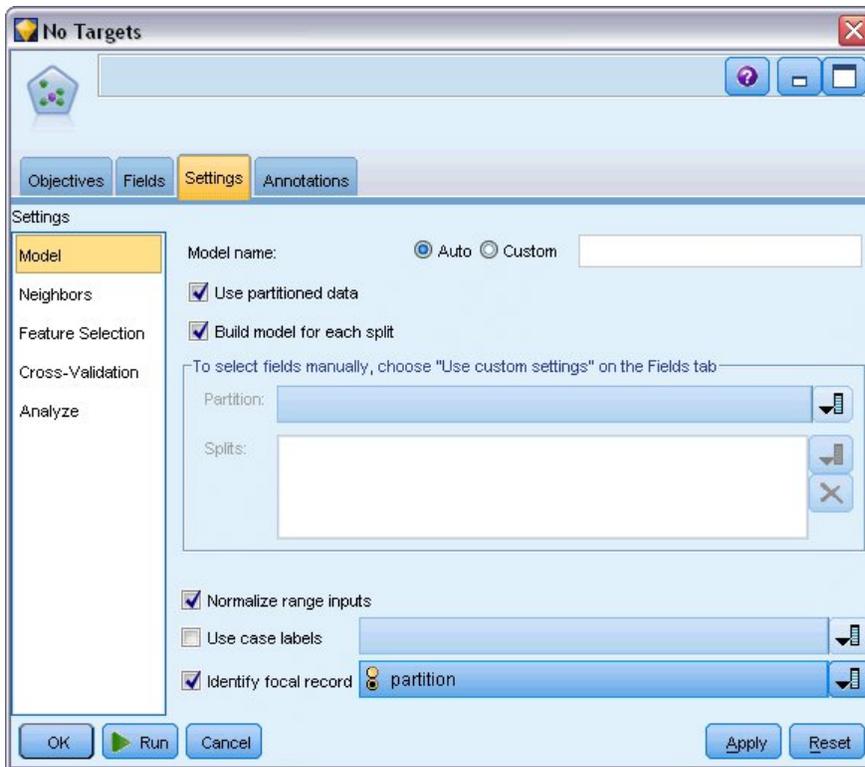


Abbildung 394. Identifizieren der Fokusdatensätze mithilfe des Felds "partition"

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 Flagfelds stellen wir sicher, dass Datensätze, in denen für dieses Feld der Wert 1 festgelegt ist, zu unseren Fokusdatensätzen werden.

Wir haben bereits gesehen, dass *newCar* und *newTruck* als einzige Datensätze den Wert 1 in diesem Feld aufweisen und damit unsere Fokusdatensätze sein werden.

14. Aktivieren Sie im Bereich **Modell** der Registerkarte **Einstellungen** das Kontrollkästchen **Fokusdatensatz identifizieren**.
15. Wählen Sie **partition** aus der Dropdown-Liste für dieses Feld.
16. Klicken Sie auf die Schaltfläche **Ausführen**.

Untersuchen der Ausgabe

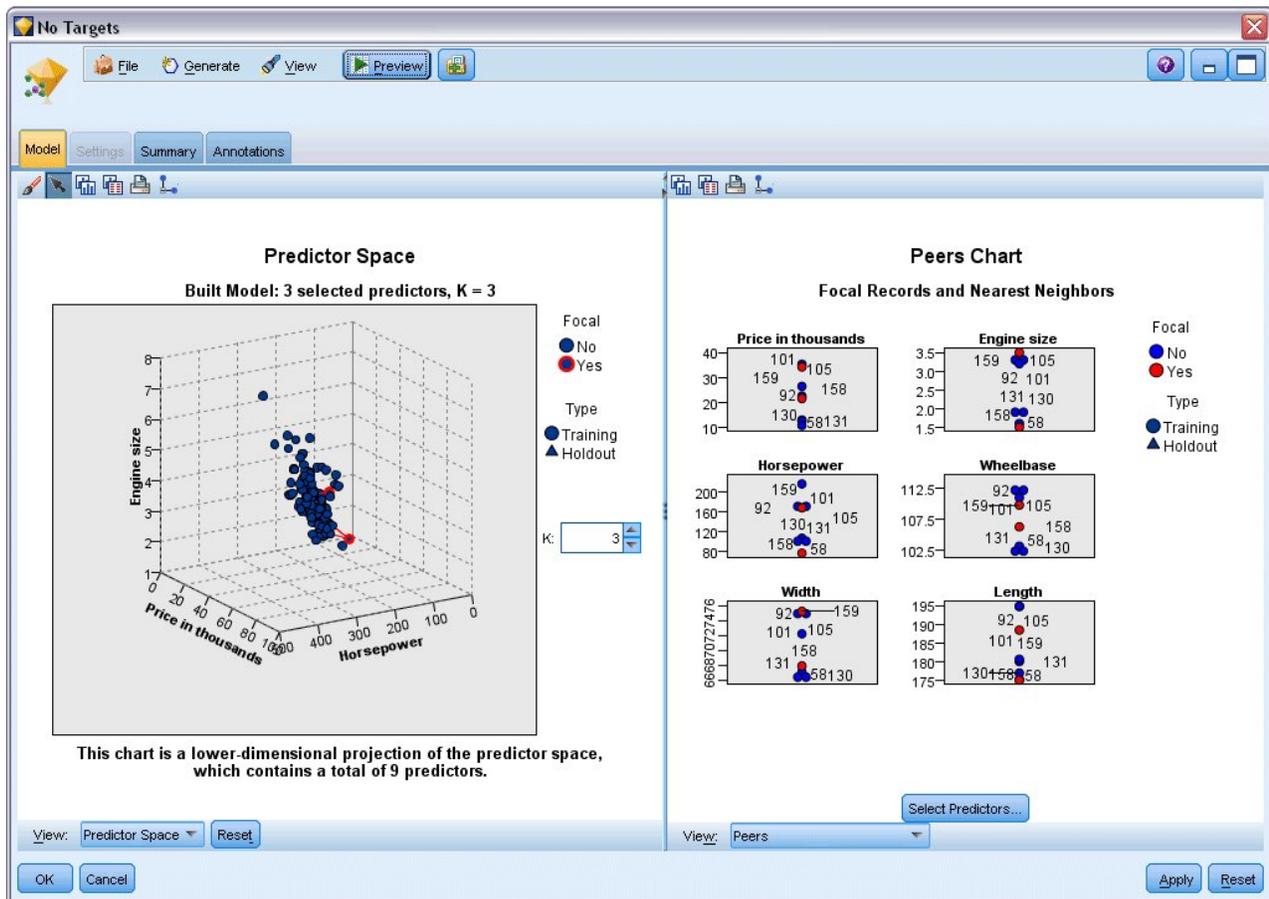
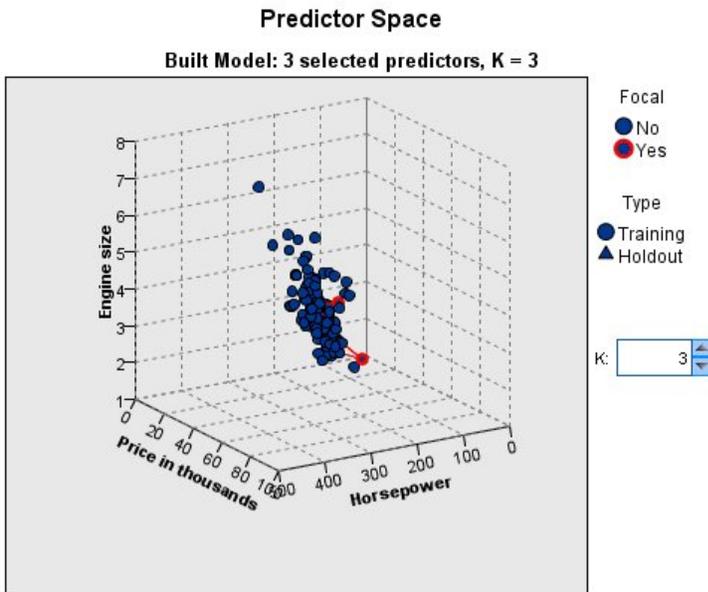


Abbildung 395. Das Fenster "Modellviewer"

Im Streamerstellungsbereich und in der Modellalette wurde ein Modellnugget erstellt. Öffnen Sie eines der Nuggets, um die Anzeige des Modellviewers aufzurufen, 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 einem bestimmten Merkmal des Modells an, wenn Sie einen Teil der Hauptansicht ansteuern.

Prädiktorbereich



This chart is a lower-dimensional projection of the predictor space, which contains a total of 9 predictors.

Abbildung 396. Prädiktorbereichsdiagramm

Dieses Prädiktorbereichsdiagramm ist ein interaktives 3-D-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.

Peerdiagramm

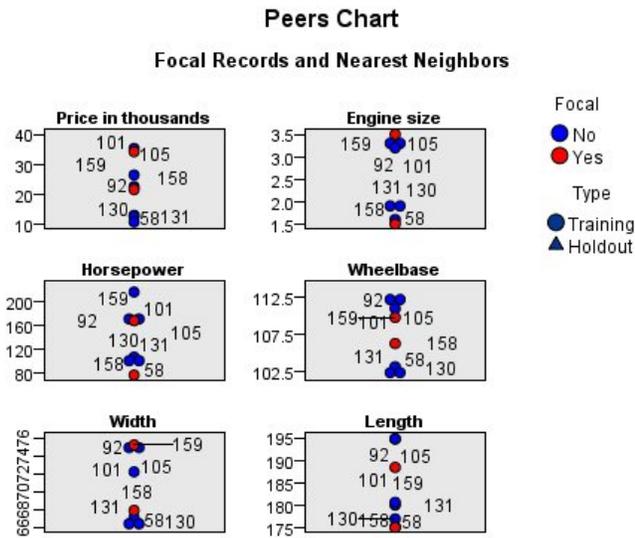


Abbildung 397. Peerdiagramm

Die Standardhilfsansicht ist das Peerdiagramm, 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 Identifizierung die Ausgabe aus dem Tabellenknoten.

Wenn die Ausgabe des Tabellenknotens noch verfügbar ist:

1. Klicken Sie auf die Registerkarte **Ausgabe** des Managerbereichs oben rechts im IBM SPSS Modeler-Hauptfenster.
2. Doppelklicken Sie auf den Eintrag **Tabelle (16 Felder, 159 Datensätze)**.

Wenn die Ausgabe des Tabellenknotens nicht mehr verfügbar ist:

3. Öffnen Sie im IBM SPSS Modeler-Hauptfenster den Tabellenknoten.
4. Klicken Sie auf **Ausführen**.

	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...

Abbildung 398. Identifizieren von Datensätzen nach Datensatznummer

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.

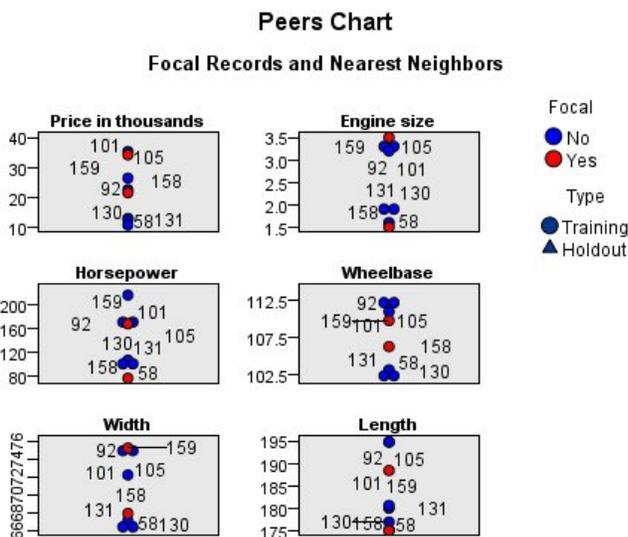


Abbildung 399. Vergleich der Funktionen auf dem Peerdigramm

Dadurch können wir auf dem Peerdigramm 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 Peerdiagramm ist ein wenig unübersichtlich, deshalb wollen wir zu einer einfacheren Ansicht wechseln.

5. Klicken Sie auf die Dropdown-Liste **Ansicht** am unteren Rand des Peerdiagramms (auf den Eintrag, der gerade **Peers** lautet).
6. Wählen Sie **Nachbar und Abstandstabelle** aus.

Nachbar und Abstandstabelle

k Nearest Neighbors and Distances
Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distances	
	1	2	3	1	2
158	131	130	58	0.979	0.990
159	105	92	101	0.580	0.634

Abbildung 400. Nachbar und Abstandstabelle

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 Dataset vergleichen können. Zudem wurden für zwei sehr unterschiedliche Holdout-Datensätze die Fälle berechnet, die diesen Holdouts am ähnlichsten sind.

Bemerkungen

Diese Informationen wurden für weltweit angebotene Produkte und Dienstleistungen erarbeitet.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite „Copyright and trademark information“ unter www.ibm.com/legal/copytrade.shtml.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Marken oder eingetragene Marken der Intel Corporation oder der Tochtergesellschaften des Unternehmens in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA, anderen Ländern oder beidem.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein.

Index

A

- Ableitungsknoten 83
- Abwärtssuche
 - Entscheidungslistenmodelle 108
- Analyseknoten 89
- Anmeldung bei IBM SPSS Modeler Server 8
- Anpassungsgüte
 - in verallgemeinerten linearen Modellen 275, 279
- Anwendungsbeispiele 3
- Ausführung anhalten 15
- Ausgabe 13
- Ausschneiden 15

B

- Bedingungsüberwachung 229
- Befehlszeile
 - IBM SPSS Modeler starten 7
- Beispiele
 - Anwendungshandbuch 3
 - Bayes-Netz 205, 213
 - Bedingungsüberwachung 229
 - Diskriminanzanalyse 235
 - Eingabezeichenfolgenlänge reduzieren 99
 - Einzelhandelsanalyse 225
 - Katalogverkäufe 179
 - KNN 337
 - multinomiale logistische Regression 131, 139
 - neue Fahrzeugangebote bewerten 337
 - SVM 287
 - Telekommunikation 131, 139, 151, 171, 235
 - Übersicht 5
 - Umcodierungsknoten 99
 - Warenkorbanalyse 329
 - Zeichenfolgenlänge reduzieren 99
 - Zellprobenklassifikation 287
- Benutzer-ID
 - IBM SPSS Modeler Server 8

C

- CLEM
 - Einführung 20
- Codierungen für kategoriale Variablen
 - in der Cox-Regression 301
- Coordinator of Processes 9
- COP 9
- Cox-Regression
 - Codierungen für kategoriale Variablen 301
 - Hazard-Kurve 306
 - Überlebenskurve 305
 - Variablenauswahl 302
 - zensierte Fälle 300

CRISP-DM 15

D

- Daten
 - anzeigen 78
 - einlesen 75
 - Manipulation 83
 - Modellierung 86, 88, 89
- Diagrammknoten 81
- Diskriminanzanalyse
 - Eigenwerte 242
 - Klassifikationstabelle 245
 - schrittweise Methoden 241
 - Strukturmatrix 243
 - Territorien 244
 - Wilks-Lambda 242
- Dokumentation 3
- Domänenname (Windows)
 - IBM SPSS Modeler Server 8
- Drucken 20
 - Streams 18

E

- Eigenwerte
 - in der Diskriminanzanalyse 242
- Einfügen 15
- Einführung
 - IBM SPSS Modeler 7
- Einzelhandelsanalyse 225
- Entscheidungslistenknoten
 - Beispielanwendung 105
- Entscheidungslistenmodelle
 - Beispielanwendung 105
 - benutzerdefinierte Maße mithilfe von Excel 121
 - erzeugen 129
 - Excel-Vorlage ändern 127
 - mit Excel verbinden 121
 - Sitzungsinformationen speichern 129
- Entscheidungslistenviewer 108
- Erstellungsbereich 11
- Excel
 - Entscheidungslistenvorlage ändern 127
 - mit Entscheidungslistenmodellen verbinden 121
- Expression Builder 83

F

- Felder
 - Auswahl für die Analyse 91
 - Rangordnung der Wichtigkeit 91
 - Screening 91
- Filtern 86

G

- Gammaregression
 - in verallgemeinerten linearen Modellen 281
- Generierte Modelle (Palette) 13
- Größe ändern 17
- Gruppierte Überlebensdaten
 - in verallgemeinerten linearen Modellen 247

H

- Hauptfenster 11
- Hazard-Kurve
 - in der Cox-Regression 306
- Hostname
 - IBM SPSS Modeler Server 8, 9
- Hotkeys 19

I

- IBM SPSS Modeler 1, 11
 - Dokumentation 3
 - erste Schritte 7
 - über Befehlszeile ausführen 7
 - Übersicht 7
- IBM SPSS Modeler Server 1
 - Benutzer-ID 8
 - Domänenname (Windows) 8
 - Hostname 8, 9
 - Kennwort 8
 - Portnummer 8, 9
- IBM SPSS Modeler Server-Verbindungen hinzufügen 9
- Intervallzensierte Überlebensdaten
 - in verallgemeinerten linearen Modellen 247

K

- Kennwort
 - IBM SPSS Modeler Server 8
- Klassen 15
- Klassifikationstabelle
 - in der Diskriminanzanalyse 245
- Knoten 7
- Knoten für lernfähiges Antwortmodell
 - Beispielanwendung 193
 - Beispielstream zum Erstellen 194
 - Modell durchsuchen 198
 - Stream erstellen 194
- Kopieren 15

M

- Manager 13
- Maus
 - in IBM SPSS Modeler verwenden 18

- Mehrere IBM SPSS Modeler-Sitzungen 10
- Merkmalauswahlknoten
 - Rangordnung von Prädiktoren 91
 - Screening von Prädiktoren 91
 - Wichtigkeit 91
- Merkmalauswahlmodelle 91
- Microsoft Excel
 - Entscheidungslistenvorlagen ändern 127
 - mit Entscheidungslistenmodellen verbinden 121
- Minimierung 17
- Mining-Aufgaben
 - Entscheidungslistenmodelle 108
- Mittelwerte von Kovariaten
 - in der Cox-Regression 304
- Mittlere Maustaste
 - Simulierung 18
- Modellierung 86, 88, 89

N

- Negative binomiale Regression
 - in verallgemeinerten linearen Modellen 277
- Netzdiagrammknoten 81
- Nuggets
 - definiert 13

O

- Omnibus-Test
 - in verallgemeinerten linearen Modellen 275
- Omnibus-Tests
 - in der Cox-Regression 302

P

- Paletten 11
- Parameterschätzungen
 - in verallgemeinerten linearen Modellen 253, 264, 276, 285
- Poisson-Regression
 - in verallgemeinerten linearen Modellen 271
- Portnummer
 - IBM SPSS Modeler Server 8, 9
- Prädiktoren
 - Auswahl für die Analyse 91
 - Rangordnung der Wichtigkeit 91
 - Screening 91
- Projekte 15

Q

- Quellenknoten 75

R

- Rangordnung von Prädiktoren 91
- Rest
 - Entscheidungslistenmodelle 108
- Rückgängig 15

S

- Schrittweise Methoden
 - in der Cox-Regression 302
 - in der Diskriminanzanalyse 241
- Screening von Prädiktoren 91
- Scripts 20
- Segmente
 - aus Scoring ausschließen 108
 - Entscheidungslistenmodelle 108
- Server
 - Anmeldung 8
 - nach COP für Server suchen 9
 - Verbindungen hinzufügen 9
- Single Sign-on 8
- Skalierung der Streams für Ansicht 18
- SLRM-Knoten
 - Beispielanwendung 193
 - Beispielstream zum Erstellen 194
 - Modell durchsuchen 198
 - Stream erstellen 194
- Stream 11
- Streams 7
 - erstellen 75
 - Skalierung für Ansicht 18
- Strukturmatrix
 - in der Diskriminanzanalyse 243
- Suche mit geringer Wahrscheinlichkeit
 - Entscheidungslistenmodelle 108
- Suche nach COP für Verbindungen 9
- Symbole
 - Optionen festlegen 18
- Symboleiste 15

T

- Tabellenknoten 78
- Tastenkombinationen
 - Tastatur 19
- Temp-Verzeichnis 10
- Territorien
 - in der Diskriminanzanalyse 244
- Tests der Modelleffekte
 - in verallgemeinerten linearen Modellen 251, 263, 276

U

- Überlebenskurven
 - in der Cox-Regression 305

V

- var.-Dateiknoten 75
- Verallgemeinerte lineare Modelle
 - Anpassungsgüte 275, 279
 - Omnibus-Test 275
 - Parameterschätzungen 253, 264, 276, 285
 - Poisson-Regression 271
 - Tests der Modelleffekte 251, 263, 276
 - verwandte Prozeduren 270, 280, 285
- Verbindungen
 - Server-Cluster 9
 - zu IBM SPSS Modeler Server 8, 9

- Viewer "Interaktive Liste"
 - arbeiten mit 108
 - Beispielanwendung 108
 - Vorschaufenster 108
- Visuelle Programmierung 11
- Vorbereiten 83

W

- Warenkorbanalyse 329
- Wichtigkeit
 - Rangordnung von Prädiktoren 91
- Wilks-Lambda
 - in der Diskriminanzanalyse 242

Z

- Zensierte Fälle
 - in der Cox-Regression 300
- Zoomen 15

