IBM SPSS Modeler 18.3 - Handbuch zu datenbankinternem Mining

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen” auf Seite 107 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 18, Release 3, Modifikation 0 von IBM® SPSS Modeler und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Inhaltsverzeichnis

Vorwort

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

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

Informationen zu IBM Business Analytics

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

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

Technical Support

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

Kapitel 1. Informationen zu IBM SPSS Modeler

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

IBM SPSS Modeler bietet eine Vielzahl von Modellierungsmethoden, 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 in https://www.ibm.com/analytics/us/en/technology/spss/.

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 (im Lieferumfang von IBM SPSS Deployment Manager enthalten)
- 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 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 Modus für verteilte Analysen 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 Data-

sets, 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 Benutzerschnittstelle zur Verwaltung einer Vielzahl der SPSS Modeler Server-Konfigurationsoptionen, die auch mithilfe einer Optionsdatei konfiguriert werden können. Die Konsole gehört zum Lieferumfang von IBM SPSS Deployment Manager, kann zum Überwachen und Konfigurieren Ihrer SPSS Modeler Server-Installationen verwendet werden und stehen 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. SPSS Modeler Server ist für die Verwendung von SPSS Modeler Batch 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.

Weitere Informationen zu SPSS Modeler Solution Publisher finden Sie in der Dokumentation zu IBM SPSS Collaboration and Deployment Services. Die IBM Dokumentation zu IBM SPSS Collaboration and Deployment Services enthält die Abschnitte "IBM SPSS-Modeler-Lösungs-Verlag" und "IBM SPSS Analytics Toolkit."

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 Repository von IBM SPSS Collaboration and Deployment Services 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 sowie für die Arbeit mit unstrukturierten Textdaten erweitert. SPSS Modeler Premium schließt IBM SPSS Modeler Text Analytics ein:

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 vorhandenen 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 Subscription

IBM SPSS Modeler Subscription stellt dieselbe Vorhersageanalysefunktionalität bereit wie der konventionelle IBM SPSS Modeler-Client. Mit der Subscription-Edition können Sie regelmäßig Produktaktualisierungen herunterladen.

Dokumentation

Dokumentation ist über das Hilfemenü in SPSS Modeler verfügbar. Dadurch wird die IBM Online-Dokumentation geöffnet, die außerhalb des Produkts stets verfügbar ist.

Die vollständige Dokumentation für die einzelnen Produkte (einschließlich Installationsanweisungen) ist über den Produktdownload in einem separaten komprimierten Ordner auch im PDF-Format verfügbar. Die aktuellen PDF-Dokumente können auch über das Web unter https://www.ibm.com/support/pages/spss-modeler-183-documentation heruntergeladen werden.

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, das Arbeiten mit Projekten und Berichten sowie das Packen von Streams für die Bereitstellung in IBM SPSS Collaboration and Deployment Services 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 Anwendungshandbuch.** Die Beispiele in diesem Handbuch bieten eine kurze, gezielte Einführung in bestimmte Modellierungsmethoden und -verfahren. Eine Online-Version dieses Handbuchs kann auch über das Hilfemenü aufgerufen werden. Weitere Informationen finden Sie im Abschnitt „Anwendungsbeispiele” auf Seite 4.
- **IBM SPSS Modeler Python Handbuch für Scripterstellung und Automatisierung.** Informationen zur Automatisierung des Systems über Python-Scripting, 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 als Schritte bei der Verarbeitung von Jobs im IBM SPSS 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 Deployment Manager Benutzerhandbuch.** Informationen zur Verwendung der zum Lieferumfang von Deployment Manager gehörenden Benutzerschnittstelle der Administrationskonsole zum Überwachen und Konfigurieren von IBM SPSS Modeler Server.
- **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:

- **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.

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 können jedoch auch auf reale Anwendungen übertragen werden.

Klicken Sie im Menü "Hilfe" in SPSS Modeler auf die Option **Anwendungsbeispiele**, um auf die Beispiele zuzugreifen.

Die Datendateien und Beispielstreams wurden im Ordner `Demos`, einem Unterordner des Produktinstallationsverzeichnisses, installiert. Weitere Informationen finden Sie in „Ordner "Demos"” auf Seite 4.

Beispiele für die Datenbankmodellierung. Die Beispiele finden Sie im Handbuch *IBM SPSS Modeler Datenbankinternes Mining*.

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

Ordner "Demos"

Die in den Anwendungsbeispielen verwendeten Datendateien und Beispielstreams werden im Ordner `Demos`, einem Unterordner des Produktinstallationsverzeichnisses (z. B. `C:\Programme\IBM\SPSS\Modeler\<version>\Demos`) installiert. Auf diesen Ordner können Sie auch über die Programmgruppe SPSS Modeler im Windows-Startmenü oder durch Klicken auf `Demos` in der Liste der zuletzt angezeigten Verzeichnisse im Dialogfeld **Datei** > **Stream öffnen** zugreifen.

Lizenzüberwachung

Bei der Verwendung von SPSS Modeler wird die Lizenznutzung überwacht und in regelmäßigen Intervallen protokolliert. Es werden die Lizenzmetriken *AUTHORIZED_USER* und *CONCURRENT_USER* protokolliert und der Typ der protokollierten Metrik ist von Ihrem Lizenztyp für SPSS Modeler abhängig.

Die erstellten Protokolldateien können vom Produkt IBM License Metric Tool verarbeitet werden, über das Sie Lizenznutzungsberichte generieren können.

Die Lizenzprotokolldateien werden im selben Verzeichnis erstellt, in dem SPSS Modeler Client-Protokolldateien aufgezeichnet werden (standardmäßig in `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`).

Kapitel 2. Datenbankinternes Mining

Übersicht über die Datenbankmodellierung

IBM SPSS Modeler Server unterstützt die Integration in Data-Mining-Tools und Datenmodellierungstools von Datenbankanbietern wie IBM Netezza, Oracle Data Miner und Microsoft Analysis Services. Sie können Modelle erstellen, scoren und in der Datenbank speichern, ohne dazu die IBM SPSS Modeler-Anwendung verlassen zu müssen. Damit können Sie die analytischen Funktionen und die Benutzerfreundlichkeit des IBM SPSS Modeler-Desktops mit der Leistungsstärke einer Datenbank kombinieren und gleichzeitig die datenbankeigenen Algorithmen nutzen, die von diesen Herstellern angeboten werden. Die Modelle werden innerhalb der Datenbank erstellt. Anschließend können Sie sie auf normale Weise über die IBM SPSS Modeler-Benutzerschnittstelle durchsuchen und scoren und bei Bedarf ihre Bereitstellung mithilfe von IBM SPSS Modeler Solution Publisher durchführen. Die unterstützten Algorithmen befinden sich in der Datenbankmodellierungspalette von IBM SPSS Modeler.

Der Zugriff auf datenbankeigene Algorithmen mithilfe von IBM SPSS Modeler bietet mehrere Vorteile:

- Datenbankeigene Algorithmen sind häufig eng mit dem Datenbankserver integriert und bieten u. U. eine verbesserte Leistung.
- Modelle, die innerhalb der Datenbank erstellt und gespeichert werden, können einfacher bereitgestellt und mit allen Anwendungen, die Zugriff auf die Datenbank haben, gemeinsam genutzt werden.

SQL-Generierung. Die Modellierung innerhalb der Datenbank ist nicht dasselbe wie die SQL-Generierung, die auch als "SQL-Pushback" bekannt ist. Mit dieser Funktion können Sie SQL-Anweisungen für native IBM SPSS Modeler-Operationen generieren, die dann zur Leistungsverbesserung per Pushback in die Datenbank zurückübertragen (und somit dort ausgeführt) werden können. Die Zusammenführungs-, Aggregat- und Auswahlknoten generieren beispielsweise jeweils SQL-Code, der auf diese Weise per Pushback an die Datenbank zurückübertragen werden kann. Die Verwendung von SQL-Generierung in Verbindung mit Datenbankmodellierung kann zu Streams führen, die von Anfang bis Ende in der Datenbank ausgeführt werden können, was erhebliche Leistungssteigerungen gegenüber in IBM SPSS Modeler ausgeführten Streams mit sich bringt.

Anmerkung: Datenbankmodellierung und SQL-Optimierung erfordern, dass auf dem IBM SPSS Modeler-Computer IBM SPSS Modeler Server-Konnektivität aktiviert ist. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus IBM SPSS Modeler per Pushback übertragen und auf IBM SPSS Modeler Server zugreifen. Wählen Sie zur Überprüfung des aktuellen Lizenzstatus Folgendes im Menü von IBM SPSS Modeler aus.

Hilfe > **Info...** > **Weitere Details**

Wenn Konnektivität aktiviert ist, wird auf der Registerkarte "Lizenzstatus" die Option **Serveraktivierung** angezeigt.

Informationen zu den unterstützten Algorithmen finden Sie in den nachfolgenden, herstellerspezifischen Abschnitten.

Anforderungen

Für die Datenbankmodellierung benötigen Sie das folgende Setup:

- Eine ODBC-Verbindung zu einer geeigneten Datenbank sowie die Installation der jeweils erforderlichen Analysekomponente (Microsoft Analysis Services oder Oracle Data Miner).).
- In IBM SPSS Modeler muss im Dialogfeld "Hilfsanwendungen" die Datenbankmodellierung aktiviert sein (**Extras > Hilfsanwendungen**).
- In IBM SPSS Modeler sowie in IBM SPSS Modeler Server (sofern verwendet) sollten im Dialogfeld "Benutzeroptionen" die Einstellungen **SQL generieren** und **SQL-Optimierung** aktiviert sein. Beachten Sie,

dass SQL-Optimierung für die Datenbankmodellierung nicht zwingend erforderlich ist, jedoch aus Leistungsgründen dringend empfohlen wird.

Anmerkung: Datenbankmodellierung und SQL-Optimierung erfordern, dass auf dem IBM SPSS Modeler-Computer IBM SPSS Modeler Server-Konnektivität aktiviert ist. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus IBM SPSS Modeler per Pushback übertragen und auf IBM SPSS Modeler Server zugreifen. Wählen Sie zur Überprüfung des aktuellen Lizenzstatus Folgendes im Menü von IBM SPSS Modeler aus.

Hilfe > **Info...** > **Weitere Details**

Wenn Konnektivität aktiviert ist, wird auf der Registerkarte "Lizenzstatus" die Option **Serveraktivierung** angezeigt.

Detaillierte Informationen finden Sie in den nachfolgenden, herstellerspezifischen Abschnitten.

Erstellen eines Modells

Der Vorgang des Erstellens und Scorens von Modellen mithilfe von Datenbankalgorithmen ähnelt anderen Data-Mining-Typen in IBM SPSS Modeler. Die allgemeinen Abläufe bei der Arbeit mit Knoten und Modellierungs-"Nuggets" ähneln der Arbeit mit anderen Streams in IBM SPSS Modeler. Der einzige Unterschied liegt darin, dass die eigentliche Verarbeitung und Modellerstellung an die Datenbank zurückgegeben werden.

Ein Datenbankmodellierungsstream ist konzeptionell mit anderen Datenstreams in IBM SPSS Modeler identisch. Dieser Stream führt jedoch alle Operationen in einer Datenbank aus, auch z. B. die Modellerstellung mithilfe des Knotens für Microsoft-Entscheidungsbäume. Wenn Sie den Stream ausführen, weist IBM SPSS Modeler die Datenbank an, das aus dem Stream resultierende Modell zu erstellen und zu speichern und die zugehörigen Details nach IBM SPSS Modeler herunterzuladen. Die im Stream violett eingezeichneten Knoten werden in der Datenbank ausgeführt.

Datenaufbereitung

Unabhängig davon, ob datenbankinterne Algorithmen verwendet werden oder nicht, sollten Sie die Datenaufbereitungsaufgaben nach Möglichkeit an die Datenbank zurückgeben, um so die Leistung zu steigern.

- Sind die Originaldaten in der Datenbank gespeichert, besteht das Ziel darin, die Daten dort zu behalten. Stellen Sie hierzu sicher, dass alle erforderlichen vorgeordneten Operationen in SQL umgewandelt werden können. Auf diese Weise vermeiden Sie, dass die Daten in IBM SPSS Modeler heruntergeladen werden (und somit ein Engpass entsteht, der den gesamten Leistungszuwachs zunichte machen würde), und Sie sorgen dafür, dass der gesamte Stream in der Datenbank ausgeführt wird.
- Sind die Originaldaten *nicht* in der Datenbank gespeichert, kann die Datenbankmodellierung dennoch verwendet werden. In diesem Fall wird die Datenaufbereitung in IBM SPSS Modeler ausgeführt und die vorbereiteten Daten werden automatisch an die Datenbank zum Erstellen des Modells hochgeladen.

Modellscoring

Modelle, die in IBM SPSS Modeler mit datenbankinternem Mining generiert werden, unterscheiden sich von regulären IBM SPSS Modeler-Modellen. Obwohl sie im Modellmanager als generierte "Modellnuggets" angezeigt werden, handelt es sich jedoch tatsächlich um ferne Modelle, die auf dem fernen Data-Mining-Server oder dem fernen Datenbankserver gespeichert werden. In IBM SPSS Modeler werden lediglich Verweise auf die fernen Modelle angezeigt. Mit anderen Worten: Das angezeigte IBM SPSS Modeler-Modell ist eine Modellhülle, die Informationen wie den Hostnamen des Datenbankservers, den Datenbanknamen sowie den Modellnamen enthält. Dies ist eine wichtige Unterscheidung, die beim Durchsuchen und Scoren von Modellen, die mit datenbankinternen Algorithmen erstellt wurden, berücksichtigt werden muss.

Sobald Sie ein Modell erstellt haben, können Sie es wie jedes andere in IBM SPSS Modeler generierte Modell dem Stream zum Zwecke des Scorens hinzufügen. Das gesamte Scoring erfolgt innerhalb der Datenbank, auch wenn dies bei vorgeordneten Operationen nicht der Fall ist. (Vorgeordnete Operationen können weiterhin nach Möglichkeit per Pushback an die Datenbank zurückübertragen werden, um die Leis-

tungsfähigkeit zu verbessern, dies ist jedoch keine Voraussetzung, damit das Scoring stattfinden kann.) Außerdem können Sie das generierte Modell in den meisten Fällen mithilfe des vom Datenbankanbieter bereitgestellten Standardbrowsers durchsuchen.

Für das Durchsuchen und das Scoring ist jeweils eine Live-Verbindung zu dem Server, auf dem Oracle Data Miner bzw. Microsoft Analysis Services ausgeführt wird, erforderlich.

Anzeigen von Ergebnissen und Festlegen von Einstellungen

Doppelklicken Sie zum Anzeigen von Ergebnissen und zum Festlegen von Einstellungen für das Scoring auf das Modell im Streamerstellungsbereich. Sie können auch mit der rechten Maustaste auf das Modell klicken und **Durchsuchen** oder **Bearbeiten** wählen. Bestimmte Einstellungen hängen vom Typ des Modells ab.

Exportieren und Speichern von Datenbankmodellen

Datenbankmodelle können wie andere in IBM SPSS Modeler erstellte Modelle und Übersichten mit den Optionen im Menü "Datei" aus dem Modellbrowser exportiert werden.

1. Wählen Sie im Menü "Datei" des Modellbrowsers eine der folgenden Optionen aus:

- **Text exportieren** exportiert die Modellübersicht in eine Textdatei.
- **HTML exportieren** exportiert die Modellübersicht in eine HTML-Datei.
- **PMML exportieren** (nur für IBM Db2 IM-Modelle unterstützt) exportiert das Modell als Predictive Model Markup Language (PMML), das mit anderer PMML-kompatibler Software verwendet werden kann.

Anmerkung: Sie können ein generiertes Modell auch mit **Knoten speichern** im Menü "Datei" speichern.

Modellkonsistenz

IBM SPSS Modeler speichert für jedes generierte Datenbankmodell unter demselben Namen, der in der Datenbank gespeichert ist, eine Beschreibung der Modellstruktur zusammen mit einem Verweis auf das Modell. Auf der Registerkarte "Server" eines generierten Modells wird ein eindeutiger für das Modell generierter Schlüssel angezeigt, der mit dem tatsächlichen Modell in der Datenbank übereinstimmt.

IBM SPSS Modeler überprüft anhand dieses zufällig erstellten Schlüssels, ob das Modell noch konsistent ist. Dieser Schlüssel wird bei der Modellerstellung in der Beschreibung des Modells gespeichert. Es empfiehlt sich, vor der Ausführung eines Bereitstellungsstreams zu überprüfen, ob die Schlüssel übereinstimmen.

1. Klicken Sie auf die Schaltfläche **Überprüfen**, um durch einen Vergleich der Modellbeschreibung des in der Datenbank gespeicherten Modells mit dem von IBM SPSS Modeler gespeicherten Zufallsschlüssel die Konsistenz des Modells zu überprüfen. Wird das Datenbankmodell nicht gefunden oder stimmen die beiden Schlüssel nicht überein, wird ein Fehler ausgegeben.

Anzeigen und Exportieren von generiertem SQL-Code

Der generierte SQL-Code kann vor der Ausführung angezeigt werden, was für die Fehlersuche hilfreich sein kann.

Kapitel 3. Datenbankmodellierung mit Microsoft Analysis Services

IBM SPSS Modeler und Microsoft Analysis Services

IBM SPSS Modeler unterstützt die Integration in Microsoft SQL Server Analysis Services. Diese Funktionalität ist in IBM SPSS Modeler als Modellierungsknoten implementiert und über die Datenbankmodellierungspalette verfügbar. Falls die Palette nicht sichtbar ist, aktivieren Sie im Dialogfeld "Hilfsanwendungen" auf der Registerkarte "Microsoft" die MS Analysis Services-Integration. Weitere Informationen finden Sie im Thema „Aktivieren der Integration in Analysis Services” auf Seite 13.

IBM SPSS Modeler unterstützt die Integration der folgenden Analysis Services-Algorithmen:

- Entscheidungsbäume
- Clustering
- Assoziationsregeln
- Naive Bayes
- Lineare Regression
- Neuronales Netz
- Logistische Regression
- Zeitreihen
- Sequenzclustering

Die folgende Abbildung veranschaulicht den Fluss der Daten vom Client zum Server, wobei das datenbankinterne Mining von IBM SPSS Modeler Server verwaltet wird. Die Modellerstellung erfolgt mit Analysis Services. Das resultierende Modell wird in Analysis Services gespeichert. Ein Verweis auf dieses Modell bleibt in den IBM SPSS Modeler-Streams erhalten. Das Modell wird dann aus Analysis Services entweder an Microsoft SQL Server oder an IBM SPSS Modeler zum Zwecke des Scoring heruntergeladen.

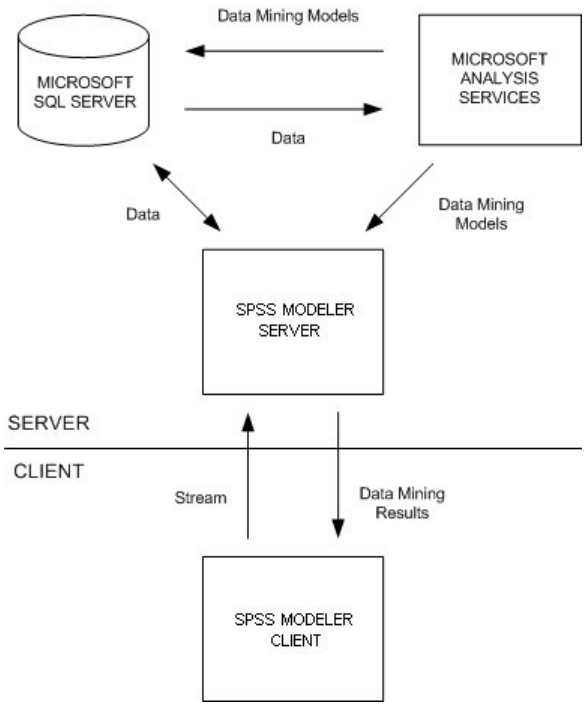

Abbildung 1. Datenfluss zwischen IBM SPSS Modeler, Microsoft SQL Server und Microsoft Analysis Services bei der Modellerstellung

Hinweis: IBM SPSS Modeler Server ist nicht erforderlich, kann jedoch verwendet werden. Der IBM SPSS Modeler-Client kann datenbankinterne Mining-Berechnungen verarbeiten.

Anforderungen für die Integration in Microsoft Analysis Services

Für die Modellerstellung innerhalb der Datenbank unter Verwendung von Analysis Services-Algorithmen mit IBM SPSS Modeler gelten die folgenden Voraussetzungen. Wenden Sie sich gegebenenfalls an Ihren Datenbankadministrator, um sicherzustellen, dass diese Bedingungen erfüllt sind.

- IBM SPSS Modeler wird im Rahmen einer IBM SPSS Modeler Server-Installation (verteilter Modus) unter Windows ausgeführt. UNIX-Plattformen werden bei dieser Integration in Analysis Services nicht unterstützt.

  Wichtig: Die IBM SPSS Modeler-Benutzer müssen eine ODBC-Verbindung mithilfe des SQL Native Client-Treibers konfigurieren, der unter der in *Additional IBM SPSS Modeler Server Requirements* (Zusätzliche Anforderungen) weiter unten angegebenen URL bei Microsoft erhältlich ist. *Der im Lieferumfang von IBM SPSS Data Access Pack enthaltene (und normalerweise für andere Verwendungszwecke von IBM SPSS Modeler empfohlene) Treiber wird hierfür nicht empfohlen.* Der Treiber sollte für die Verwendung von SQL Server mit aktivierter Funktion **Integrierte Windows-Authentifizierung** konfiguriert sein, da IBM SPSS Modeler keine SQL Server-Authentifizierung unterstützt. Wenn Sie Fragen zur Erstellung oder Einstellung von Berechtigungen für ODBC-Datenquellen haben, wenden Sie sich an Ihren Datenbankadministrator.
- SQL Server muss installiert sein, jedoch nicht unbedingt auf demselben Host wie IBM SPSS Modeler. IBM SPSS Modeler-Benutzer müssen über ausreichende Berechtigungen zum Lesen und Schreiben von Daten sowie zum Erstellen und Verwerfen von Tabellen und Ansichten verfügen.

  Anmerkung: SQL Server Enterprise Edition wird empfohlen. Die Enterprise Edition bietet zusätzliche Flexibilität durch Bereitstellung erweiterter Parameter zur Feinabstimmung der Algorithmusergebnisse. Die Version Standard Edition enthält dieselben Parameter, der Benutzer kann jedoch einige der erweiterten Parameter nicht bearbeiten.
- Microsoft SQL Server Analysis Services muss auf demselben Host wie SQL Server installiert sein.

Weitere IBM SPSS Modeler Server-Anforderungen

Um Analysis Services-Algorithmen mit IBM SPSS Modeler Server verwenden zu können, müssen folgende Komponenten auf dem IBM SPSS Modeler Server-Hostsystem installiert sein.

Anmerkung: Wenn SQL Server auf demselben Host wie IBM SPSS Modeler Server installiert ist, sind diese Komponenten bereits verfügbar.

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider. (Wählen Sie unbedingt die korrekte Version für Ihr Betriebssystem aus.)
- Microsoft SQL Server Native Client. (Wählen Sie unbedingt die korrekte Version für Ihr Betriebssystem aus.)
- Bei der Verwendung von Microsoft SQL Server 2008 oder 2012 ist möglicherweise auch Microsoft Core XML Services (MSXML) 6.0 erforderlich.

Um diese Komponenten herunterzuladen, navigieren Sie zu *www.microsoft.com/downloads*, suchen Sie **.NET Framework** bzw. (für alle anderen Komponenten) **SQL Server Feature Pack** und wählen Sie das neueste Paket für Ihre SQL Server-Version aus.

Möglicherweise müssen zunächst andere Pakete installiert werden. Diese sollten ebenfalls auf der Microsoft Downloads-Website erhältlich sein.

Weitere IBM SPSS Modeler-Anforderungen

Um Analysis Services-Algorithmen mit IBM SPSS Modeler verwenden zu können, müssen dieselben Komponenten installiert sein, wie oben angegeben. Darüber hinaus sind auf dem Client folgende Komponenten erforderlich:

- Microsoft SQL Server Datamining Viewer Controls. (Wählen Sie unbedingt die korrekte Version für Ihr Betriebssystem aus.) Hierfür ist außerdem folgende Komponente erforderlich:
- Microsoft ADOMD.NET

Um diese Komponenten herunterzuladen, navigieren Sie zu *www.microsoft.com/downloads*, suchen Sie **SQL Server Feature Pack** und wählen Sie das neueste Paket für Ihre SQL Server-Version aus.

Anmerkung: Datenbankmodellierung und SQL-Optimierung erfordern, dass auf dem IBM SPSS Modeler-Computer IBM SPSS Modeler Server-Konnektivität aktiviert ist. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus IBM SPSS Modeler per Pushback übertragen und auf IBM SPSS Modeler Server zugreifen. Wählen Sie zur Überprüfung des aktuellen Lizenzstatus Folgendes im Menü von IBM SPSS Modeler aus.

Hilfe > **Info...** > **Weitere Details**

Wenn Konnektivität aktiviert ist, wird auf der Registerkarte "Lizenzstatus" die Option **Serveraktivierung** angezeigt.

Aktivieren der Integration in Analysis Services

Um die Integration von IBM SPSS Modeler in Analysis Services zu ermöglichen, müssen Sie SQL Server und Analysis Services konfigurieren, eine ODBC-Datenquelle erstellen, im IBM SPSS Modeler-Dialogfeld "Hilfsanwendungen" die Integration aktivieren und SQL-Generierung und -Optimierung aktivieren.

Hinweis: Microsoft SQL Server und Microsoft Analysis Services müssen verfügbar sein. Weitere Informationen finden Sie im Thema „Anforderungen für die Integration in Microsoft Analysis Services” auf Seite 12.

Konfigurieren von SQL Server

Konfigurieren Sie SQL Server so, dass die Möglichkeit des Scoring innerhalb der Datenbank zugelassen wird.

1. Erstellen Sie auf dem SQL Server-Hostcomputer den folgenden Registrierungsschlüssel:

```
HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP
```

2. Fügen Sie diesem Schlüssel den folgenden DWORD-Wert hinzu:

```
AllowInProcess 1
```

3. Starten Sie SQL Server nach dieser Änderung neu.

Konfigurieren von Analysis Services

Bevor IBM SPSS Modeler mit Analysis Services kommunizieren kann, müssen zunächst zwei Einstellungen im Dialogfeld mit den Eigenschaften von Analysis Services manuell konfiguriert werden:

1. Melden Sie sich über MS SQL Server Management Studio beim Analysis Server an.
2. Öffnen Sie das Dialogfeld mit den Eigenschaften. Klicken Sie hierzu mit der rechten Maustaste auf den Servernamen und wählen Sie die Option **Eigenschaften** aus.
3. Aktivieren Sie das Kontrollkästchen **Erweiterte (Alle) Eigenschaften anzeigen**.
4. Ändern Sie die folgenden Eigenschaften:

- Ändern Sie den Wert für `DataMining\AllowAdHocOpenRowsetQueries` in `True` (der Standardwert ist `False`).
- Ändern Sie den Wert für `DataMining\AllowProvidersInOpenRowset` in `[all]` (hier gibt es keinen Standardwert).

Erstellen eines ODBC-DSN für SQL Server

Damit Sie in einer Datenbank lesen oder in sie schreiben können, muss eine ODBC-Datenquelle für die entsprechende Datenbank mit den erforderlichen Lese- und Schreibberechtigungen installiert und konfiguriert sein. Der Microsoft SQL Native Client ODBC-Treiber ist erforderlich und wird automatisch gemeinsam mit SQL Server installiert. *Der im Lieferumfang von IBM SPSS Data Access Pack enthaltene (und normalerweise für andere Verwendungszwecke von IBM SPSS Modeler empfohlene) Treiber wird hierfür nicht empfohlen.* Wenn sich IBM SPSS Modeler und SQL Server auf unterschiedlichen Hosts befinden, können Sie den Microsoft SQL Native Client ODBC-Treiber herunterladen. Weitere Informationen finden Sie im Thema „Anforderungen für die Integration in Microsoft Analysis Services” auf Seite 12.

Wenn Sie Fragen zur Erstellung oder Einstellung von Berechtigungen für ODBC-Datenquellen haben, wenden Sie sich an Ihren Datenbankadministrator.

1. Erstellen Sie mit dem Treiber für Microsoft SQL Native Client ODBC einen ODBC-DSN, der auf die im Data-Mining-Vorgang verwendete SQL Server-Datenbank verweist. Die restlichen Standardeinstellungen des Treibers sollten unverändert beibehalten werden.
2. Stellen Sie für diesen DSN sicher, dass die Option **Integrierte Windows-Authentifizierung** aktiviert ist.

- Wenn IBM SPSS Modeler und IBM SPSS Modeler Server auf unterschiedlichen Hosts ausgeführt werden, müssen Sie auf beiden Hosts denselben ODBC-DSN erstellen. Stellen Sie sicher, dass auf den Hosts jeweils derselbe DSN-Name verwendet wird.

Aktivieren der Analysis Services-Integration in IBM SPSS Modeler

Damit Analysis Services in IBM SPSS Modeler verwendet werden kann, müssen Sie zunächst im Dialogfeld "Hilfsanwendungen" einige Angaben zum Server machen.

1. Wählen Sie in den IBM SPSS Modeler-Menüs Folgendes aus:

   Tools > **Optionen** > **Hilfsanwendungen**

2. Klicken Sie auf die Registerkarte **Microsoft**.

- **Microsoft Analysis Services-Integration aktivieren.** Aktiviert die Datenbankmodellierungspalette (sofern nicht bereits angezeigt) am unteren Rand des IBM SPSS Modeler-Fensters und fügt die Knoten für die Analysis Services-Algorithmen hinzu.
- **Analyseserver-Host.** Geben Sie den Namen des Computers an, auf dem Analysis Services ausgeführt wird.
- **Analyseserverdatenbank.** Wählen Sie die gewünschte Datenbank aus, indem Sie auf die Schaltfläche mit Auslassungspunkten (...) klicken. Es wird ein weiteres Dialogfeld geöffnet, in dem Sie eine der ver-

fügbaren Datenbanken auswählen können. In der Liste werden die Datenbanken aufgeführt, die für den angegebenen Analyseserver verfügbar sind. Da Microsoft Analysis Services Data-Mining-Modelle in benannten Datenbanken speichert, sollten Sie die entsprechende Datenbank wählen, in der die mit IBM SPSS Modeler erstellten Microsoft-Modelle gespeichert sind.

- **SQL Server-Verbindung.** Geben Sie die DSN-Informationen an, die von der SQL Server-Datenbank zum Speichern der Daten verwendet werden, die an den Analyseserver weitergeleitet werden sollen. Wählen Sie die ODBC-Datenquelle aus, aus der die Daten für die Erstellung von Analysis Services Data-Mining-Modellen bereitgestellt werden. Wenn Sie Analysis Services-Modelle aus Daten von Flatfiles oder ODBC-Datenquellen erstellen, werden die Daten automatisch in eine temporäre Tabelle hochgeladen, die in der SQL Server-Datenbank erstellt wird, auf die diese ODBC-Datenquelle verweist.
- **Warnen, wenn ein Data-Mining-Modell überschrieben würde.** Wählen Sie diese Option aus, um sicherzustellen, dass in der Datenbank gespeicherte Modelle nicht von IBM SPSS Modeler überschrieben werden, ohne dass eine Warnung ausgegeben wird.

Hinweis: Im Dialogfeld "Hilfsanwendungen" vorgenommene Einstellungen können innerhalb der verschiedenen Analysis Services-Knoten überschrieben werden.

Aktivieren der SQL-Generierung und -Optimierung

1. Wählen Sie in den IBM SPSS Modeler-Menüs Folgendes aus:

   Tools > **Streameigenschaften** > **Optionen**
2. Klicken Sie im Navigationsbereich auf die Option **Optimierung**.
3. Überzeugen Sie sich, dass die Option **SQL generieren** aktiviert ist. Diese Einstellung ist für die Datenbankmodellierung erforderlich.
4. Wählen Sie **SQL-Generierung optimieren** und **Andere Ausführung optimieren** aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Erstellen von Modellen mit Analysis Services

Für die Modellierung von Analysis Services muss sich das Trainingsdataset in einer Tabelle oder Ansicht innerhalb der SQL Server-Datenbank befinden. Wenn sich die Daten nicht in SQL Server befinden oder wenn die Daten wegen einer Datenaufbereitung, die nicht in SQL Server erfolgen kann, in IBM SPSS Modeler verarbeitet werden müssen, werden diese vor der Modellierung automatisch in eine temporäre SQL Server-Tabelle geladen.

Verwalten von Analysis Services-Modellen

Bei der Bildung eines Analysis Services-Modells mit IBM SPSS Modeler wird in IBM SPSS Modeler ein Modell erstellt und außerdem in der SQL Server-Datenbank ein Modell erstellt oder ersetzt. Das IBM SPSS Modeler-Modell stellt einen Bezug zum Inhalt eines in einem Datenbankserver gespeicherten Datenbankmodells her.IBM SPSS Modeler kann eine Konsistenzprüfung durchführen, indem sowohl im IBM SPSS Modeler-Modell als auch im SQL Server-Modell eine identische, generierte Modellschlüsselzeichenfolge gespeichert wird.

Der MS-Modellierungsknoten **Entscheidungsbaum** wird für Vorhersagemodelle mit kategorialen und stetigen Attributen verwendet. Bei kategorialen Attributen erstellt der Knoten Vorhersagen auf der Grundlage der Beziehungen zwischen den Eingabespalten in einem Dataset. Beispiel: Wenn prognostiziert werden soll, welche Kunden mit hoher Wahrscheinlichkeit ein Fahrrad kaufen und neun von zehn jüngeren Kunden ein Fahrrad kaufen, jedoch nur zwei von zehn älteren Kunden, folgert der Knoten, dass das Alter ein guter Prädiktor für den Fahrradkauf ist. Der Entscheidungsbaum erstellt seine Vorhersagen dann auf der Grundlage dieser Tendenz hin zu einem bestimmten Ergebnis. Bei stetigen Attributen verwendet der Algorithmus lineare Regression, um zu ermitteln, an welcher Stelle sich ein Entscheidungsbaum aufspaltet. Wenn mehrere Spalten auf "vorhersagbar" gesetzt sind oder wenn die Eingangsdaten eine verschachtelte Tabelle enthalten, die auf "vorhersagbar" gesetzt ist, erstellt der Knoten einen gesonderten Entscheidungsbaum für jede vorhersagbare Spalte.

Der Modellierungsknoten **MS Clustering** verwendet iterative Verfahren zur Gruppierung von Fällen in einem Dataset in Clustern, die ähnliche Merkmale enthalten. Diese Gruppierungen sind sinnvoll für die Exploration von Daten, die Identifizierung von Anomalien in den Daten und die Erstellung von Vorhersagen. Clustering-Modelle identifizieren Beziehungen in einem Dataset, die sich möglicherweise nicht logisch durch Fallbeobachtungen ableiten lassen. Sie können beispielsweise durch Logik feststellen, dass Personen, die mit dem Fahrrad zum Arbeitsplatz pendeln, normalerweise nicht sonderlich weit von ihrem Arbeitsplatz entfernt wohnen. Der Algorithmus kann jedoch noch andere Merkmale von Fahrradpendlern ermitteln, die nicht so offensichtlich sind. Der Clusterknoten unterscheidet sich darin von anderen Data-Mining-Knoten, dass kein Zielfeld angegeben ist. Der Clusterknoten trainiert das Modell ausschließlich ausgehend von den Beziehungen, die in den Daten vorliegen, und von den Clustern, die der Knoten identifiziert.

Der Modellierungsknoten **MS-Assoziationsregeln** ist nützlich für Empfehlungsengines. Eine Empfehlungsengine empfiehlt Kunden Produkte auf der Grundlage der Artikel, die sie bereits erworben oder an denen sie Interesse bekundet haben. Assoziationsmodelle werden für Datasets erstellt, die IDs sowohl für die einzelnen Fälle aufweisen als auch für die Elemente, die diese Fälle enthalten. Eine Gruppe von Elementen in einem Fall wird als **Elementsatz** bezeichnet. Assoziationsmodelle bestehen aus einer Reihe von Elementsätzen und den Regeln, die beschreiben, wie diese Elemente innerhalb der Fälle in Gruppen zusammengefasst werden. Die Regeln, die der Algorithmus ermittelt, können verwendet werden, um anhand der Artikel, die sich bereits im Einkaufswagen des Kunden befinden, vorherzusagen, welche Artikel er voraussichtlich in der Zukunft erwerben wird.

Der MS-Modellierungsknoten **Naive Bayes** von Analysis Services berechnet die bedingte Wahrscheinlichkeit zwischen Ziel- und Prädiktorfeldern und geht dabei davon aus, dass die Spalten unabhängig sind. Das Modell wird als "naiv" bezeichnet, weil es alle vorgeschlagenen Vorhersagevariablen als voneinander unabhängig behandelt. Diese Methode erfordert weniger Berechnungsaufwand als die anderen Analysis Services-Algorithmen und ist daher nützlich für die schnelle Ermittlung von Beziehungen in den vorbereitenden Phasen der Modellierung. Mit diesem Knoten können Sie erste Explorationen der Daten vornehmen und anschließend die Ergebnisse anwenden, um zusätzliche Modelle mit anderen Knoten zu erstellen, deren Berechnung länger dauert, die jedoch zu genaueren Ergebnissen führen.

Der MS-Modellierungsknoten **Lineare Regression** ist eine Abwandlung des Knotens "Entscheidungsbäume", bei dem der Parameter `MINIMUM_LEAF_CASES` auf größer oder gleich der Gesamtzahl der Fälle im Datensatz gesetzt ist, die der Knoten für das Trainieren des Mining-Modells verwendet. Wenn der Parameter so gesetzt ist, erstellt der Knoten nie eine Aufteilung und führt also eine lineare Regression durch.

Der MS-Modellierungsknoten **Neuronales Netz** ähnelt dem MS-Knoten "Entscheidungsbäume" dahingehend, dass der MS-Knoten "Neuronales Netz" Wahrscheinlichkeiten für jeden möglichen Status des Eingabeattributs berechnet, wenn jeder Status des vorhersagbaren Attributs vorliegt. Später können Sie mithilfe dieser Wahrscheinlichkeiten auf der Grundlage der Eingabeattribute ein Ergebnis des vorhergesagten Attributs prognostizieren.

Der MS-Modellierungsknoten **Logistische Regression** ist eine Abwandlung des MS-Knotens "Neuronales Netz", bei dem der Parameter `HIDDEN_NODE_RATIO` auf 0 gesetzt ist. Diese Einstellung erstellt ein neuronales Netzmodell, das keine verdeckte Schicht enthält und daher der logistischen Regression entspricht.

Der Modellierungsknoten **MS Time Series** bietet Regressionsalgorithmen, die zur Vorhersage von stetigen Werten wie z. B. Produktverkäufen im Laufe der Zeit optimiert sind. Während andere Microsoft-Algorithmen, z. B. Entscheidungsbäume, zusätzliche Spalten mit neuen Informationen erfordern, um einen Trend vorherzusagen, verzichtet ein Zeitreihenmodell darauf. Ein Zeitreihenmodell kann Trends allein auf der Basis des ursprünglichen Datasets vorhersagen, mit dem das Modell erstellt wurde. Sie können dem Modell auch neue Daten hinzufügen, wenn Sie eine Vorhersage treffen, und automatisch die neuen Daten in der Trendanalyse berücksichtigen. Weitere Informationen finden Sie im Thema „Knoten "MS Time Series"” auf Seite 19.

Der MS-Modellierungsknoten **Sequenzclustering** identifiziert geordnete Sequenzen in Daten und kombiniert die Ergebnisse dieser Analyse mit Clustering-Techniken, um Cluster auf der Grundlage der Sequenzen und anderer Attribute zu generieren. Weitere Informationen finden Sie im Thema „MS-Sequenzclustering-Knoten” auf Seite 21.

Sie können von der Datenbankmodellierungspalette am unteren Rand des IBM SPSS Modeler-Fensters aus auf alle Knoten zugreifen.

Gemeinsame Einstellungen für alle Algorithmenknoten

Folgende Einstellungen haben alle Analysis Services-Algorithmen gemeinsam.

Serveroptionen

Auf der Registerkarte "Server" können Sie den Analyseserver-Host und die Analyseserver-Datenbank sowie die SQL Server-Datenquelle konfigurieren. Die hier festgelegten Optionen überschreiben die Optionen, die auf der Registerkarte "Microsoft" im Dialogfeld "Hilfsanwendungen" festgelegt wurden. Weitere Informationen finden Sie im Thema „Aktivieren der Integration in Analysis Services” auf Seite 13.

Hinweis: Beim Scoring von Analysis Services-Modellen steht eine Variante dieser Registerkarte zur Verfügung. Weitere Informationen finden Sie im Thema „Analysis Services-Modellnugget - Registerkarte "Server"” auf Seite 22.

Modelloptionen

Um das grundlegendste Modell erstellen zu können, müssen Sie auf der Registerkarte "Modell" Optionen festlegen, bevor Sie weitere Schritte durchführen. Die Scoring-Methode und andere erweiterte Optionen werden auf der Registerkarte "Experten" festgelegt.

Die folgenden grundlegenden Modellierungsoptionen sind verfügbar:

Modellname. Gibt den Namen des Modells an, das beim Ausführen des Knotens erstellt wird.

- **Auto.** Generiert den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen oder dem Namen des Modelltyps in Fällen, in denen kein Ziel angegeben ist (z. B. Clustering-Modelle).
- **Benutzerdefiniert.** Hier können Sie einen benutzerdefinierten Namen für das erstellte Modell angeben.

Partitionierte Daten verwenden. Teilt die Daten basierend auf dem aktuellen Partitionsfeld in separate Subsets oder Stichproben für das Training, das Testen und die Validierung auf. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer separaten Stichprobe testen, erhalten Sie einen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln. Wenn im Stream kein Partitionsfeld angegeben ist, wird diese Option ignoriert.

Mit Drillthrough. Wenn angezeigt, können Sie mithilfe dieser Option das Modell abfragen, um Einzelheiten über die darin enthaltenen Fälle zu erfahren.

Eindeutiges Feld. Wählen Sie aus der Dropdownliste ein Feld aus, das jeden Fall eindeutig identifiziert. Im Normalfall ist dies ein ID-Feld, wie z. B. **CustomerID**.

MS-Entscheidungsbäume - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

MS-Clustering - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

MS Naive Bayes - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

MS Lineare Regression - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

MS Neuronales Netz - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

MS Logistic Regression - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

Knoten "MS-Assoziationsregeln"

Der Modellierungsknoten "MS-Assoziationsregeln" ist nützlich für Empfehlungsengines. Eine Empfehlungsengine empfiehlt Kunden Produkte auf der Grundlage der Artikel, die sie bereits erworben oder an denen sie Interesse bekundet haben. Assoziationsmodelle werden für Datasets erstellt, die IDs sowohl für die einzelnen Fälle aufweisen als auch für die Elemente, die diese Fälle enthalten. Eine Gruppe von Elementen in einem Fall wird als **Elementsatz** bezeichnet.

Assoziationsmodelle bestehen aus einer Reihe von Elementsätzen und den Regeln, die beschreiben, wie diese Elemente innerhalb der Fälle in Gruppen zusammengefasst werden. Die Regeln, die der Algorithmus ermittelt, können verwendet werden, um anhand der Artikel, die sich bereits im Einkaufswagen des Kunden befinden, vorherzusagen, welche Artikel er voraussichtlich in der Zukunft erwerben wird.

Für Daten in Tabellenformat erstellt der Algorithmus Werte zur Darstellung der Wahrscheinlichkeit ($MP-*Feld*) für jede generierte Empfehlung ($M-*Feld*). Für Daten in Transaktionsformat werden Werte für Unterstützung ($MS-*Feld*), Wahrscheinlichkeit ($MP-*Feld*) und angepasste Wahrscheinlichkeit ($MAP-*Feld*) für jede generierte Empfehlung ($M-*Feld*) erstellt.

Anforderungen

Die Anforderungen für ein transaktionsorientiertes Assoziationsmodell sehen wie folgt aus:

- **Eindeutiges Feld.** Ein Assoziationsregelmodell erfordert einen Schlüssel, der Datensätze eindeutig identifiziert.
- **ID-Feld.** Beim Aufbau eines MS-Assoziationsregelmodells mit Daten im Transaktionsformat ist ein ID-Feld erforderlich, das jede Transaktion identifiziert. ID-Felder können auf dieselben Werte gesetzt werden wie das eindeutige Feld.
- **Mindestens ein Eingabefeld.** Der Assoziationsregelalgorithmus verlangt mindestens ein Eingabefeld.
- **Zielfeld.** Beim Erstellen eines MS-Assoziationsmodells mit Transaktionsdaten muss das Zielfeld das Transaktionsfeld sein, z. B. Produkte, die ein Benutzer gekauft hat.

MS-Assoziationsregeln - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

Knoten "MS Time Series"

Der Modellierungsknoten "MS Time Series" unterstützt zwei Arten der Vorhersage:

- Zukunftsvorhersagen
- Historische Vorhersagen

Zukunftsvorhersagen schätzen Zielfeldwerte für eine angegebene Anzahl an Zeitspannen über das Ende Ihrer historischen Daten hinaus und werden immer ausgeführt. **Historische Vorhersagen** sind geschätzte Zielfeldwerte für eine angegebene Anzahl an Zeitspannen, für die Ihre historischen Daten die tatsächlichen Werte enthalten. Sie können mithilfe historischer Vorhersagen die Qualität des Modells beurteilen, indem Sie die tatsächlichen historischen Werte mit den vorhergesagten Werten vergleichen. Der Wert des Anfangspunkts für die Vorhersagen bestimmt, ob historische Vorhersagen ausgeführt werden.

Im Unterschied zum IBM SPSS Modeler-Zeitreihenknoten benötigt der Knoten "MS Time Series" keinen vorangehenden Zeitintervallknoten. Ein weiterer Unterschied besteht darin, dass Werte standardmäßig

nur für die vorhergesagten Zeilen erzeugt werden, nicht für alle historischen Zeilen in den Zeitreihendaten.

Anforderungen

Die Anforderungen für ein MS Time Series-Modell sehen wie folgt aus:

- **Einzelnes Schlüsselzeitfeld.** Jedes Modell muss ein Zahlen- oder Datumsfeld enthalten, das als die Fallreihe verwendet wird und das Zeitintervall definiert, welches das Modell verwendet. Der Datentyp für das Schlüsselzeitfeld kann entweder Datum/Uhrzeit oder Zahl sein. Jedoch muss das Feld bestimmte stetige Werte enthalten, die für jede Reihe eindeutig sein müssen.
- **Einzelnes Zielfeld.** Sie können in jedem Modell nur ein Zielfeld angeben. Der Datentyp des Zielfelds muss stetige Werte haben. Sie können beispielsweise vorhersagen, wie numerische Attribute wie Einnahmen, Umsatz oder Temperatur sich im Laufe der Zeit ändern. Jedoch können Sie kein Feld verwenden, das kategoriale Werte als Zielfeld enthält, z. B. Kaufstatus oder Bildungsniveau.
- **Mindestens ein Eingabefeld.** Der MS Time Series-Algorithmus verlangt mindestens ein Eingabefeld. Der Datentyp des Eingabefelds muss stetige Werte haben. Nicht stetige Eingabefelder werden bei der Erstellung des Modells ignoriert.
- **Dataset muss sortiert sein.** Das Eingabedataset muss (nach dem Schlüsselzeitfeld) sortiert sein, ansonsten wird die Modellerstellung mit einem Fehler abgebrochen.

MS Time Series - Modelloptionen

Modellname. Gibt den Namen des Modells an, das beim Ausführen des Knotens erstellt wird.

- **Auto.** Generiert den Modellnamen automatisch basierend auf den Ziel- oder ID-Feldnamen oder dem Namen des Modelltyps in Fällen, in denen kein Ziel angegeben ist (z. B. Clustering-Modelle).
- **Benutzerdefiniert.** Hier können Sie einen benutzerdefinierten Namen für das erstellte Modell angeben.

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

Mit Drillthrough. Wenn angezeigt, können Sie mithilfe dieser Option das Modell abfragen, um Einzelheiten über die darin enthaltenen Fälle zu erfahren.

Eindeutiges Feld. Wählen Sie aus der Dropdown-Liste das Schlüsselzeitfeld aus, das zur Erstellung des Zeitreihenmodells verwendet wird.

MS Time Series - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

Für historische Vorhersagen entscheidet sich die Anzahl der historischen Schritte, die im Scoring-Ergebnis berücksichtigt werden können, durch den Wert von (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP). Standardmäßig beträgt die Begrenzung 10, d. h., nur zehn historische Vorhersagen werden getroffen. In diesem Fall tritt z. B. ein Fehler auf, wenn Sie einen Wert kleiner als -10 für **Historische Vorhersagen** auf der Registerkarte "Einstellungen" des Modellnuggets eingeben (siehe „Modellnugget "MS Time Series" - Registerkarte "Einstellungen"” auf Seite 24). Wenn Sie mehr historische Vorhersagen sehen möchten, können Sie den Wert von HISTORIC_MODEL_COUNT oder HISTORIC_MODEL_GAP erhöhen, wodurch sich allerdings auch die Erstellungsdauer für das Modell verlängert.

MS Time Series - Einstellungsoptionen

Schätzung beginnen. Geben Sie die Zeitperiode an, in der Vorhersagen beginnen sollen.

- **Beginn ab: Neue Vorhersage.** Die Zeitperiode, in der zukünftige Vorhersagen beginnen sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen

Daten bei 12.99 enden und Ihre Vorhersagen 01.00 beginnen sollen, verwenden Sie den Wert 1. Wenn die Vorhersagen jedoch 03.00 beginnen sollen, verwenden Sie den Wert 3.

- **Beginn ab: Historische Vorhersage.** Die Zeitperiode, in der historische Vorhersagen beginnen sollen, ausgedrückt als negativer Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Sie historische Vorhersagen für die letzten fünf Zeitperioden Ihrer Daten erstellen wollen, verwenden Sie den Wert -5.

Schätzung beenden. Geben Sie die Zeitperiode an, in der Vorhersagen enden sollen.

- **Vorhersageschritt beenden.** Die Zeitperiode, in der zukünftige Vorhersagen enden sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Ihre Vorhersagen 6.00 enden sollen, verwenden Sie hier den Wert 6. Für zukünftige Vorhersagen muss der Wert stets größer oder gleich dem **Start**-Wert sein.

MS-Sequenzclustering-Knoten

Der MS-Sequenzclustering-Knoten verwendet einen Sequenzanalyse-Algorithmus zur Exploration von Daten, die Ereignisse enthalten, die durch nachfolgende Pfade bzw. *Sequenzen* verknüpft werden können. Einige Beispiele dafür können die Klickpfade sein, die angelegt werden, wenn Benutzer in einer Website navigieren oder suchen, oder die Reihenfolge, in der ein Kunde Artikel in seinen Einkaufswagen bei einem Online-Händler legt. Der Algorithmus findet die häufigsten Sequenzen durch Gruppierung bzw. *Clustering* von Sequenzen, die identisch sind.

Anforderungen

Die Anforderungen für ein Microsoft Sequenzclustering-Modell sehen wie folgt aus:

- **ID-Feld.** Der Microsoft-Sequenzclustering-Algorithmus erfordert, dass die Sequenzinformationen in Transaktionsformat gespeichert sind. Dafür ist ein ID-Feld erforderlich, das jede Transaktion identifiziert.
- **Mindestens ein Eingabefeld.** Der Algorithmus verlangt mindestens ein Eingabefeld.
- **Sequenzfeld.** Der Algorithmus erfordert auch ein Sequenz-ID-Feld mit einem Messniveau des Typs "Stetig". Sie können beispielsweise eine Webseiten-ID, eine ganze Zahl oder eine Zeichenfolge verwenden, solange das Feld Ereignisse in einer Sequenz identifiziert. Nur eine Sequenz-ID ist pro Sequenz zulässig und nur ein Sequenztyp ist pro Modell erlaubt. Das Sequenzfeld muss sich von den Feldern "ID" und "Eindeutig" unterscheiden.
- **Zielfeld.** Ein Zielfeld ist beim Erstellen eines Sequenzclustering-Modells erforderlich.
- **Eindeutiges Feld.** Ein Sequenzclustering-Modell erfordert ein Schlüsselfeld, das Datensätze eindeutig identifiziert. Sie können das Feld "Eindeutig" auf denselben Wert setzen wie das Feld "ID".

MS-Sequenzclustering - Feldoptionen

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

Bevor Sie ein Sequenzclustering-Modell erstellen können, müssen Sie festlegen, welche Felder als Ziele und als Eingaben verwendet werden sollen. Beachten Sie, dass Sie für den MS-Sequenzclustering-Knoten keine Feldinformationen aus einem vorgeordneten Typenknoten können. Sie müssen die Feldeinstellungen hier angeben.

ID. Wählen Sie ein ID-Feld aus der Liste aus. Als ID-Feld können numerische oder symbolische Felder verwendet werden. Jeder eindeutige Wert in diesem Feld sollte eine bestimmte Analyseeinheit darstellen. Bei einer Warenkorbanwendung könnte z. B. jede ID einen einzelnen Kunden darstellen. Für eine Webprotokoll-Analyseanwendung könnte jede ID einen Computer (nach IP-Adresse) oder einen Benutzer (nach Anmeldedaten) darstellen.

Eingaben. Wählen Sie das Eingabefeld bzw. die Eingabefelder für das Modell aus. Diese Felder enthalten die in der Sequenzmodellierung interessanten Ereignisse.

Sequenz. Wählen Sie ein Feld aus der Liste aus, das als Sequenz-ID-Feld verwendet werden soll. Sie können beispielsweise eine Webseiten-ID, eine ganze Zahl oder eine Zeichenfolge verwenden, solange das

Feld Ereignisse in einer Sequenz identifiziert. Nur eine Sequenz-ID ist pro Sequenz zulässig und nur ein Sequenztyp ist pro Modell erlaubt. Das Feld "Sequenz" muss sich vom Feld "ID" (in dieser Registerkarte) und vom Feld "Eindeutig" (auf der Registerkarte "Modell") unterscheiden.

Ziel. Wählen Sie ein Feld aus, das als Zielfeld verwendet werden soll, d. h. das Feld, dessen Wert Sie auf der Grundlage der Sequenzdaten vorhersagen möchten.

MS-Sequenzclustering - Expertenoptionen

Die auf der Registerkarte "Experten" verfügbaren Optionen können je nach der Struktur des ausgewählten Streams variieren. Vollständige Details zu den für den ausgewählten Modellknoten von Analysis Server verfügbaren Expertenoptionen finden Sie in der Hilfe zu den einzelnen Feldern in der Benutzerschnittstelle.

Scoring von Analysis Services-Modellen

Das Modellscoring erfolgt in SQL Server und wird durch Analysis Services durchgeführt. Wenn die Daten aus IBM SPSS Modeler stammen oder wenn sie in IBM SPSS Modeler vorbereitet werden müssen, muss das Dataset gegebenenfalls in eine temporäre Tabelle geladen werden. Bei Modellen, die Sie mithilfe des datenbankinternen Minings von IBM SPSS Modeler aus erstellen, handelt es sich tatsächlich um ferne Modelle, die auf dem fernen Data-Mining- oder Datenbankserver gespeichert werden. Dies ist eine wichtige Unterscheidung, die beim Durchsuchen und Scoren von mit Microsoft Analysis Services-Algorithmen erstellten Modellen berücksichtigt werden muss.

In IBM SPSS Modeler wird in der Regel nur eine Vorhersage mit der zugehörigen Wahrscheinlichkeit oder Konfidenz erstellt.

Beispiele zum Modellscoring finden Sie in „Beispiele für das Mining mit Analysis Services” auf Seite 25.

Gemeinsame Einstellungen für alle Analysis Services-Modelle

Folgende Einstellungen haben alle Analysis Services-Modelle gemeinsam.

Analysis Services-Modellnugget - Registerkarte "Server"

Auf der Registerkarte "Server" können Verbindungen für das datenbankinterne Mining angegeben werden. Auf der Registerkarte wird auch der eindeutige Modellschlüssel angegeben. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip generiert und sowohl im Modell in IBM SPSS Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

Auf der Registerkarte "Server" können Sie den Analyseserver-Host und die Analyseserver-Datenbank sowie die SQL Server-Datenquelle für den Scoring-Vorgang konfigurieren. Die hier festgelegten Optionen überschreiben die Optionen, die in den Dialogfeldern "Hilfsanwendungen" oder "Modell erstellen" in IBM SPSS Modeler festgelegt wurden. Weitere Informationen finden Sie im Thema „Aktivieren der Integration in Analysis Services” auf Seite 13.

Modell-GUID. Hier wird der Modellschlüssel angezeigt. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip generiert und sowohl im Modell in IBM SPSS Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

Überprüfen. Klicken Sie auf diese Schaltfläche, um den Modellschlüssel mit dem Schlüssel des in der Analysis Services-Datenbank gespeicherten Modells zu vergleichen. Dadurch können Sie sicherstellen, dass das Modell noch im Analyseserver vorhanden ist und dass sich die Struktur des Modells nicht verändert hat.

Anmerkung: Die Schaltfläche "Überprüfen" ist nur für Modelle verfügbar, die dem Streamerstellungsbereich zur Vorbereitung auf das Scoring hinzugefügt werden. Schlägt die Überprüfung fehl, stellen Sie fest, ob das Modell gelöscht oder durch ein anderes Modell auf dem Server ersetzt wurde.

Ansicht. Klicken Sie hier, um eine grafische Darstellung des Entscheidungsbaummodells zu erhalten. Der Entscheidungsbaumviewer steht auch für andere Entscheidungsbaumalgorithmen in IBM SPSS Modeler zur Verfügung, wobei die Funktionalität immer gleich ist.

Analysis Services-Modellnugget - Registerkarte "Übersicht"

Auf der Registerkarte "Übersicht" eines Modellnuggets finden Sie Informationen zum Modell selbst (*Analyse*), den im Modell verwendeten Feldern (*Felder*), den beim Erstellen des Modells verwendeten Einstellungen (*Aufbaueinstellungen*) und zum Modelltraining (*Trainingsübersicht*).

Beim ersten Durchsuchen des Knotens sind die Ergebnisse der Registerkarte "Übersicht" reduziert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie sie mithilfe des Erweiterungssteuerelements links neben dem betreffenden Element oder klicken Sie auf die Schaltfläche **Alles anzeigen**, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement die gewünschten Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche **Alles ausblenden** alle Ergebnisse ausblenden.

Analyse. Zeigt Informationen zum jeweiligen Modell an. Wenn Sie einen Analyseknoten ausgeführt haben, der an dieses Modellnugget angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt.

Felder. Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden.

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

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

MS Time Series - Modellnugget

Das Modell "MS Time Series" erzeugt Werte nur für die vorhergesagten Zeitperioden, nicht für die historischen Daten.

In der folgenden Tabelle sind die Felder aufgeführt, die dem Modell hinzugefügt werden.

Tabelle 1. Dem Modell hinzugefügte Felder

Feldname	**Beschreibung**
$M-*feld*	Vorhergesagter Wert für *feld*
$Var-*feld*	Berechnete Varianz für *feld*
$Stdev-*feld*	Standardabweichung für *feld*

Modellnugget "MS Time Series" - Registerkarte "Server"

Auf der Registerkarte "Server" können Verbindungen für das datenbankinterne Mining angegeben werden. Auf der Registerkarte wird auch der eindeutige Modellschlüssel angegeben. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip generiert und sowohl im Modell in IBM SPSS Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

Auf der Registerkarte "Server" können Sie den Analyseserver-Host und die Analyseserver-Datenbank sowie die SQL Server-Datenquelle für den Scoring-Vorgang konfigurieren. Die hier festgelegten Optionen überschreiben die Optionen, die in den Dialogfeldern "Hilfsanwendungen" oder "Modell erstellen" in IBM SPSS Modeler festgelegt wurden. Weitere Informationen finden Sie im Thema „Aktivieren der Integration in Analysis Services” auf Seite 13.

Modell-GUID. Hier wird der Modellschlüssel angezeigt. Der Schlüssel wird bei der Modellerstellung nach dem Zufallsprinzip generiert und sowohl im Modell in IBM SPSS Modeler als auch in der Beschreibung des Modellobjekts gespeichert, das in der Analysis Services-Datenbank gespeichert ist.

Überprüfen. Klicken Sie auf diese Schaltfläche, um den Modellschlüssel mit dem Schlüssel des in der Analysis Services-Datenbank gespeicherten Modells zu vergleichen. Dadurch können Sie sicherstellen, dass das Modell noch im Analyseserver vorhanden ist und dass sich die Struktur des Modells nicht verändert hat.

Anmerkung: Die Schaltfläche "Überprüfen" ist nur für Modelle verfügbar, die dem Streamerstellungsbereich zur Vorbereitung auf das Scoring hinzugefügt werden. Schlägt die Überprüfung fehl, stellen Sie fest, ob das Modell gelöscht oder durch ein anderes Modell auf dem Server ersetzt wurde.

Ansicht. Klicken Sie hier, um eine grafische Darstellung des Zeitreihenmodells zu erhalten. Analysis Services zeigt das fertiggestellte Modell als Baumstruktur an. Sie können auch ein Diagramm mit dem historischen Wert des Zielfelds im Laufe der Zeit zusammen mit vorhergesagten zukünftigen Werten anzeigen.

Weitere Informationen finden Sie in der Beschreibung für den Zeitreihen-Viewer in der MSDN-Bibliothek unter http://msdn.microsoft.com/en-us/library/ms175331.aspx.

Modellnugget "MS Time Series" - Registerkarte "Einstellungen"

Schätzung beginnen. Geben Sie die Zeitperiode an, in der Vorhersagen beginnen sollen.

- **Beginn ab: Neue Vorhersage.** Die Zeitperiode, in der zukünftige Vorhersagen beginnen sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Ihre Vorhersagen 01.00 beginnen sollen, verwenden Sie den Wert 1. Wenn die Vorhersagen jedoch 03.00 beginnen sollen, verwenden Sie den Wert 3.
- **Beginn ab: Historische Vorhersage.** Die Zeitperiode, in der historische Vorhersagen beginnen sollen, ausgedrückt als negativer Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Sie historische Vorhersagen für die letzten fünf Zeitperioden Ihrer Daten erstellen wollen, verwenden Sie den Wert -5.

Schätzung beenden. Geben Sie die Zeitperiode an, in der Vorhersagen enden sollen.

- **Vorhersageschritt beenden.** Die Zeitperiode, in der zukünftige Vorhersagen enden sollen, ausgedrückt als Versatz von der letzten Zeitperiode Ihrer historischen Daten. Beispiel: Wenn Ihre historischen Daten bei 12.99 enden und Ihre Vorhersagen 6.00 enden sollen, verwenden Sie hier den Wert 6. Für zukünftige Vorhersagen muss der Wert stets größer oder gleich dem **Start**-Wert sein.

MS-Sequenzclustering-Modellnugget

In der folgenden Tabelle sind die Felder aufgeführt, die dem Microsoft-Sequenzclustering-Modell hinzugefügt werden (dabei steht *Feld* für den Namen des Zielfelds).

Tabelle 2. Dem Modell hinzugefügte Felder

Feldname	Beschreibung
$MC-*feld*	Vorhersage des Clusters, dem diese Sequenz angehört.
$MCP-*feld*	Wahrscheinlichkeit, dass diese Sequenz dem vorhergesagten Cluster angehört.
$MS-*feld*	Vorhergesagter Wert für *feld*
$MSP-*feld*	Wahrscheinlichkeit, dass der Wert von $MS-*feld* korrekt ist.

Exportieren von Modellen und Generieren von Knoten

Sie können eine Modellübersicht und -struktur als Text- oder HTML-Datei exportieren. Sie können die benötigten Auswahl- und Filterknoten generieren.

Ähnlich wie andere Modellnuggets in IBM SPSS Modeler unterstützen die Microsoft Analysis Services-Modellnuggets die direkte Generierung von Datensatz- und Feldoperationsknoten. Mit den Optionen im Menü "Generieren" des Modellnuggets können Sie die folgenden Knoten generieren:

- Auswahlknoten (nur wenn auf der Registerkarte "Modell" ein Element ausgewählt ist)
- Filterknoten

Beispiele für das Mining mit Analysis Services

Im Lieferumfang sind einige Beispielstreams enthalten, die die Verwendung von MS Analysis Services Data Mining mit IBM SPSS Modeler demonstrieren. Diese Streams befinden sich im IBM SPSS Modeler-Installationsordner unter:

\Demos\Database_Modelling\Microsoft

Hinweis: Der Ordner "Demos" kann über die Programmgruppe "IBM SPSS Modeler" im Windows-Startmenü aufgerufen werden.

Beispielstreams: Entscheidungsbäume

Die folgenden Streams können zusammen in Folge verwendet werden als Beispiel für einen Datenbankmining-Prozess, bei dem der Entscheidungsbaumalgorithmus von MS Analysis Services verwendet wird.

Tabelle 3. Entscheidungsbäume - Beispielstreams

Stream	**Beschreibung**
1_upload_data.str	Bereinigt Daten und lädt sie aus einer Flatfile in die Datenbank.
2_explore_data.str	Bietet ein Beispiel für die Datenuntersuchung mit IBM SPSS Modeler.
3_build_model.str	Erstellt das Modell unter Verwendung des datenbankeigenen Algorithmus.
4_evaluate_model.str	Wird als Beispiel für die Modellevaluierung mit IBM SPSS Modeler verwendet.
5_deploy_model.str	Verwendet das Modell für datenbankinternes Scoring.

Hinweis: Um das Beispiel auszuführen, müssen die Streams in der richtigen Reihenfolge ausgeführt werden. Außerdem müssen die Quellen- und Modellierungsknoten in den einzelnen Streams aktualisiert werden, um auf eine gültige Datenquelle für die zu verwendende Datenbank zu verweisen.

Das in den Beispielstreams verwendete Dataset bezieht sich auf Kreditkartenanwendungen und stellt ein Klassifizierungsproblem mit einer Mischung aus kategorialen und stetigen Prädiktoren dar. Weitere Informationen zu diesem Dataset finden Sie in der Datei *crx.names* in demselben Ordner wie die Beispielstreams.

Diese Daten stehen im UCI Machine Learning Repository unter *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* zur Verfügung.

Beispielstream: Hochladen von Daten

Der erste Beispielstream, *1_upload_data.str*, wird verwendet, um Daten aus einer Flatfile zu bereinigen und in SQL Server hochzuladen.

Da für das Data-Mining mit Analysis Services ein Schlüsselfeld erforderlich ist, fügt dieser erste Stream im Dataset mithilfe eines Ableitungsknotens ein neues Feld mit dem Namen *KEY* und den eindeutigen Werten *1,2,3* hinzu (unter Verwendung der IBM SPSS Modeler-Funktion `@INDEX`).

Der nachfolgende Füllerknoten ist für die Behandlung von fehlenden Werten zuständig und ersetzt leere, aus der Textdatei *crx.data* eingelesene Felder durch *NULL*-Werte.

Beispielstream: Datenexploration

Der zweite Beispielstream, *2_explore_data.str*, soll zeigen, wie mithilfe eines Data Audit-Knotens ein allgemeiner Überblick über die Daten (einschließlich statistischer Funktionen und Diagramme) gewonnen werden kann.

Wenn Sie im Data Audit-Bericht auf ein Diagramm doppelklicken, wird ein detaillierteres Diagramm angezeigt, in dem Sie einzelne Felder eingehender untersuchen können.

Beispielstream: Erstellen des Modells

Der dritte Beispielstream, *3_build_model.str*, veranschaulicht die Modellerstellung in IBM SPSS Modeler. Sie können das Datenbankmodell an den Stream anhängen und darauf doppelklicken, um Einstellungen für die Erstellung festzulegen.

Auf der Registerkarte "Modell" des Dialogfelds können Sie Folgendes festlegen:

1. Wählen Sie das Feld **Key** als eindeutiges ID-Feld aus.

Auf der Registerkarte "Experten" können Sie die Einstellungen für die Modellerstellung verfeinern.

Stellen Sie vor dem Ausführen des Streams sicher, das Sie die richtige Datenbank für die Modellerstellung angegeben haben. Verwenden Sie die Registerkarte "Server", um beliebige Einstellungen zu berichtigen.

Beispielstream: Auswerten des Modells

Der vierte Beispielstream, *4_evaluate_model.str*, veranschaulicht die Vorteile der Verwendung von IBM SPSS Modeler für die Modellierung innerhalb der Datenbank. Sobald Sie das Modell ausgeführt haben, können Sie es wieder zu Ihrem Datenstream hinzufügen und das Modell mit verschiedenen von IBM SPSS Modeler bereitgestellten Tools evaluieren.

Anzeigen der Modellierungsergebnisse

Sie können durch einen Doppelklick auf das Modellnugget Ihre Ergebnisse untersuchen. Auf der Registerkarte "Übersicht" werden die Ergebnisse in einer Baumansicht angezeigt. Mit der Schaltfläche **Ansicht** auf der Registerkarte "Server" öffnen Sie eine grafische Darstellung des Entscheidungsbaummodells.

Evaluieren der Modellergebnisse

Der Analyseknoten im Beispielstream erstellt eine Fehlklassifizierungstabelle, aus der das Muster der Übereinstimmungen zwischen jedem vorhergesagten Feld und dem zugehörigen Zielfeld ersichtlich wird. Führen Sie den Analyseknoten aus, um die Ergebnisse anzuzeigen.

Der Evaluierungsknoten im Beispielstream kann ein Gewinndiagramm erstellen, das die Verbesserungen der Vorhersagegenauigkeit durch das Modell aufzeigt. Führen Sie den Evaluierungsknoten aus, um die Ergebnisse anzuzeigen.

Beispielstream: Bereitstellen des Modells

Sobald Sie mit der Genauigkeit des Modells zufrieden sind, können Sie es für die Verwendung mit externen Anwendungen oder für eine erneute Veröffentlichung in der Datenbank bereitstellen. Im letzten Beispielstream, *5_deploy_model.str*, werden Daten aus der Tabelle `CREDIT` gelesen und dann mit dem Datenbankexportknoten gescort und in der Tabelle `CREDITSCORES` veröffentlicht.

Durch Ausführen des Streams wird folgender SQL-Code generiert:

```
DROP TABLE CREDITSCORES
CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" f
```

```
loat,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varcha
r(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"fiel
d12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varch
ar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )

INSERT INTO CREDITSCORES ("field1","field2","field3","field4","field5","field6","field7","field8",
"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
    T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
    T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
    T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
    SELECT CONVERT(NVARCHAR,[TA].[field1])  AS C0, CONVERT(NVARCHAR,[TA].[field2])  AS C1,
        [TA].[field3] AS C2, CONVERT(NVARCHAR,[TA].[field4])  AS C3,
        CONVERT(NVARCHAR,[TA].[field5])  AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
        CONVERT(NVARCHAR,[TA].[field7])  AS C6, [TA].[field8] AS C7,
        CONVERT(NVARCHAR,[TA].[field9])  AS C8, CONVERT(NVARCHAR,[TA].[field10])  AS C9,
        [TA].[field11] AS C10, CONVERT(NVARCHAR,[TA].[field12])  AS C11,
        CONVERT(NVARCHAR,[TA].[field13])  AS C12, [TA].[field14] AS C13,
        [TA].[field15] AS C14, CONVERT(NVARCHAR,[TA].[field16])  AS C15,
        [TA].[KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16])  AS C17,
        [TA].[$MC-field16] AS C18
    FROM openrowset('MSOLAP',
        'Datasource=localhost;Initial catalog=FoodMart 2000',
        'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
            [T].[C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
            [T].[C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
            [T].[C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
            [T].[C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
            [T].[C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
            PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
        FROM [CREDIT1] PREDICTION JOIN
            openrowset(''MSDASQL'',
            ''Dsn=LocalServer;Uid=;pwd='',''SELECT T0."field1" AS C0,T0."field2" AS C1,
            T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
            T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
            T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
            T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
            T0."KEY" AS C16 FROM "dbo".CREDITDATA T0'') AS [T]
        ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
            and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
            and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
            and [T].[C14] = [CREDIT1].[field15]') AS [TA]
)  T0
```

Kapitel 4. Datenbankmodellierung mit Oracle Data Mining

Informationen zu Oracle Data Mining

IBM SPSS Modeler unterstützt die Integration in Oracle Data Mining (ODM), das eine Serie von eng in Oracle RDBMS integrierten Data-Mining-Algorithmen bietet. Der Zugriff auf diese Funktionen erfolgt über die grafische Benutzerschnittstelle und die am Workflow orientierte Entwicklungsumgebung von IBM SPSS Modeler. So können Kunden die Data-Mining-Algorithmen von ODM verwenden.

IBM SPSS Modeler unterstützt die Integration folgender Oracle Data Mining-Algorithmen:

- Naive Bayes
- Adaptive Bayes
- Support Vector Machine (SVM)
- Verallgemeinerte lineare Modelle (GLM)*
- Entscheidungsbaum
- O-Cluster
- K-Means
- Nonnegative Matrix Factorization (NMF)
- Apriori
- Minimum Descriptor Length (MDL)
- Attribute Importance (AI)

* nur 11*g* R1

Voraussetzungen für die Integration in Oracle

Für die datenbankinterne Modellierung mit Oracle Data Mining gelten die folgenden Voraussetzungen. Wenden Sie sich gegebenenfalls an Ihren Datenbankadministrator, um sicherzustellen, dass diese Bedingungen erfüllt sind.

- Ausführung von IBM SPSS Modeler im lokalen Modus oder im Rahmen einer IBM SPSS Modeler Server-Installation unter Windows oder UNIX.
- Oracle 10*g* R2 oder 11*g* R1 (10.2 Database oder höher) mit der Option für Oracle Data Mining.

Anmerkung: 10*g* R2 stellt Unterstützung für alle Datenbankmodellierungsalgorithmen bereit, ausgenommen für "Verallgemeinerte lineare Modelle" (erfordern 11*g* R1).

- Eine ODBC-Datenquelle für die Verbindung mit Oracle wie unten beschrieben.

Anmerkung: Datenbankmodellierung und SQL-Optimierung erfordern, dass auf dem IBM SPSS Modeler-Computer IBM SPSS Modeler Server-Konnektivität aktiviert ist. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus IBM SPSS Modeler per Pushback übertragen und auf IBM SPSS Modeler Server zugreifen. Wählen Sie zur Überprüfung des aktuellen Lizenzstatus Folgendes im Menü von IBM SPSS Modeler aus.

Hilfe > **Info...** > **Weitere Details**

Wenn Konnektivität aktiviert ist, wird auf der Registerkarte "Lizenzstatus" die Option **Serveraktivierung** angezeigt.

Aktivieren der Integration in Oracle

Um die Integration von IBM SPSS Modeler in Oracle Data Mining zu ermöglichen, müssen Sie Oracle konfigurieren, eine ODBC-Datenquelle erstellen, im IBM SPSS Modeler-Dialogfeld "Hilfsanwendungen" die Integration aktivieren und die SQL-Generierung und -Optimierung aktivieren.

Konfigurieren von Oracle

Informationen zur Installation und Konfiguration von Oracle Data Mining finden Sie in der Oracle-Dokumentation. Weitere Details finden Sie insbesondere in *Oracle Administrator's Guide*.

Erstellen einer ODBC-Datenquelle für Oracle

Um die Verbindung zwischen Oracle und IBM SPSS Modeler zu aktivieren, müssen Sie einen ODBC-Datenquellennamen (DSN) erstellen.

Bevor Sie einen DSN erstellen, sollten Sie grundlegende Kenntnisse über ODBC-Datenquellen und -Treiber sowie über Datenbankunterstützung in IBM SPSS Modeler besitzen.

Wenn Sie mit IBM SPSS Modeler Server im verteilten Modus arbeiten, müssen Sie den DSN auf dem Server-Computer erstellen. Wenn Sie im lokalen (Client-)Modus arbeiten, müssen Sie auf dem Client-Computer einen DSN erstellen.

1. Installieren Sie die ODBC-Treiber. Diese Treiber finden Sie auf dem zu dieser Version gehörenden IBM SPSS Data Access Pack-Installationsmedium. Führen Sie die Datei *setup.exe* aus, um das Installationsprogramm zu starten, und wählen Sie alle relevanten Treiber aus. Befolgen Sie die Anweisungen auf dem Bildschirm, um die Treiber zu installieren.
   a. Erstellen Sie den DSN.

      Anmerkung: Die Menüfolge ist von der jeweils vorliegenden Windows-Version abhängig.

      - **Windows XP.** Wählen Sie im Menü "Start" die Option **Systemsteuerung** aus. Doppelklicken Sie auf **Verwaltung** und dann auf **Datenquellen (ODBC)**.
      - **Windows Vista.** Wählen Sie im Menü "Start" die Option **Systemsteuerung** und dann **Systemwartung** aus. Doppelklicken Sie auf **Verwaltung**, wählen Sie dann **Datenquellen (ODBC)** aus und klicken Sie auf **Öffnen**.
      - **Windows 7.** Wählen Sie im Menü "Start" die Option **Systemsteuerung**, dann **System & Sicherheit** und anschließend **Verwaltung** aus. Wählen Sie **Datenquellen (ODBC)** aus und klicken Sie dann auf **Öffnen**.

   b. Wechseln Sie zur Registerkarte **System-DSN** und klicken Sie auf **Hinzufügen**.
2. Wählen Sie den Treiber **SPSS OEM 6.0 Oracle Wire Protocol** aus.
3. Klicken Sie auf **Fertigstellen**.
4. Geben Sie in der Anzeige "ODBC Oracle Wire Protocol Driver Setup" einen Datenquellennamen Ihrer Wahl, den Hostnamen des Oracle-Servers, die Portnummer für die Verbindung und die SID für die verwendete Oracle-Instanz ein.

   Hostnamen, Port und SID finden Sie auf dem Serversystem in der Datei *tnsnames.ora*, sofern Sie TNS mit einer *tnsnames.ora*-Datei implementiert haben. Weitere Informationen erhalten Sie von Ihrem Oracle-Administrator.
5. Klicken Sie auf die Schaltfläche **Test**, um die Verbindung zu testen.

Aktivieren der Oracle Data Mining-Integration in IBM SPSS Modeler

1. Wählen Sie in den IBM SPSS Modeler-Menüs Folgendes aus:

   Tools > **Optionen** > **Hilfsanwendungen**
2. Klicken Sie auf die Registerkarte **Oracle**.

Oracle Data Mining-Integration aktivieren. Aktiviert die Datenbankmodellierungspalette (sofern nicht bereits angezeigt) am unteren Rand des IBM SPSS Modeler-Fensters und fügt die Knoten für die Oracle Data Mining-Algorithmen hinzu.

Oracle-Verbindung. Legen Sie die Oracle ODBC-Datenquelle fest, die bei der Bildung und dem Speichern der Modelle als Standard verwendet wird, und geben Sie einen gültigen Benutzernamen und ein Kennwort ein. Bei den einzelnen Modellierungsknoten und Modellnuggets kann diese Einstellung überschrieben werden.

Hinweis: Die für die Modellierung verwendete Datenbankverbindung kann, muss aber nicht mit der für den Datenzugriff verwendeten übereinstimmen. Sie können beispielsweise einen Stream einsetzen, der auf die Daten einer Oracle-Datenbank zugreift, die Daten für die Bereinigung und sonstige Bearbeitungen in IBM SPSS Modeler herunterlädt und dann zur Modellierung in eine andere Oracle-Datenbank lädt. Alternativ können sich die Originaldaten in einer Flatfile oder einer anderen Oracle-externen Quelle befinden. In diesem Fall müssen sie zur Modellierung in Oracle geladen werden. In allen Fällen werden die Daten automatisch in eine temporäre Tabelle geladen, die in der für die Modellierung verwendeten Datenbank angelegt wird.

Warnen, wenn ein Oracle Data Mining-Modell überschrieben würde. Wählen Sie diese Option aus, um sicherzustellen, dass in der Datenbank gespeicherte Modelle nicht von IBM SPSS Modeler überschrieben werden, ohne dass eine Warnung ausgegeben wird.

Oracle Data Mining-Modelle auflisten. Zeigt die verfügbaren Data-Mining-Modelle an.

Start von Oracle Data Miner aktivieren. (optional) Wenn diese Option aktiviert ist, kann IBM SPSS Modeler die Anwendung Oracle Data Miner starten. Weitere Informationen finden Sie in „Oracle Data Miner" auf Seite 48.

Pfad für ausführbare Datei von Oracle Data Miner. (optional) Gibt den physischen Speicherort der ausführbaren Oracle Data Miner-Datei für Windows an (zum Beispiel *C:\odm\bin\odminerw.exe*). Oracle Data Miner wird nicht zusammen mit IBM SPSS Modeler installiert. Die korrekte Version muss von der Oracle-Website (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) heruntergeladen und auf dem Client installiert werden.

Aktivieren der SQL-Generierung und -Optimierung

1. Wählen Sie in den IBM SPSS Modeler-Menüs Folgendes aus:

   Tools > **Streameigenschaften** > **Optionen**
2. Klicken Sie im Navigationsbereich auf die Option **Optimierung**.
3. Überzeugen Sie sich, dass die Option **SQL generieren** aktiviert ist. Diese Einstellung ist für die Datenbankmodellierung erforderlich.
4. Wählen Sie **SQL-Generierung optimieren** und **Andere Ausführung optimieren** aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Modellierung mit Oracle Data Mining

Die Oracle-Modellierungsknoten arbeiten bis auf einige wenige Ausnahmen in IBM SPSS Modeler genau wie andere Modellierungsknoten. Über die Datenbankmodellierungspalette am unteren Rand des IBM SPSS Modeler-Fensters können Sie auf diese Knoten zugreifen.

Erläuterung der Daten

Für Oracle müssen kategoriale Daten in einem Zeichenfolgenformat (entweder CHAR oder VARCHAR2) gespeichert sein. Demzufolge erlaubt IBM SPSS Modeler nicht, dass numerische Speicherfelder mit dem Messniveau *Flag* oder *Nominal* (kategorial) als Eingabe für ODM-Modelle verwendet werden. Gegebenenfalls können Nummern in IBM SPSS Modeler mit dem Umcodierungsknoten in Zeichenfolgen konvertiert werden.

Zielfeld. In ODM-Klassifizierungsmodellen kann nur ein Feld als Ausgabefeld (Ziel) ausgewählt werden.

Modellname. Ab Oracle 11*g* R1 ist `unique` ein Schlüsselwort und kann nicht als Name für benutzerdefinierte Modelle verwendet werden.

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Allgemeine Kommentare

- Für von Oracle Data Mining erstellte Modelle bietet IBM SPSS Modeler keinen PMML-Export/-Import.
- Das Modellscoring erfolgt immer innerhalb von ODM. Wenn die Daten aus IBM SPSS Modeler stammen oder dort vorbereitet werden müssen, muss das Dataset gegebenenfalls in eine temporäre Tabelle geladen werden.
- In IBM SPSS Modeler wird in der Regel nur eine Vorhersage mit der zugehörigen Wahrscheinlichkeit oder Konfidenz erstellt.
- IBM SPSS Modeler beschränkt die Anzahl der Felder, die beim Erstellen und Scoren von Modellen verwendet werden können, auf 1.000.
- IBM SPSS Modeler kann ODM-Modelle mit IBM SPSS Modeler Solution Publisher aus zur Ausführung veröffentlichten Streams heraus scoren.

Oracle-Modelle - Serveroptionen

Legen Sie die Oracle-Verbindung fest, die zum Hochladen der für die Modellierung verwendeten Daten verwendet wird. Gegebenenfalls können Sie auf der Registerkarte "Server" für jeden Modellierungsknoten eine Verbindung auswählen, mit der die im Dialogfeld "Hilfsanwendungen" angegebene Standard-Oracle-Verbindung überschrieben wird. Weitere Informationen finden Sie im Thema „Aktivieren der Integration in Oracle” auf Seite 30.

Kommentare

- Die für die Modellierung verwendete Verbindung kann mit der im Quellenknoten für einen Stream verwendeten Verbindung identisch sein. Sie können beispielsweise einen Stream einsetzen, der auf die Daten einer Oracle-Datenbank zugreift, die Daten für die Bereinigung und sonstige Bearbeitungen in IBM SPSS Modeler herunterlädt und dann zur Modellierung in eine andere Oracle-Datenbank lädt.
- Der Name der ODBC-Datenquelle wird in jeden IBM SPSS Modeler-Stream eingebettet. Wenn ein auf einem Host erstellter Stream auf einem anderen Host ausgeführt wird, muss der Name der Datenquelle auf beiden Hosts identisch sein. Alternativ kann für jeden Quellen- oder Modellierungsknoten auf der Registerkarte "Server" eine andere Datenquelle ausgewählt werden.

Fehlklassifizierungskosten

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

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

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

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

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

Hiinweis: Nur im Entscheidungsbaummodell können die Kosten zum Zeitpunkt der Erstellung angegeben werden.

Oracle Naive Bayes

Naive Bayes ist ein bekannter Algorithmus für Klassifizierungsprobleme. Das Modell wird als *naive* bezeichnet, weil es alle vorgeschlagenen Vorhersagevariablen als voneinander unabhängig behandelt. Naive Bayes ist ein schneller, skalierbarer Algorithmus, der für Kombinationen von Attributen und für das Zielattribut bedingte Wahrscheinlichkeiten berechnet. Aus den Trainingsdaten wird eine unabhängige Wahrscheinlichkeit ermittelt. Diese liefert die Wahrscheinlichkeit jeder Zielklasse anhand des Vorkommens der einzelnen Wertekategorien aus jeder einzelnen Eingabevariablen.

- Die Kreuzvalidierung wird eingesetzt, um die Modellgenauigkeit mit denselben Daten zu testen, die zur Modellierung verwendet wurden. Dies ist insbesondere dann nützlich, wenn die Anzahl der für die Modellierung verfügbaren Fälle gering ist.
- Die Modellausgabe kann in einem Matrixformat durchsucht werden. Bei den in der Matrix vorhandenen Zahlen handelt es sich um bedingte Wahrscheinlichkeiten, die sich auf die vorhergesagten Fälle (Spalten) und die Variablen-Wert-Kombinationen der Prädiktoren (Zeilen) beziehen.

Naive Bayes - Modelloptionen

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

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Naive Bayes - Expertenoptionen

Wenn das Modell erstellt wird, werden einzelne Prädiktorattributwerte oder -wertpaare ignoriert, wenn ein bestimmter Wert oder ein Wertpaar in den Trainingsdaten nicht häufig genug vorkommt. Die Schwellenwerte für ignorierte Werte werden als Anteilswerte der Anzahl der in den Trainingsdaten vorhandenen Datensätze angegeben. Die Anpassung dieser Schwellenwerte kann das Rauschen reduzieren und die Voraussetzungen des Modells verbessern, sodass es für andere Datasets verallgemeinert werden kann.

- **Singleton-Schwellenwert.** Legt den Schwellenwert für einen bestimmten Prädiktorattributwert fest. Die Häufigkeit des Vorkommens eines bestimmten Werts muss gleich oder höher sein als der angegebene Anteilswert. Ansonsten wird der Wert ignoriert.
- **Pairwise-Schwellenwert.** Legt den Schwellenwert für ein bestimmtes Attribut und ein Prädiktorwertpaar fest. Die Häufigkeit des Vorkommens eines bestimmten Wertpaars muss gleich oder höher sein als der angegebene Anteilswert. Ansonsten wird das Paar ignoriert.

Vorhersagewahrscheinlichkeit. Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Vorhersage für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie **Auswählen** wählen, auf die Schaltfläche **Angeben** klicken, eines der möglichen Ergebnisse wählen und dann auf **Einfügen** klicken.

Vorhersageset verwenden. Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

Oracle Adaptive Bayes

Adaptive Bayes Network (ABN) bildet anhand der Minimum Description Length (MDL) und der automatischen Merkmalauswahl Bayesian Network Classifier. ABN funktioniert in bestimmten Situationen gut, in denen Naive Bayes wenig erreicht. In den meisten anderen Fällen funktioniert ABN mindestens genauso gut, die Leistung kann allerdings etwas geringer sein. Mit dem ABN-Algorithmus können Baumtypen von erweiterten, auf Bayes basierenden Modellen gebildet werden, zu denen vereinfachte Entscheidungsbaummodelle (Einzelfunktion), reduzierte Naive Bayes-Modelle und verstärkte Multifunktionsmodelle gehören.

Anmerkung: Der Algorithmus Oracle Adaptive Bayes wurde in Oracle 12C entfernt und wird in IBM SPSS Modeler bei Verwendung von Oracle 12C nicht unterstützt. Siehe http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726.

Generierte Modelle

Im Einzelfunktionsmodus erstellt ABN einen vereinfachten Entscheidungsbaum, der auf einem Set lesbarer Regeln basiert, über die Fachanwender oder Analysten die Grundlage der Vorhersagen des Modells nachvollziehen und anderen entsprechend erläutern können. Dies kann sich im Vergleich zu Naive Bayes- oder Multifunktionsmodellen als signifikanter Vorteil erweisen. Diese Regeln können wie ein Standardregelset in IBM SPSS Modeler durchsucht werden. Ein einfaches Regelset könnte folgendermaßen aussehen:

```
IF MARITAL_STATUS = "Married"
AND EDUCATION_NUM = "13-16"
THEN CHURN= "TRUE"
Confidence = .78, Support = 570 cases
```

Reduzierte Naive Bayes- und Multifunktionsmodelle können nicht in IBM SPSS Modeler durchsucht werden.

Adaptive Bayes - Modelloptionen

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

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Modelltyp

Zum Erstellen des Modells stehen drei verschiedene Modi zur Auswahl.

- **Multimerkmal.** Erstellt und vergleicht mehrere Modelle, einschließlich eines NB-Modells sowie Einzel- und Multifunktionsmodellen für die Produktwahrscheinlichkeit. Hierbei handelt es sich um den umfassendsten Modus, dessen Berechnung demnach in der Regel auch am längsten dauert. Regeln werden nur dann erzeugt, wenn sich das Einzelfunktionsmodell am besten eignet. Wenn ein Multifunktions- oder ein NB-Modell ausgewählt wird, werden keine Regeln erzeugt.
- **Einzelmerkmal.** Erstellt einen vereinfachten Entscheidungsbaum, der auf einem Set von Regeln basiert. Jede Regel enthält eine Bedingung und Wahrscheinlichkeiten, die jedem Ergebnis zugeordnet sind. Die Regeln sind untereinander exklusiv und liegen in einer für Menschen lesbaren Form vor, was sich gegenüber Naive Bayes- und Multifunktionsmodellen als signifikanter Vorteil erweisen kann.
- **Naive Bayes.** Erstellt ein einzelnes NB-Modell und vergleicht dieses mit dem globalen Stichprobenvorgänger (die Verteilung der Zielwerte in der globalen Stichprobe). Das NB-Modell wird nur dann ausgegeben, wenn es sich für die Zielwerte als besserer Prädiktor erweist als der globale Vorgänger. Andernfalls wird das Modell nicht ausgegeben.

Adaptive Bayes - Expertenoptionen

Ausführungszeit beschränken. Über diese Option können Sie eine maximale Erstellungszeit in Minuten angeben. Damit können Sie Modelle in kürzerer Zeit erstellen, wenngleich das daraus resultierende Modell weniger genau sein kann. An jeder Etappe des Modellierungsverfahrens prüft der Algorithmus, bevor er fortfährt, ob er in der Lage ist, die nächste Etappe innerhalb der vorgegebenen Zeit abzuschließen, und liefert bei Erreichen der Zeitgrenze das beste verfügbare Modell zurück.

Max. Prädiktoren. Mit dieser Option können Sie die Komplexität des Modells einschränken und die Leistung verbessern, indem Sie die Anzahl der verwendeten Prädiktoren beschränken. Prädiktoren werden auf der Grundlage einer MDL-Messung ihrer Korrelation mit dem Ziel eingestuft. Diese Einstufung bestimmt die Wahrscheinlichkeit, dass sie in das Modell aufgenommen werden.

Max. Naive Bayes-Prädiktoren. Diese Option legt die maximale Anzahl der Prädiktoren fest, die im Naive Bayes-Modell verwendet werden.

Oracle Support Vector Machine (SVM)

Support Vector Machine (SVM) ist ein Klassifizierungs- und Regressionsalgorithmus, der eine Theorie des maschinellen Lernens verwendet, um die Vorhersagegenauigkeit zu maximieren, ohne die Daten übermäßig anzupassen. SVM verwendet eine optionale nicht lineare Transformation der Trainingsdaten, an die sich die Suche nach Regressionsgleichungen in den transformierten Daten anschließt, mit denen die Klassen getrennt werden (für kategoriale Ziele) oder das Ziel angepasst wird (für stetige Ziele). Die Oracle-Implementierung von SVM ermöglicht das Erstellen von Modellen unter Verwendung eines der zwei verfügbaren Kerne - des linearen oder des gaußschen Kerns. Der lineare Kern verzichtet auf die gesamte nicht lineare Transformation und liefert als Modellergebnis im Grunde ein Regressionsmodell.

Weitere Informationen finden Sie in den Veröffentlichungen *Oracle Data Mining Application Developer's Guide* und *Oracle Data Mining Concepts*.

Oracle SVM - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Aktives Lernen. Bietet eine Möglichkeit für den Umgang mit großen Aufbausets. Beim aktiven Lernen erstellt der Algorithmus ein erstes Modell anhand einer kleinen Stichprobe, bevor er das Modell auf das gesamte Trainingsdataset anwendet, und aktualisiert dann schrittweise die Stichprobe und das Modell anhand der Ergebnisse. Der Zyklus wird wiederholt, bis das Modell gegen die Trainingsdaten konvergiert oder bis die maximal zulässige Anzahl an Support-Vektoren erreicht wurde.

Kernfunktion. Wählen Sie **Linear** oder **Gaußsch** aus oder übernehmen Sie den Standard **Systembestimmt**, damit das System den geeignetsten Kern wählt. Gaußsche Kerne sind in der Lage, komplexere Beziehungen zu lernen, benötigen aber in der Regel mehr Rechenzeit. Sie können mit dem linearen Kern beginnen und den gaußschen Kern nur dann ausprobieren, wenn der lineare Kern keine gute Anpassung findet. Dies passiert häufiger mit einem Regressionsmodell, bei dem sich die Auswahl des Kerns stärker auswirkt. Denken Sie außerdem daran, dass mit dem gaußschen Kern erstellte SVM-Modelle nicht in IBM SPSS Modeler durchsucht werden können. Mit dem linearen Kern erstellte Modelle können genauso in IBM SPSS Modeler durchsucht werden wie Standardregressionsmodelle.

Normalisierungsmethode. Legt die Normalisierungsmethode für stetige Eingabe- und Zielfelder fest. Zur Auswahl stehen **Z-Score**, **Min-Max** oder **Keine**. Oracle führt die Normalisierung automatisch durch, wenn das Kontrollkästchen **Automatische Datenaufbereitung** aktiviert ist. Inaktivieren Sie dieses Kontrollkästchen, um die Normalisierungsmethode manuell auszuwählen.

Oracle SVM - Expertenoptionen

Kerncachegröße. Legt die Größe des während der Modellierung zum Speichern berechneter Kerne verwendbaren Caches in Byte fest. Es liegt auf der Hand, dass ein größerer Cache in der Regel zu einer schnelleren Modellierung führt. Der Standardwert ist 50MB.

Konvergenztoleranz. Legt den zulässigen Toleranzwert für den Abschluss der Modellierung fest. Der Wert muss zwischen 0 und 1 liegen. Der Standardwert ist 0,001. Höhere Werte führen zu schneller erstellten, aber weniger genauen Modellen.

Standardabweichung festlegen. Legt den Standardabweichungsparameter fest, der vom gaußschen Kern verwendet wird. Dieser Parameter wirkt sich auf das Verhältnis zwischen Modellkomplexität und der Möglichkeit der Verallgemeinerung auf andere Datasets aus (zu große und zu geringe Datenanpassung). Ein höherer Standardabweichungswert begünstigt eine zu geringe Anpassung. Standardmäßig wird dieser Parameter anhand der Trainingsdaten geschätzt.

Epsilon festlegen. Nur für Regressionsmodelle. Legt den Wert des Intervalls der Fehler fest, die bei der Bildung nicht Epsilon-sensitiver Modelle zulässig sind. Letztlich wird dadurch zwischen kleinen Fehlern (die ignoriert werden) und großen Fehlern (die nicht ignoriert werden) unterschieden. Der Wert muss zwischen 0 und 1 liegen. Standardmäßig wird dieser Wert aus den Trainingsdaten geschätzt.

Komplexitätsfaktor festlegen. Legt den Komplexitätsfaktor fest, der für den Ausgleich zwischen Modellfehler (gegen die Trainingsdaten gemessen) und Modellkomplexität sorgt und so eine zu große oder zu geringe Anpassung der Daten vermeidet. Ein höherer Wert stuft Fehler schwerwiegender ein, was das Risiko einer zu großen Anpassung der Daten birgt. Ein geringer Wert stuft Fehler weniger schwerwiegend ein und kann zu einer geringen Anpassung führen.

Ausreißerquote angeben. Gibt die gewünschte Quote an Ausreißern in den Trainingsdaten an. Nur gültig für SVM-Modellen mit einer einzigen Klasse. Die Verwendung mit der Einstellung **Komplexitätsfaktor festlegen** ist nicht möglich.

Vorhersagewahrscheinlichkeit. Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Vorhersage für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie **Auswählen** wählen, auf die Schaltfläche **Angeben** klicken, eines der möglichen Ergebnisse wählen und dann auf **Einfügen** klicken.

Vorhersageset verwenden. Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

Oracle SVM - Gewichtungsoptionen

In einem Klassifizierungsmodell können Sie mithilfe von Gewichtungen die relative Wichtigkeit der verschiedenen möglichen Zielwerte angeben. Dies kann beispielsweise dann sinnvoll sein, wenn die Datenpunkte in den Trainingsdaten nicht realistisch auf die verschiedenen Kategorien verteilt sind. Mit Gewichtungen können Sie das Modell verzerren, um einen Ausgleich für diejenigen Kategorien zu bewirken, die in den Daten unterrepräsentiert sind. Durch die Erhöhung der Gewichtung für einen Zielwert sollte der Prozentsatz der richtigen Vorhersagen für die betreffende Kategorie erhöht werden.

Es gibt drei Methoden zur Festlegung von Gewichtungen:

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

Die Summe der Gewichtungen aller Kategorien sollte den Wert 1,0 ergeben. Wenn sie keine Summe von 1,0 bilden, wird eine Warnnachricht ausgegeben und es besteht die Möglichkeit, die Werte automatisch normalisieren zu lassen. Diese automatische Anpassung behält die Anteile über die Kategorien hinweg bei, während die Gewichtungsbeschränkung erzwungen wird. Sie können diese Anpassung jederzeit durchführen, indem Sie auf die Schaltfläche **Normalisieren** klicken. Um die Tabelle auf gleiche Werte für alle Kategorien zurückzusetzen, klicken Sie auf die Schaltfläche **Gleichsetzen**.

Oracle GLM-Modelle

(Nur 11g) Verallgemeinerte lineare Modelle (GLM - Generalized Linear Model) lockern die restriktiven Annahmen der linearen Modelle. Dazu gehören beispielsweise die Annahmen, dass die Zielvariable eine Normalverteilung hat und dass die Wirkung der Prädiktoren auf die Zielvariable in ihrem Wesen linear ist. Ein verallgemeinertes lineares Modell eignet sich für Vorhersagen, in denen das Ziel wahrscheinlich eine nicht normale Verteilung hat, z. B. eine Multinomial- oder eine Poisson-Verteilung. Ebenso ist ein verallgemeinertes lineares Modell nützlich, wenn die Beziehung oder die Verknüpfung zwischen den Prädiktoren und dem Ziel wahrscheinlich nicht linear ist.

Weitere Informationen finden Sie in den Veröffentlichungen *Oracle Data Mining Application Developer's Guide* und *Oracle Data Mining Concepts*.

Oracle GLM - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Normalisierungsmethode. Legt die Normalisierungsmethode für stetige Eingabe- und Zielfelder fest. Zur Auswahl stehen **Z-Score**, **Min-Max** oder **Keine**. Oracle führt die Normalisierung automatisch durch, wenn

das Kontrollkästchen **Automatische Datenaufbereitung** aktiviert ist. Inaktivieren Sie dieses Kontrollkästchen, um die Normalisierungsmethode manuell auszuwählen.

Behandlung fehlende Werte. Gibt an, wie fehlende Werte in den Eingabedaten verarbeitet werden sollen:

- **Ersetzen durch Mittelwert oder Modalwert** ersetzt fehlende Werte von numerischen Attributen durch den Mittelwert und fehlende Werte von kategorialen Attributen durch den Modalwert.
- **Nur vollständige Datensätze verwenden** ignoriert Datensätze, in denen Werte fehlen.

Oracle GLM - Expertenoptionen

Zeilengewichtungen verwenden. Markieren Sie dieses Kontrollkästchen, um die benachbarte Dropdown-Liste zu aktivieren, aus der Sie eine Spalte mit einem Gewichtungsfaktor für die Zeilen wählen können.

Zeilendiagnose in Tabelle speichern. Markieren Sie dieses Kontrollkästchen, um das benachbarte Textfeld zu aktivieren, in das Sie den Namen einer Tabelle eingeben können, die Diagnosedaten auf Zeilenebene enthalten soll.

Koeffizientkonfidenzniveau. Der Sicherheitsgrad (von 0,0 bis 1,0), in dem der vorhergesagte Wert für das Ziel innerhalb eines Konfidenzintervalls liegt, das vom Modell berechnet wurde. Konfidenzgrenzen werden mit den Koeffizientenstatistiken zurückgegeben.

Referenzkategorie für Ziel. Wählen Sie **Benutzerdefiniert** aus, um für das Zielfeld einen Wert als Referenzkategorie zu verwenden, oder behalten Sie den Standardwert **Auto**.

Ridge-Regression. Bei der Ridge-Regression handelt es sich um eine Technik zur Kompensierung der Situation, in der ein zu hoher Korrelationsgrad bei den Variablen besteht. Mithilfe der Option **Auto** können Sie erlauben, dass der Algorithmus diese Technik verwendet. Sie können sie aber auch manuell über die Optionen **Inaktivieren** und **Aktivieren** steuern. Wenn Sie sich für die manuelle Aktivierung der Ridge-Regression entscheiden, können Sie den Standardwert des Systems für den Rigde-Parameter überschreiben, indem Sie einen Wert in das benachbarte Feld eingeben.

VIF für Ridge-Regression generieren. Markieren Sie dieses Kontrollkästchen zur Erzeugung von VIF-Statistiken (Variance Inflation Factor), wenn die Ridge für lineare Regression verwendet wird.

Vorhersagewahrscheinlichkeit. Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Vorhersage für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie **Auswählen** wählen, auf die Schaltfläche **Angeben** klicken, eines der möglichen Ergebnisse wählen und dann auf **Einfügen** klicken.

Vorhersageset verwenden. Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

Oracle GLM - Gewichtungsoptionen

In einem Klassifizierungsmodell können Sie mithilfe von Gewichtungen die relative Wichtigkeit der verschiedenen möglichen Zielwerte angeben. Dies kann beispielsweise dann sinnvoll sein, wenn die Datenpunkte in den Trainingsdaten nicht realistisch auf die verschiedenen Kategorien verteilt sind. Mit Gewichtungen können Sie das Modell verzerren, um einen Ausgleich für diejenigen Kategorien zu bewirken, die in den Daten unterrepräsentiert sind. Durch die Erhöhung der Gewichtung für einen Zielwert sollte der Prozentsatz der richtigen Vorhersagen für die betreffende Kategorie erhöht werden.

Es gibt drei Methoden zur Festlegung von Gewichtungen:

- **Auf Trainingsdaten basierend.** Dies ist die Standardeinstellung. Die Gewichtungen basieren auf den relativen Häufigkeiten der Kategorien in den Trainingsdaten.
- **Für alle Klassen gleich.** Gewichtungen für alle Kategorien werden als $1/k$ definiert, wobei k die Zahl der Zielkategorien darstellt.
- **Benutzerdefiniert.** Sie können eigene Gewichtungen angeben. Die Startwerte für Gewichtungen werden für alle Klassen gleich gesetzt. Sie können die Gewichtungen für einzelne Kategorien auf benutzer-

definierte Werte einstellen. Um die Gewichtung einer bestimmten Kategorie anzupassen, wählen Sie in der Tabelle die Gewichtungszelle aus, die der gewünschten Kategorie entspricht, löschen den Inhalt der Zelle und geben den gewünschten Wert ein.

Die Summe der Gewichtungen aller Kategorien sollte den Wert 1,0 ergeben. Wenn sie keine Summe von 1,0 bilden, wird eine Warnnachricht ausgegeben und es besteht die Möglichkeit, die Werte automatisch normalisieren zu lassen. Diese automatische Anpassung behält die Anteile über die Kategorien hinweg bei, während die Gewichtungsbeschränkung erzwungen wird. Sie können diese Anpassung jederzeit durchführen, indem Sie auf die Schaltfläche **Normalisieren** klicken. Um die Tabelle auf gleiche Werte für alle Kategorien zurückzusetzen, klicken Sie auf die Schaltfläche **Gleichsetzen**.

Oracle Decision Tree

Oracle Data Mining bietet eine klassische Entscheidungsbaumfunktion, die auf dem beliebten Algorithmus mit Klassifizierungs- und Regressionsbäumen beruht. Das ODM-Entscheidungsbaummodell enthält vollständige Informationen zu jedem Knoten, einschließlich Konfidenz, Support und Splitting-Kriterium. Für jeden Knoten kann die vollständige Regel angezeigt werden. Außerdem wird für jeden Knoten ein Ersatzattribut angegeben, das verwendet wird, wenn das Modell auf einen Fall mit fehlenden Werten angewendet wird.

Entscheidungsbäume sind beliebt, weil sie universell einsetzbar, leicht anzuwenden und leicht zu verstehen sind. Entscheidungsbäume sichten alle potenziellen Eingabeattribute auf der Suche nach dem besten "Splitter", also dem besten Trennwert für Attribute (z. B. `ALTER > 55`), der die nachgeordneten Datensätze in homogenere Grundgesamtheiten aufteilt. Bei jeder Split-Entscheidung wiederholt ODM den Prozess, indem der gesamte Baum erweitert wird und "Endblätter" erstellt werden, die ähnliche Grundgesamtheiten von Datensätzen, Elementen bzw. Personen darstellen. Ausgehend vom Stammknoten des Baums (z. B. der gesamten Grundgesamtheit) bieten Entscheidungsbäume von Menschen lesbare Regeln mit Anweisungen vom Typ `WENN A, dann B`. Diese Entscheidungsbaumregeln geben außerdem Support und Konfidenz für jeden Baumknoten an.

Auch Adaptive Bayes Networks können kurze und einfache Regeln bieten, die dazu beitragen können, Erklärungen für jede Vorhersage zu finden, Entscheidungsbäume jedoch bieten vollständige Oracle Data Mining-Regeln für jede Aufteilungsentscheidung. Entscheidungsbäume sind außerdem hilfreich bei der Entwicklung detaillierter Profile für die besten Kunden, gesunde Patienten, Faktoren im Zusammenhang mit Betrug usw.

Entscheidungsbaum - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Unreinheitsmetrik. Gibt an, welche Metrik für die Ermittlung der besten Testfrage für die Aufteilung der Daten an den einzelnen Knoten verwendet wird. Der beste Splitter und Split-Wert sind diejenigen, die zu der größten Zunahme an Zielwerthomogenität für die Elemente im Knoten führen. Homogenität wird in Übereinstimmung mit einer Metrik gemessen. Die Metriken **Gini** und **Entropie** werden unterstützt.

Entscheidungsbäume - Expertenoptionen

Maximale Tiefe. Legt die maximale Tiefe des zu erstellenden Baummodells fest.

Mindestprozentsatz der Datensätze in einem Knoten. Legt den Prozentsatz der Mindestanzahl an Datensätzen pro Knoten fest.

Mindestprozentsatz der Datensätze für eine Aufteilung. Legt die Mindestanzahl an Datensätzen in einem übergeordneten Knoten als Prozentsatz der Gesamtzahl der zum Trainieren des Modells verwendeten Datensätze fest. Es wird nicht versucht, einen Split durchzuführen, wenn die Anzahl der Datensätze unterhalb dieses Prozentsatzes liegt.

Mindestdatensätze in einem Knoten. Legt die Mindestanzahl an auszugebenden Datensätzen fest.

Mindestdatensätze für eine Aufteilung. Legt die Mindestanzahl der Datensätze in einem übergeordneten Knoten als Wert fest. Es wird nicht versucht, einen Split durchzuführen, wenn die Anzahl der Datensätze unterhalb dieses Werts liegt.

Regel-ID. Wenn diese Option aktiviert ist, wird eine Zeichenfolge in das Modell aufgenommen, die den Knoten im Baum angibt, bei dem eine bestimmte Aufteilung (Split) vorgenommen werden soll.

Vorhersagewahrscheinlichkeit. Ermöglicht dem Modell, die Wahrscheinlichkeit einer korrekten Vorhersage für ein mögliches Ergebnis des Zielfelds zu umfassen. Sie aktivieren diese Option, indem Sie **Auswählen** wählen, auf die Schaltfläche **Angeben** klicken, eines der möglichen Ergebnisse wählen und dann auf **Einfügen** klicken.

Vorhersageset verwenden. Generiert eine Tabelle von allen möglichen Resultaten für alle möglichen Ergebnisse des Zielfelds.

Oracle O-Cluster

Der Algorithmus "Oracle-O-Cluster" identifiziert natürlich vorkommende Gruppierungen in einer Datengesamtheit. Clustering mit orthogonaler Partitionierung (O-Cluster) ist ein Oracle-eigener Clustering-Algorithmus, der ein auf einem hierarchischen Raster beruhendes Clustering-Modell erstellt, d. h., es erstellt achsenparallele (orthogonale) Partitionen im Bereich des Eingabeattributraums. Der Algorithmus arbeitet rekursiv. Die entstehende hierarchische Struktur stellt ein unregelmäßiges Raster dar, das den Attributraum in Cluster zerlegt.

Der O-Clusteralgorithmus kann sowohl numerische als auch kategoriale Attribute verarbeiten und ODM wählt automatisch die besten Clusterdefinitionen aus. ODM stellt Informationen zu Clusterdetails und Clusterregeln sowie Werte für den Clusterschwerpunkt (Zentroid) bereit und kann zum Scoren einer Grundgesamtheit in Bezug auf ihre Clusterzugehörigkeit verwendet werden.

O-Cluster - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Maximale Anzahl von Clustern. Legt die maximale Anzahl der generierten Cluster fest.

O-Cluster - Expertenoptionen

Maximaler Puffer. Legt die maximale Puffergröße fest.

Sensitivität. Legt einen Anteil fest, der die für die Abtrennung eines neuen Clusters erforderliche Spitzendichte angibt. Der Anteil steht in Bezug zur globalen einheitlichen Dichte.

Oracle-K-Means

Der Algorithmus "Oracle-K-Means" identifiziert natürlich vorkommende Gruppierungen in einer Datengesamtheit. Der K-Means-Algorithmus ist ein distanzbasierter Clusteralgorithmus, der die Daten in eine zuvor festgelegte Anzahl an Clustern einteilt (vorausgesetzt, dass genügend unterschiedliche Fälle vorhanden sind). Distanzbasierte Algorithmen beruhen auf einer Distanzmetrik (Funktion) zur Messung der Ähnlichkeit zwischen Datenpunkten. Datenpunkte werden dem nächsten Cluster gemäß der verwendeten Distanzmetrik zugewiesen. ODM bietet eine erweiterte Version von K-Means.

Der K-Means-Algorithmus unterstützt hierarchische Cluster, verarbeitet numerische und kategoriale Attribute und teilt die Grundgesamtheit in die vom Benutzer angegebene Anzahl an Clustern auf. ODM stellt Informationen zu Clusterdetails und Clusterregeln sowie Werte für den Clusterschwerpunkt (Zentroid) bereit und kann zum Scoren einer Grundgesamtheit in Bezug auf ihre Clusterzugehörigkeit verwendet werden.

K-Means - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Anzahl der Cluster. Legt die Anzahl der generierten Cluster fest.

Distanzfunktion. Gibt an, welche Distanzfunktion für K-Means-Clustering verwendet wird.

Split-Kriterium. Gibt an, welches Split-Kriterium für K-Means-Clustering verwendet wird.

Normalisierungsmethode. Legt die Normalisierungsmethode für stetige Eingabe- und Zielfelder fest. Zur Auswahl stehen **Z-Score**, **Min-Max** oder **Keine**.

K-Means - Expertenoptionen

Iterationen. Legt die Anzahl der Iterationen für den K-Means-Algorithmus fest.

Konvergenztoleranz. Legt die Konvergenztoleranz für den K-Means-Algorithmus fest.

Anzahl der Klassen. Gibt die Anzahl der Klassen in dem von K-Means erstellten Attributhistogramm an. Die Klassengrenzen für die einzelnen Attribute werden global für das gesamte Trainingsdataset berechnet. Die Klassiermethode ist "Gleiche Breite". Alle Attribute haben dieselbe Anzahl von Klassen, mit Ausnahme von Attributen mit einem einzelnen Wert, die nur eine einzige Klasse aufweisen.

Blockerweiterung. Legt den Erweiterungsfaktor für Arbeitsspeicher fest, der für die Aufnahme der Clusterdaten zugeordnet wird.

Mindestprozentsatz für Attributsupport. Legt den Anteil der Attributwerte fest, die nicht null sein müssen, damit das Attribut in die Regelbeschreibung für den Cluster aufgenommen wird. Wenn der Parameterwert bei Daten mit fehlenden Werten zu hoch festgelegt wird, kann dies zu sehr kurzen oder sogar leeren Regeln führen.

Oracle-NMF (Nonnegative Matrix Factorization)

NMF (Nonnegative Matrix Factorization) dient zur Verkleinerung eines großen Datasets in repräsentative Attribute. NMF ähnelt vom Konzept her der Hauptkomponentenanalyse (Principal Components Analysis, PCA), kann jedoch mit größeren Attributmengen und einem additiven Darstellungsmodell umgehen und ist somit ein leistungsstarker, hochmoderner Data-Mining-Algorithmus, der für eine Vielzahl von Verwendungsfällen eingesetzt werden kann.

NMF kann verwendet werden, um große Datenmengen, beispielsweise Textdaten, in kleinere, dünner besetzte Darstellungen zu reduzieren, die die Dimensionalität der Daten verringern (dieselben Informationen können unter Verwendung von wesentlich weniger Variablen beibehalten werden). Die Ausgabe der NMF-Modelle kann mithilfe von Techniken für überwachtes Lernen, wie SVMs, oder nicht überwachtes Lernen, wie Clustering-Verfahren, analysiert werden. Oracle Data Mining verwendet NMF- und SVM-Algorithmen für das Mining von unstrukturierten Textdaten.

NMF - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Normalisierungsmethode. Legt die Normalisierungsmethode für stetige Eingabe- und Zielfelder fest. Zur Auswahl stehen **Z-Score**, **Min-Max** oder **Keine**. Oracle führt die Normalisierung automatisch durch, wenn das Kontrollkästchen **Automatische Datenaufbereitung** aktiviert ist. Inaktivieren Sie dieses Kontrollkästchen, um die Normalisierungsmethode manuell auszuwählen.

NMF - Expertenoptionen

Anzahl der Merkmale angeben. Dient zur Angabe der Anzahl der zu extrahierenden Merkmale.

Startwert für Zufallszahlen. Legt den Zufallsstartwert für den NMF-Algorithmus fest.

Anzahl der Iterationen. Legt die Anzahl der Iterationen für den NMF-Algorithmus fest.

Konvergenztoleranz. Legt die Konvergenztoleranz für den NMF-Algorithmus fest.

Alle Merkmale anzeigen. Zeigt die Merkmal-ID und die Konfidenz für alle Merkmale an und nicht nur die Werte für das beste Merkmal.

Oracle Apriori

Der Apriori-Algorithmus erkennt Assoziationsregeln in den Daten. Beispiel: "Falls ein Kunde einen Rasierer und After-Shave-Lotion kauft, dann kauft er auch mit einer Wahrscheinlichkeit von 80 % Rasiercreme." Das Association-Mining-Problem lässt sich in zwei Teilprobleme zerlegen:

- Ermitteln aller Elementkombinationen, der sogenannten "Frequent Itemsets" (häufig vorkommende Elementmengen), deren Support größer ist als der minimale Support.
- Verwenden der Frequent Itemsets zum Generieren der gewünschten Regeln. Es gilt also: Wenn ABC und BC häufig vorkommen, dann gilt die Regel "A impliziert BC" immer dann, wenn das Verhältnis von `support(ABC)` zu `support(BC)` mindestens so groß ist wie die minimale Konfidenz. Beachten Sie, dass die Regel über den Mindestsupport verfügt, da ABCD häufig vorkommt. ODM Association unterstützt nur Regeln mit einem einzigen Sukzedens (ABC impliziert D).

Die Anzahl der Frequent Itemsets richtet sich nach den Parametern für den minimalen Support. Die Anzahl der generierten Regeln richtet sich nach der Anzahl der Frequent Itemsets und dem Konfidenzparameter. Wenn der Konfidenzparameter zu hoch festgelegt ist, kann es vorkommen, dass zwar Frequent Itemsets im Assoziationsmodell vorliegen, aber keine Regeln.

ODM verwendet eine SQL-basierte Implementierung des Apriori-Algorithmus. Die Schritte zur Kandidatengenerierung und zur Support-Zählung werden mithilfe von SQL-Abfragen implementiert. Es werden keine spezialisierten, arbeitsspeicherinternen Datenstrukturen verwendet. Die SQL-Abfragen werden durch verschiedene Hinweise so optimiert, dass sie effizient auf dem Datenbankserver ausgeführt werden.

Apriori - Feldoptionen

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

Bevor Sie ein "Apriori"-Model erstellen können, müssen Sie festlegen, welche Felder als relevante Elemente bei der Assoziationsmodellierung verwendet werden sollen.

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

Benutzerdefinierte Einstellungen verwenden. Diese Option weist den Knoten an, die hier angegebenen Feldinformationen zu verwenden und nicht die in einem vorgelagerten Typknoten angegebenen. Geben Sie nach Auswahl dieser Option die restlichen Felder im Dialogfeld an. (Diese hängen davon ab, ob Sie das Transaktionsformat verwenden.)

Wenn Sie das Transaktionsformat *nicht verwenden*, geben Sie Folgendes an:

- **Eingaben.** Wählen Sie die Eingabefelder aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.
- **Partition.** In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen.

Wenn Sie das Transaktionsformat *verwenden*, geben Sie Folgendes an:

Transaktionsformat verwenden. Verwenden Sie diese Option, wenn Sie Daten so transformieren möchten, dass nicht mehr eine Zeile pro Element, sondern eine Zeile pro Fall verwendet wird.

Durch Auswahl dieser Option werden die Feldsteuerelemente im unteren Bereich dieses Dialogfelds verändert:

Geben Sie beim Transaktionsformat Folgendes an:

- **ID.** Wählen Sie ein ID-Feld aus der Liste aus. Als ID-Feld können numerische oder symbolische Felder verwendet werden. Jeder eindeutige Wert in diesem Feld sollte eine bestimmte Analyseeinheit darstellen. Bei einer Warenkorbanwendung könnte z. B. jede ID einen einzelnen Kunden darstellen. Für eine Webprotokoll-Analyseanwendung könnte jede ID einen Computer (nach IP-Adresse) oder einen Benutzer (nach Anmeldedaten) darstellen.
- **Inhalt.** Geben Sie das Inhaltsfeld für das Modell an. Dieses Feld enthält das Element, das für die Assoziationsmodellierung relevant ist.
- **Partition.** In diesem Feld können Sie ein Feld angeben, das verwendet wird, um die Daten in getrennte Stichproben für die Trainings-, Test- und Validierungsphase der Modellbildung aufzuteilen. Indem Sie mit einer Stichprobe das Modell erstellen und es mit einer anderen Stichprobe testen, erhalten Sie ei-

nen guten Hinweis dafür, wie gut das Modell sich für größere Datasets verallgemeinern lässt, die den aktuellen Daten ähneln. Wenn mehrere Partitionsfelder mithilfe von Typ- oder Partitionsknoten erstellt wurden, muss in jedem Modellierungsknoten, der die Partitionierung verwendet, auf der Registerkarte "Felder" ein einzelnes Partitionsfeld ausgewählt werden. (Wenn nur eine einzige Partition vorhanden ist, wird diese immer automatisch verwendet, wenn die Partitionierung aktiviert ist.) Beachten Sie außerdem, dass die Partitionierung auch auf der Registerkarte "Modelloptionen" des Knotens aktiviert sein muss, wenn die ausgewählte Partitionierung in Ihrer Analyse angewendet werden soll. (Wenn die Auswahl der Option aufgehoben ist, kann die Partitionierung ohne Änderung der Feldeinstellungen inaktiviert werden.)

Apriori - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Maximale Regellänge. Legt die maximale Anzahl an Vorbedingungen für jede beliebige Regel als ganze Zahl von 2 bis 20 fest. Auf diese Weise können Sie die Komplexität der Regeln begrenzen. Wenn Regeln zu komplex oder zu spezifisch sind oder das Training des Regelsets zu viel Zeit in Anspruch nimmt, sollten Sie diese Einstellung reduzieren.

Minimale Konfidenz. Legt das minimale Konfidenzniveau mit einem Wert zwischen 0 und 1 fest. Regeln mit einer niedrigeren Konfidenz als das angegebene Kriterium werden verworfen.

Minimale Unterstützung. Legt die Untergrenze für Support als Wert zwischen 0 und 1 fest. Apriori erkennt Muster mit einer Häufigkeit über der Untergrenze für Support.

Oracle Minimum Description Length (MDL)

Der Algorithmus "Oracle Minimum Description Length (MDL)" dient zur Ermittlung der Attribute, die den größten Einfluss auf ein Zielattribut haben. Wenn Sie wissen, welche Attribute den größten Einfluss haben, haben Sie häufig einen besseren Einblick in Ihre Geschäftstätigkeiten, können diese besser verwalten und einfacher Aktivitäten modellieren. Außerdem können diese Attribute die Datentypen anzeigen, durch deren Hinzufügung Sie Ihre Modelle erweitern können. MDL kann verwendet werden, um zum Beispiel die Prozessattribute zu ermitteln, die für die Vorhersage der Qualität eines hergestellten Bauteils, der mit Kundenabwanderung verbundenen Faktoren oder der Gene, die mit der größten Wahrscheinlichkeit in die Behandlung einer bestimmten Krankheit eingebunden sind, am relevantesten sind.

Oracle-MDL verwirft Eingabefelder, die sie für die Vorhersage des Ziels als unwichtig erachtet. Mit den verbleibenden Eingabefeldern erstellt sie dann ein nicht verfeinertes Modellnugget, das mit einem Oracle-Modell verknüpft und in Oracle Data Miner sichtbar ist. Bei der Darstellung des Modells in Oracle Data Miner wird ein Diagramm angezeigt, das die übrigen Eingabefelder in der Reihenfolge ihrer Bedeutung zur Vorhersage des Ziels aufführt.

Negative Rangfolge bedeutet Rauschen. Eingabefelder, die bei null oder darunter eingeordnet werden, tragen nicht zur Vorhersage bei und sollten wahrscheinlich aus den Daten entfernt werden.

So zeigen Sie das Diagramm an:

1. Klicken Sie mit der rechten Maustaste auf das nicht verfeinerte Modellnugget in der Modellpalette und wählen Sie **Durchsuchen** aus.
2. Klicken Sie im Modellfenster auf die Schaltfläche zum Start von Oracle Data Miner.
3. Stellen Sie eine Verbindung zu Oracle Data Miner her. Weitere Informationen finden Sie im Thema „Oracle Data Miner" auf Seite 48.
4. Erweitern Sie im Oracle Data Miner-Navigationsfenster den Bereich **Modelle** und dann **Attribut-Wichtigkeit**.
5. Wählen Sie das relevante Oracle-Modell aus (es hat denselben Namen wie in IBM SPSS Modeler angegebene Zielfeld). Wenn Sie nicht sicher sind, ob es das korrekte Modell ist, wählen Sie den Ordner "Attribut-Wichtigkeit" aus und suchen Sie ein Modell nach Erstellungsdatum.

MDL - Modelloptionen

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

Eindeutiges Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. IBM SPSS Modeler schreibt vor, dass dieses Schlüsselfeld zwingend numerisch sein muss.

Anmerkung: Dieses Feld ist für alle Oracle-Knoten mit Ausnahme von Oracle Adaptive Bayes, Oracle-O-Cluster und Oracle-Apriori optional.

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Oracle Attribute Importance (AI)

Das Ziel der Attributwichtigkeit ist es, die Attribute im Dataset zu finden, die mit dem Ergebnis zusammenhängen, sowie das Maß, in dem sie das Endergebnis beeinflussen. Der Knoten "Oracle Attribute Importance" analysiert Daten, findet Muster und sagt Ergebnisse mit einem entsprechenden Niveau an Zuverlässigkeit voraus.

Alle Modelloptionen

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

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

Autom. Datenaufbereitung. (Nur 11*g*) Aktiviert (Standard) oder inaktiviert den automatisierten Datenaufbereitungsmodus von Oracle Data Mining. Wenn dieses Kästchen markiert ist, führt ODM automatisch die vom Algorithmus geforderten Datenumformungen durch. Weitere Informationen finden Sie in *Oracle Data Mining Concepts*.

Alle Auswahloptionen

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

Die folgenden Optionen sind verfügbar:

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

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

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

Das Zielfeld bleibt unabhängig von der Auswahl immer erhalten.

AI-Modellnugget - Registerkarte "Modell"

Auf der Registerkarte "Modell" für ein Oracle AI-Modellnugget werden die Rangwertung und die Bedeutsamkeit für alle Eingaben angezeigt. Außerdem haben Sie die Möglichkeit, mithilfe der Kontrollkästchen in der Spalte auf der linken Seite Felder für die Filterung auszuwählen. Bei der Ausführung des Streams werden nur die aktivierten Felder sowie die Zielvorhersage beibehalten. Die anderen Eingabefelder werden verworfen. Die Standardauswahl beruht auf den im Modellierungsknoten angegebenen Optionen, Sie können jedoch nach Bedarf weitere Felder auswählen bzw. deren Auswahl aufheben.

- Um die Liste nach Rang, Feldname, Wichtigkeit oder einer anderen der angezeigten Spalten zu sortieren, klicken Sie auf die Spaltenüberschrift. Wählen Sie alternativ das gewünschte Element neben der Schaltfläche "Sortieren nach" aus. Mit den nach unten bzw. oben zeigenden Pfeilen können Sie die Sortierrichtung ändern.
- Sie können mithilfe der Symbolleiste alle Felder aktivieren bzw. inaktivieren und auf das Dialogfeld "Felder markieren" zugreifen, in dem Sie Felder nach Rangordnung oder Wichtigkeit auswählen können. Zum Erweitern der Auswahl können Sie auch die Umschalt- oder Steuertaste drücken, während Sie auf Felder klicken.
- Die Schwellenwerte für die Einordnung von Eingaben als "bedeutsam", "marginal" bzw. "unbedeutend" werden in der Legende unterhalb der Tabelle angezeigt. Diese Werte werden im Modellierungsknoten angegeben.

Verwalten von Oracle-Modellen

Oracle-Modelle werden genau wie andere IBM SPSS Modeler-Modelle zur Modellpalette hinzugefügt und können fast genauso verwendet werden. Es gibt jedoch einige wichtige Unterschiede, die sich daraus ergeben, dass zurzeit jedes in IBM SPSS Modeler erstellte Oracle-Modell auf ein in einem Datenbank-Server gespeichertes Modell verweist.

Oracle-Modellnugget - Registerkarte "Server"

Bei der Bildung eines ODM-Modells mit IBM SPSS Modeler wird einerseits in IBM SPSS Modeler ein Modell erzeugt und andererseits in der Oracle-Datenbank ein Modell erzeugt oder ersetzt. Ein IBM SPSS Modeler-Modell dieser Art verweist auf den Inhalt eines in einem Datenbankserver gespeicherten Datenbankmodells. IBM SPSS Modeler kann eine Konsistenzprüfung durchführen, indem sowohl im IBM SPSS Modeler-Modell als auch im Oracle-Modell eine identische, generierte **Motellschlüsselzeichenfolge** gespeichert wird.

Die Schlüsselzeichenfolge wird für jedes Oracle-Modell im Dialogfeld "Modelle auflisten" in der Spalte *Modellinformationen* angezeigt. Die Schlüsselzeichenfolge eines IBM SPSS Modeler-Modells wird auf der Registerkarte "Server" eines IBM SPSS Modeler-Modells (wenn es sich in einem Stream befindet) als **Modellschlüssel** ausgegeben.

Mit der Schaltfläche "Überprüfen" auf der Registerkarte "Server" eines Modellnuggets wird ermittelt, ob die Modellschlüssel des IBM SPSS Modeler-Modells und des Oracle-Modells übereinstimmen. Wenn in Oracle kein Modell mit demselben Namen gefunden wird oder wenn der Modellschlüssel nicht übereinstimmt, bedeutet dies, dass das Oracle-Modell seit der Bildung des IBM SPSS Modeler-Modells gelöscht oder neu erstellt wurde.

Oracle-Modellnugget - Registerkarte "Übersicht"

Auf der Registerkarte "Übersicht" eines Modellnuggets finden Sie Informationen zum Modell selbst (*Analyse*), den im Modell verwendeten Feldern (*Felder*), den beim Erstellen des Modells verwendeten Einstellungen (*Aufbaueinstellungen*) und zum Modelltraining (*Trainingsübersicht*).

Beim ersten Durchsuchen des Knotens sind die Ergebnisse der Registerkarte "Übersicht" reduziert. Um die für Sie relevanten Ergebnisse anzuzeigen, vergrößern Sie sie mithilfe des Erweiterungssteuerelements links neben dem betreffenden Element oder klicken Sie auf die Schaltfläche **Alles anzeigen**, um alle Ergebnisse anzuzeigen. Um die Ergebnisse nach der Betrachtung wieder auszublenden, können Sie mit dem Erweiterungssteuerelement die gewünschten Ergebnisse reduzieren. Alternativ können Sie mit der Schaltfläche **Alles ausblenden** alle Ergebnisse ausblenden.

Analyse. Zeigt Informationen zum jeweiligen Modell an. Wenn Sie einen Analyseknoten ausgeführt haben, der an dieses Modellnugget angehängt ist, werden die Informationen aus dieser Analyse ebenfalls in diesem Abschnitt angezeigt.

Felder. Listet die Felder auf, die bei der Erstellung des Modells als Ziel bzw. als Eingaben verwendet werden.

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

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

Oracle-Modellnugget - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" des Modellnuggets können Sie die Einstellung bestimmter Optionen für den Modellierungsknoten zu Scoring-Zwecken überschreiben.

Oracle Decision Tree

Fehlklassifizierungskosten verwenden. Bestimmt, ob Fehlklassifizierungskosten im Oracle-Entscheidungsbaummodell verwendet werden. Weitere Informationen finden Sie im Thema „Fehlklassifizierungskosten” auf Seite 32.

Regel-ID. Wenn ausgewählt (markiert), wird dem Oracle-Entscheidungsbaummodell eine Regel-ID-Spalte hinzugefügt. Die Regel-ID identifiziert den Knoten in der Struktur, an dem eine bestimmte Aufteilung erfolgt.

Oracle NMF

Alle Merkmale anzeigen. Wenn ausgewählt (markiert), wird die Merkmal-ID und die Konfidenz für alle Merkmale im Oracle-NMF-Modell angezeigt und nicht nur die Werte für das beste Merkmal.

Auflisten der Oracle-Modelle

Die Schaltfläche "Oracle Data Mining-Modelle auflisten" öffnet ein Dialogfeld, in dem die vorhandenen Datenbankmodelle aufgelistet sind und entfernt werden können. Der Zugriff auf dieses Dialogfeld erfolgt über das Dialogfeld "Hilfsanwendungen" und über die Dialogfelder zum Erstellen, Suchen und Anwenden für mit ODM verbundene Knoten.

Zu jedem Modell werden folgende Informationen angezeigt:

- **Modellname.** Name des Modells, das zum Sortieren der Liste verwendet wird
- **Modellinformationen.** Modellschlüsselinformationen, die sich aus dem Datum und der Uhrzeit der Erstellung sowie dem Namen der Zielspalte zusammensetzen
- **Modelltyp.** Name des Algorithmus, der dieses Modell erstellt hat

Oracle Data Miner

Oracle Data Miner ist die Benutzerschnittstelle für Oracle Data Mining (ODM) und ersetzt die frühere IBM SPSS Modeler-Benutzerschnittstelle für ODM. Oracle Data Miner dient zur Erhöhung der Erfolgsquote des Analysten bei der ordnungsgemäßen Nutzung der ODM-Algorithmen. Es werden verschiedene Methoden verwendet, um diese Ziele zu erreichen:

- Die Benutzer benötigen mehr Unterstützung bei der Anwendung einer Methodologie, die sich sowohl mit der Datenaufbereitung als auch mit der Algorithmenauswahl befasst. Oracle Data Miner wird dem durch Bereitstellung von Data-Mining-Aktivitäten gerecht, mit denen die Benutzer Schritt für Schritt durch die jeweilige Methodologie geführt werden.
- Oracle Data Miner beinhaltet verbesserte und erweiterte Heuristiken in den Modellerstellungs- und Transformationsassistenten, um die Fehlerwahrscheinlichkeit bei der Angabe von Modell- und Transformationseinstellungen zu verringern.

Definieren einer Oracle Data Miner-Verbindung

1. Oracle Data Miner kann von allen Oracle-Dialogfeldern für Erstellung, Knotenanwendung und Ausgabe mithilfe der Schaltfläche **Oracle Data Miner starten** gestartet werden.

   

   Abbildung 2. Schaltfläche "Oracle Data Miner starten"
2. Das Oracle Data Miner-Dialogfeld **Edit Connection** wird dem Benutzer angezeigt, bevor die externe Oracle Data Miner-Anwendung gestartet wird (vorausgesetzt, die Option für Hilfsanwendungen wurde ordnungsgemäß definiert).

   Hinweis: Dieses Dialogfeld wird nur angezeigt, wenn kein definierter Verbindungsname vorhanden ist.

   - Geben Sie einen Data Miner-Verbindungsnamen und die entsprechenden Informationen für den Oracle 10gR1- bzw. 10gR2-Server ein. Bei dem Oracle-Server sollte es sich um denselben Server handeln, der auch in IBM SPSS Modeler angegeben wurde.
3. Das Dialogfeld **Choose Connection** des Oracle Data Miner bietet Optionen, in denen Sie angeben können, welcher Verbindungsname (im Schritt weiter oben definiert) verwendet werden soll.

Unter Oracle Data Miner auf der Oracle-Website finden Sie weitere Informationen zu Anforderungen, Installation und Verwendung von Oracle Data Miner.

Vorbereitung der Daten

Wenn Sie bei der Modellierung die mit Oracle Data Mining gelieferten Algorithmen Naive Bayes, Adaptive Bayes und Support Vector Machine verwenden, können sich zwei Arten der Datenaufbereitung als nützlich erweisen:

- **Klassierung** oder die Konvertierung fortlaufender numerischer Bereichsfelder in Kategorien für Algorithmen, die keine fortlaufenden Daten annehmen.
- **Normalisierung** oder für numerische Bereiche durchgeführte Transformationen, die dafür sorgen, dass diese ähnliche Bedeutungen und Standardabweichungen besitzen.

Klassierung

Der Klassierungsknoten von IBM SPSS Modeler bietet diverse Verfahren für Klassierungsoperationen. Eine Klassieroperation ist definiert und kann auf eines oder mehrere Felder angewendet werden. Durch Ausführung der Klassieroperation für ein Dataset werden die Schwellenwerte erstellt und das Erstellen eines IBM SPSS Modeler-Ableitungsknotens ermöglicht. Die Ableitungsoperation kann in SQL konvertiert werden und vor der Erstellung und dem Scoring des Modells angewendet werden. Dieser Ansatz erstellt eine Abhängigkeit zwischen dem Modell und dem die Klassierung durchführenden Ableitungsknoten, erlaubt aber, dass die Klassierspezifikationen von verschiedenen Modellierungsaufgaben wiederverwendet werden.

Normalisierung

Stetige Felder (numerischer Bereich), die als Eingabe für SVM-Modelle verwendet werden, sollten vor der Modellierung normalisiert werden. Bei einem Regressionsmodell muss die Normalisierung außerdem umgekehrt werden, um den Score der Modellausgabe wiederherzustellen. Als SVM-Modelleinstellungen stehen **Z-Score**, **Min-Max** oder **Keine** zur Auswahl. Die Normalisierungskoeffizienten werden von Oracle im Rahmen des Modellierungsvorgangs erzeugt und dann in IBM SPSS Modeler geladen, wo sie zusammen mit dem Modell gespeichert werden. Zum Zeitpunkt der Anwendung werden die Koeffizienten in IBM SPSS Modeler-Ableitungsausdrücke konvertiert und zur Vorbereitung der Daten für das Scoring verwendet, bevor diese an das Modell übergeben werden. In diesem Fall ist die Normalisierung eng mit der Modellierungsaufgabe verbunden.

Beispiele für Oracle Data Mining

Im Lieferumfang sind einige Beispielstreams enthalten, die die Verwendung von ODM mit IBM SPSS Modeler demonstrieren. Diese Streams befinden sich im IBM SPSS Modeler-Installationsverzeichnis unter *\Demos\Database_Modelling\Oracle Data Mining*.

Hinweis: Der Ordner "Demos" kann über die Programmgruppe "IBM SPSS Modeler" im Windows-Startmenü aufgerufen werden.

Die Streams in der folgenden Tabelle können als Sequenz gemeinsam als Beispiel für einen Database-Mining-Prozess verwendet werden, bei dem der SVM-Algorithmus (Support Vector Machine) von Oracle Data Mining verwendet wird.

Tabelle 4. Datenbankmining - Beispielstreams

Stream	**Beschreibung**
1_upload_data.str	Bereinigt Daten und lädt sie aus einer Flatfile in die Datenbank.
2_explore_data.str	Bietet ein Beispiel für die Datenuntersuchung mit IBM SPSS Modeler.
3_build_model.str	Erstellt das Modell unter Verwendung des datenbankeigenen Algorithmus.
4_evaluate_model.str	Wird als Beispiel für die Modellevaluierung mit IBM SPSS Modeler verwendet.
5_deploy_model.str	Verwendet das Modell für datenbankinternes Scoring.

Hinweis: Um das Beispiel auszuführen, müssen die Streams in der richtigen Reihenfolge ausgeführt werden. Außerdem müssen die Quellen- und Modellierungsknoten in den einzelnen Streams aktualisiert werden, um auf eine gültige Datenquelle für die zu verwendende Datenbank zu verweisen.

Das in den Beispielstreams verwendete Dataset bezieht sich auf Kreditkartenanwendungen und stellt ein Klassifizierungsproblem mit einer Mischung aus kategorialen und stetigen Prädiktoren dar. Weitere Informationen zu diesem Dataset finden Sie in der Datei *crx.names* in demselben Ordner wie die Beispielstreams.

Diese Daten stehen im UCI Machine Learning Repository unter *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* zur Verfügung.

Beispielstream: Hochladen von Daten

Der erste Beispielstream, *1_upload_data.str*, wird verwendet, um Daten aus einer Flatfile zu bereinigen und in Oracle zu laden.

Da für das Data-Mining mit Oracle ein eindeutiges ID-Feld erforderlich ist, fügt dieser erste Stream mithilfe eines Ableitungsknotens im Dataset ein neues Feld mit dem Namen *ID* und den eindeutigen Werten "1,2,3" hinzu (unter Verwendung der IBM SPSS Modeler-Funktion `@INDEX`).

Der Füllerknoten ist für die Behandlung von fehlenden Werten zuständig und ersetzt leere, aus der Textdatei *crx.data* eingelesene Felder durch *NULL*-Werte.

Beispielstream: Datenexploration

Der zweite Beispielstream, *2_explore_data.str*, soll zeigen, wie mithilfe eines Data Audit-Knotens ein allgemeiner Überblick über die Daten (einschließlich statistischer Funktionen und Diagramme) gewonnen werden kann.

Wenn Sie im Data Audit-Bericht auf ein Diagramm doppelklicken, wird ein detaillierteres Diagramm angezeigt, in dem Sie einzelne Felder eingehender untersuchen können.

Beispielstream: Erstellen des Modells

Der dritte Beispielstream, *3_build_model.str*, veranschaulicht die Modellerstellung in IBM SPSS Modeler. Doppelklicken Sie auf den Datenbankquellenknoten (mit der Beschriftung CREDIT), um die Datenquelle anzugeben. Um die Einstellungen für die Erstellung festzulegen, doppelklicken Sie auf den Erstellungsknoten (ursprünglich mit der Beschriftung "CLASS", die sich in "FIELD16" ändert, wenn die Datenquelle angegeben wurde).

Gehen Sie auf der Registerkarte "Modell" des Dialogfelds wie folgt vor:

1. Stellen Sie sicher, dass **ID** als eindeutiges Feld ausgewählt ist.
2. Stellen Sie sicher, dass als Kernfunktion **Linear** und als Normalisierungsmethode **Z-Score** ausgewählt ist.

Beispielstream: Auswerten des Modells

Der vierte Beispielstream, *4_evaluate_model.str*, veranschaulicht die Vorteile der Verwendung von IBM SPSS Modeler für die Modellierung innerhalb der Datenbank. Sobald Sie das Modell ausgeführt haben, können Sie es wieder zu Ihrem Datenstream hinzufügen und das Modell mit verschiedenen von IBM SPSS Modeler bereitgestellten Tools evaluieren.

Anzeigen der Modellierungsergebnisse

Gliedern Sie einen Tabellenknoten an das Modellnugget an, um die Ergebnisse zu untersuchen. Das Feld **$O-field16** zeigt den vorhergesagten Wert für *field16* für jeden Fall, und das Feld **$OC-field16** zeigt den Konfidenzwert für diese Vorhersage.

Evaluieren der Modellergebnisse

Sie können den Analyseknoten verwenden, um eine Fehlklassifizierungstabelle zu erstellen, aus der das Muster der Übereinstimmungen zwischen jedem vorhergesagten Feld und dem zugehörigen Zielfeld ersichtlich wird. Führen Sie den Analyseknoten aus, um die Ergebnisse anzuzeigen.

Sie können mithilfe des Evaluierungsknotens ein Gewinndiagramm erstellen, das die Verbesserungen der Vorhersagegenauigkeit durch das Modell aufzeigt. Führen Sie den Evaluierungsknoten aus, um die Ergebnisse anzuzeigen.

Beispielstream: Bereitstellen des Modells

Sobald Sie mit der Genauigkeit des Modells zufrieden sind, können Sie es für die Verwendung mit externen Anwendungen oder für eine erneute Veröffentlichung in der Datenbank bereitstellen. Im letzten Beispielstream, *5_deploy_model.str*, werden Daten aus der Tabelle `CREDITDATA` gelesen und dann mit dem Publisher-Knoten *deploy solution* gescort und in der Tabelle `CREDITSCORES` veröffentlicht.

Kapitel 5. Datenbankmodellierung mit IBM Data Warehouse und IBM Netezza Analytics

SPSS Modeler mit IBM Data Warehouse und IBM Netezza Analytics

IBM SPSS Modeler unterstützt die Integration in die Anwendungen IBM Data Warehouse und IBM Netezza Analytics, mit denen erweiterte Analyseprozesse auf IBM Servern ausgeführt werden können. Der Zugriff auf diese Funktionen erfolgt über die grafische Benutzerschnittstelle und die am Workflow orientierte Entwicklungsumgebung von IBM SPSS Modeler. So können Sie die Data-Mining-Algorithmen direkt in der IBM Netezza- oder IBM Data Warehouse-Umgebung ausführen.

SPSS Modeler unterstützt die Integration der folgenden Algorithmen von **IBM Netezza Analytics**:

- Entscheidungsbäume
- K-Means
- Two Step
- Bayes-Netz
- Naive Bayes
- KNN
- Divisives Clustering
- PCA
- Regressionsbaum
- Lineare Regression
- Zeitreihen
- Verallgemeinert linear

Weitere Informationen zu diesen Algorithmen finden Sie im *IBM Netezza Analytics Entwicklerhandbuch* und im *Referenzhandbuch zu IBM Netezza Analytics*.

SPSS Modeler unterstützt die Integration der folgenden Algorithmen von **IBM Data Warehouse** (Bayes-Netz, divisives Clustering und Zeitreihen werden nicht unterstützt):

- Entscheidungsbäume
- K-Means
- TwoStep
- Naive Bayes
- KNN
- PCA
- Regressionsbaum
- Lineare Regression
- Verallgemeinert linear

Anmerkung: AIX wird nicht unterstützt.

Integrationsanforderungen

Für die Modellierung innerhalb der Datenbank mit IBM Netezza Analytics oder IBM Data Warehouse gelten die folgenden Voraussetzungen. Wenden Sie sich gegebenenfalls an Ihren Datenbankadministrator, um sicherzustellen, dass diese Bedingungen erfüllt sind.

- IBM SPSS Modeler mit einer Installation von IBM SPSS Modeler Server unter Windows oder UNIX (ausgenommen zLinux, für das keine IBM Netezza-ODBC-Treiber zur Verfügung stehen).
- IBM Netezza Performance Server mit aktivem IBM Netezza Analytics-Paket.

  Hinweis: Die minimale Versionsnummer von Netezza Performance Server (NPS), die erforderlich ist, hängt von der erforderlichen INZA-Version ab und lautet wie folgt:

  – Alle Versionen nach NPS 6.0.0 P8 unterstützen INZA-Versionen vor 2.0.

  – Für die Verwendung von INZA 2.0 oder höher ist NPS 6.0.5 P5 oder höher erforderlich.

  Netezza Generalized Linear und Netezza Time Series benötigen zum Funktionieren INZA 2.0 und höher. Für alle anderen datenbankinternen Netezza-Knoten ist INZA 1.1 oder höher erforderlich.
- Eine ODBC-Datenquelle zum Herstellen einer Verbindung mit einer IBM Netezza-Datenbank. Weitere Informationen finden Sie im Thema „Aktivieren der Integration” auf Seite 52.
- Eine ODBC-Datenquelle zum Herstellen einer Verbindung mit einer IBM Data Warehouse-Datenbank.
- SQL-Generierung und -Optimierung aktiviert in IBM SPSS Modeler. Weitere Informationen finden Sie im Thema „Aktivieren der Integration” auf Seite 52.

Anmerkung: Datenbankmodellierung und SQL-Optimierung erfordern, dass auf dem IBM SPSS Modeler-Computer IBM SPSS Modeler Server-Konnektivität aktiviert ist. Wenn diese Einstellung aktiviert ist, können Sie auf Datenbankalgorithmen zugreifen, SQL direkt aus IBM SPSS Modeler per Pushback übertragen und auf IBM SPSS Modeler Server zugreifen. Wählen Sie zur Überprüfung des aktuellen Lizenzstatus Folgendes im Menü von IBM SPSS Modeler aus.

Hilfe > **Info...** > **Weitere Details**

Wenn Konnektivität aktiviert ist, wird auf der Registerkarte "Lizenzstatus" die Option **Serveraktivierung** angezeigt.

Aktivieren der Integration

Das Aktivieren der Integration in IBM Netezza Analytics oder IBM Data Warehouse besteht aus folgenden Schritten.

- Konfigurieren von IBM Netezza Analytics oder IBM Data Warehouse
- Erstellen einer ODBC-Datenquelle
- Aktivieren der Integration in IBM SPSS Modeler
- Aktivieren der SQL-Generierung und -Optimierung in IBM SPSS Modeler

Diese werden in den folgenden Abschnitten beschrieben.

Konfigurieren von IBM Netezza Analytics oder IBM Data Warehouse

Informationen zum Installieren und Konfigurieren von IBM Netezza Analytics oder IBM Data Warehouse finden Sie in der entsprechenden IBM Dokumentation. Beispielsweise finden Sie Informationen zu IBM Netezza Analytics im mit dem Produkt ausgelieferten *IBM Netezza Analytics Installationshandbuch*. Der Abschnitt zur *Einrichtung von Datenbankberechtigungen* in diesem Handbuch beinhaltet Details zu Scripts, die ausgeführt werden müssen, damit IBM SPSS Modeler-Streams in die Datenbank schreiben können.

Anmerkung: Wenn Sie vorhaben, Knoten zu verwenden, die auf einer Matrixberechnung beruhen, muss die Matrixengine durch Ausführung von `CALL NZM..INITIALIZE();` initialisiert werden. Andernfalls schlägt die Ausführung gespeicherter Prozeduren fehl. Die Initialisierung ist ein Setup-Schritt für jede Datenbank, der nur ein einziges Mal ausgeführt werden muss.

Erstellen einer ODBC-Datenquelle für IBM Netezza Analytics

Um die Verbindung zwischen der IBM Netezza-Datenbank und IBM SPSS Modeler zu aktivieren, müssen Sie einen ODBC-System-DSN (Data Source Name, Datenquellennamen) erstellen.

Bevor Sie einen DSN erstellen, sollten Sie grundlegende Kenntnisse über ODBC-Datenquellen und -Treiber sowie über Datenbankunterstützung in IBM SPSS Modeler besitzen.

Wenn Sie mit IBM SPSS Modeler Server im verteilten Modus arbeiten, müssen Sie den DSN auf dem Server-Computer erstellen. Wenn Sie im lokalen (Client-)Modus arbeiten, müssen Sie auf dem Client-Computer einen DSN erstellen.

Windows-Clients

1. Führen Sie von Ihrer *Netezza-Client*-CD aus die Datei *nzodbcsetup.exe* aus, um das Installationsprogramm zu starten. Befolgen Sie die Anweisungen auf dem Bildschirm, um den Treiber zu installieren. Vollständige Anweisungen finden Sie im Installations- und Konfigurationshandbuch zu IBM Netezza ODBC, JDBC und OLE DB.
   a. Erstellen Sie den DSN.

      Anmerkung: Die Menüfolge ist von der jeweils vorliegenden Windows-Version abhängig.

      - **Windows XP.** Wählen Sie im Menü "Start" die Option **Systemsteuerung** aus. Doppelklicken Sie auf **Verwaltung** und dann auf **Datenquellen (ODBC)**.
      - **Windows Vista.** Wählen Sie im Menü "Start" die Option **Systemsteuerung** und dann **Systemwartung** aus. Doppelklicken Sie auf **Verwaltung**, wählen Sie dann **Datenquellen (ODBC)** aus und klicken Sie auf **Öffnen**.
      - **Windows 7.** Wählen Sie im Menü "Start" die Option **Systemsteuerung**, dann **System & Sicherheit** und anschließend **Verwaltung** aus. Wählen Sie **Datenquellen (ODBC)** aus und klicken Sie dann auf **Öffnen**.

   b. Wechseln Sie zur Registerkarte **System-DSN** und klicken Sie auf **Hinzufügen**.
2. Wählen Sie **NetezzaSQL** aus der Liste aus und klicken Sie auf **Fertigstellen**.
3. Geben Sie auf der Registerkarte **DSN-Optionen** der Anzeige für die Netezza-ODBC-Treibereinrichtung einen Datenquellennamen Ihrer Wahl, den Hostnamen oder die IP-Adresse des IBM Netezza-Servers, die Portnummer für die Verbindung, die Datenbank der verwendeten IBM Netezza-Instanz sowie den Benutzernamen und das Kennwort für die Datenbankverbindung ein. Klicken Sie auf die Schaltfläche **Hilfe**, wenn Sie eine Erläuterung der Felder wünschen.
4. Klicken Sie auf die Schaltfläche **Verbindung testen** und stellen Sie sicher, dass Sie eine Verbindung zur Datenbank herstellen können.
5. Klicken Sie mehrere Male auf **OK**, wenn Sie eine Verbindung hergestellt haben, um die Anzeige "ODBC-Datenquellen-Administrator" zu schließen.

Windows-Server

Die Prozedur für Windows-Server ist bei Windows XP mit der Client-Prozedur identisch.

UNIX- bzw. Linux-Server

Die folgende Prozedur gilt für UNIX- bzw. Linux-Server (mit Ausnahme von zLinux, wofür keine IBM Netezza ODBC-Treiber verfügbar sind).

1. Kopieren Sie von Ihrer Netezza Client-CD/DVD die betreffende Datei `<platform>cli.package.tar.gz` in ein temporäres Verzeichnis auf dem Server.
2. Extrahieren Sie den Archivinhalt mittels der Befehle **`gunzip`** und **`untar`**.
3. Fügen Sie Ausführungsberechtigungen für das extrahierte *unpack*-Script hinzu.
4. Führen Sie das Script aus und bearbeiten Sie die Eingabeaufforderungen auf dem Bildschirm.
5. Bearbeiten Sie die Datei `modelersrv.sh` so, dass sie folgende Zeilen enthält:

```
. <SDAP-Installationspfad>/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=<SDAP-Installationspfad>; export NZ_ODBC_INI_PATH
```

Beispiel:

```
. /usr/IBM/SPSS/SDAP/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP; export NZ_ODBC_INI_PATH
```

6. Suchen Sie die Datei `/usr/local/nz/lib64/odbc.ini` und kopieren Sie ihren Inhalt in die Datei `odbc.ini`, die zusammen mit SDAP installiert wird (die Datei, die durch die Umgebungsvariable $ODBCINI definiert wird).

   Hinweis: Bei 64-Bit-Linux-Systemen verweist der Parameter **`Driver`** (Treiber) fälschlicherweise auf den 32-Bit-Treiber. Bearbeiten Sie beim Kopieren der `odbc.ini`-Inhalte im vorangegangenen Schritt den Pfad in diesem Parameter entsprechend, z. B.:

   ```
   /usr/local/nz/lib64/libnzodbc.so
   ```

7. Bearbeiten Sie die Parameter in der Netezza DSN-Definition so, dass die zu verwendende Datenbank angegeben wird.
8. Starten Sie IBM SPSS Modeler Server neu und testen Sie die Verwendung der Netezza-Knoten zum datenbankinternen Mining auf dem Client.

Aktivieren der Integration in SPSS Modeler

1. Wählen Sie im IBM SPSS Modeler-Hauptmenü Folgendes aus:

   Tools > **Optionen** > **Hilfsanwendungen**
2. Klicken Sie auf die Registerkarte **IBM Data Warehouse**.

IBM Data Warehouse Analytics-Integration aktivieren. Aktiviert die Datenbankmodellierungspalette (sofern nicht bereits angezeigt) am unteren Rand des IBM SPSS Modeler-Fensters und fügt die Knoten für die Algorithmen von IBM Data Warehouse und Netezza Data Mining hinzu.

IBM Data Warehouse-Verbindung. Klicken Sie auf die Schaltfläche **Bearbeiten** und wählen Sie die IBM Data Warehouse-Verbindungszeichenfolge aus, die Sie bei der Erstellung der ODBC-Quelle eingerichtet haben. Weitere Informationen finden Sie in der IBM Data Warehouse-Administrationskonsole.

Aktivieren der SQL-Generierung und -Optimierung

Da mit hoher Wahrscheinlichkeit mit großen Datasets gearbeitet wird, sollten Sie aus Leistungsgründen die IBM SPSS Modeler-Optionen zur SQL-Generierung und -Optimierung aktivieren.

1. Wählen Sie in den IBM SPSS Modeler-Menüs Folgendes aus:

   Tools > **Streameigenschaften** > **Optionen**
2. Klicken Sie im Navigationsbereich auf die Option **Optimierung**.
3. Überzeugen Sie sich, dass die Option **SQL generieren** aktiviert ist. Diese Einstellung ist für die Datenbankmodellierung erforderlich.
4. Wählen Sie **SQL-Generierung optimieren** und **Andere Ausführung optimieren** aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Erstellen von Modellen mit IBM Netezza Analytics und IBM Data Warehouse

Zu jedem der unterstützten Algorithmen gehört ein Modellierungsknoten. Über die Registerkarte **Datenbankmodellierung** in der Knotenpalette können Sie auf die IBM Data Warehouse- und IBM Netezza-Modellierungsknoten zugreifen.

Erläuterung der Daten

Felder in der Datenquelle können, je nach Modellierungsknoten, Variablen verschiedener Datentypen enthalten. In IBM SPSS Modeler werden Datentypen als *Messniveaus* bezeichnet. Auf der Registerkarte "Felder" des Modellierungsknotens werden Symbole verwendet, die die zulässigen Messniveautypen für die Eingabe- und Zielfelder angeben.

Zielfeld Das Zielfeld ist das Feld, dessen Wert Sie vorherzusagen versuchen. Wenn ein Ziel angegeben werden kann, kann nur eines der Quellendatenfelder als Zielfeld ausgewählt werden.

Feld für Datensatz-ID Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. Wenn die Quellendaten kein ID-Feld enthalten, können Sie dieses Feld mithilfe eines Ableitungsknotens erstellen, wie in der folgenden Prozedur gezeigt.

1. Wählen Sie den Quellenknoten aus.
2. Doppelklicken Sie auf der Registerkarte "Feldoperationen" auf den Ableitungsknoten.
3. Öffnen Sie den Ableitungsknoten, indem Sie im Erstellungsbereich auf das zugehörige Symbol doppelklicken.
4. Geben Sie im **Ableitungsfeld** (beispielsweise) `ID` ein.
5. Geben Sie `@INDEX` im Feld **Formel** ein und klicken Sie auf **OK**.
6. Verbinden Sie den Ableitungsknoten mit dem Rest des Streams.

Anmerkung: Wenn Sie aus einer Netezza-Datenbank lange numerische Daten unter Verwendung des Datentyps `NUMERIC(18,0)` abrufen, rundet SPSS Modeler die Daten gelegentlich während des Imports auf. Zur Vermeidung dieses Problems speichern Sie die Daten unter Verwendung eines der Datentypen `BIGINT` oder `NUMERIC(36,0)`.

Anmerkung: Aufgrund der Einschränkungen für die Feldtypen, die verwendet werden können, wird ein Feld mit einem Messniveau ohne Typ und der Rolle **Datensatz-ID** nicht in einem datenbankinternen Netezza-Modellierungsknoten (beispielsweise K-Means) angezeigt.

Umgang mit Nullwerten

Wenn die Eingabedaten Nullwerte enthalten, kann die Verwendung einiger Netezza-Knoten zu Fehlernachrichten oder Streams mit sehr langer Ausführungsdauer führen. Deshalb empfehlen wir, Datensätze mit Nullwerten zu entfernen. Verwenden Sie die folgende Methode.

1. Verbinden Sie einen Auswahlknoten mit dem Quellenknoten.
2. Setzen Sie die Option **Modus** des Auswahlknotens auf **Verwerfen**.
3. Geben Sie Folgendes in das Feld **Bedingung** ein:

   ```
   @NULL(Feld1) [oder @NULL(Feld2)[... oder @NULL(FeldN]])
   ```

   Achten Sie darauf, alle Eingabefelder mit aufzunehmen.
4. Verbinden Sie den Auswahlknoten mit dem Rest des Streams.

Modellausgabe

Es ist möglich, dass ein Stream, der einen Data Warehouse- oder Netezza-Modellierungsknoten enthält, bei jeder Ausführung etwas andere Ergebnisse ausgibt. Der Grund hierfür ist, dass die Reihenfolge, in der der Knoten die Quellendaten liest, nicht immer gleich ist, da die Daten vor der Modellerstellung in temporäre Tabellen eingelesen werden. Die durch diesen Effekt erzeugten Unterschiede sind jedoch vernachlässigbar.

Allgemeine Kommentare

- In IBM SPSS Collaboration and Deployment Services können keine Scoring-Konfigurationen mithilfe von Streams erstellt werden, die IBM Data Warehouse- oder IBM Netezza-Datenbankmodellierungsknoten enthalten.

- PMML-Export bzw. -Import ist für Modelle, die von den Data Warehouse- oder Netezza-Knoten erstellt wurden, nicht möglich.

Feldoptionen

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

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

Ziel. Wählen Sie ein Feld als Ziel für die Vorhersage aus. Beachten Sie bei verallgemeinerten linearen Modellen auch das Feld **Tests** in dieser Anzeige.

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

Serveroptionen

Auf der Registerkarte "Server" geben Sie die IBM Data Warehouse-Datenbank an, in der das Modell erstellt werden soll.

IBM Data Warehouse-Serverdetails. Hier geben Sie die Verbindungsdetails für die für das Modell zu verwendende Datenbank an.

- **Vorgeordnete Verbindung verwenden.** (Standardeinstellung) Verwendet die Verbindungsdetails, die in einem vorgeordneten Knoten, beispielsweise dem Datenbankquellenknoten, angegeben sind. Diese Option funktioniert nur, wenn alle vorgeordneten Knoten SQL-Pushback verwenden können. In diesem Fall müssen die Daten nicht aus der Datenbank verschoben werden, da die SQL alle vorgeordneten Knoten vollständig implementiert.
- **Daten in Verbindung verschieben**. Dient zum Verschieben der Daten in die hier angegebene Datenbank. Dadurch kann die Modellierung funktionieren, wenn sich die Daten in einer anderen IBM Data Warehouse-Datenbank, einer Datenbank eines anderen Anbieters oder in einer Flatfile befinden. Darüber hinaus werden die Daten in die hier angegebene Datenbank zurückverschoben, wenn die Daten extrahiert wurden, da ein Knoten kein SQL-Pushback durchgeführt hat. Klicken Sie auf die Schaltfläche **Bearbeiten**, um eine Verbindung zu suchen und auszuwählen.

  

  Vorsicht: IBM Netezza Analytics und IBM Data Warehouse werden in der Regel mit sehr großen Datasets verwendet. Das Übertragen großer Datenmengen zwischen Datenbanken bzw. aus einer Datenbank und wieder zurück kann sehr zeitaufwendig sein und sollte nach Möglichkeit vermieden werden.

Anmerkung: Der Name der ODBC-Datenquelle wird in jeden IBM SPSS Modeler-Stream eingebettet. Wenn ein auf einem Host erstellter Stream auf einem anderen Host ausgeführt wird, muss der Name der Datenquelle auf beiden Hosts identisch sein. Alternativ kann für jeden Quellen- oder Modellierungsknoten auf der Registerkarte "Server" eine andere Datenquelle ausgewählt werden.

Modelloptionen

Auf der Registerkarte "Modelloptionen" können Sie bestimmen, ob Sie einen Namen für das Modell angeben oder automatisch erstellen lassen möchten. Sie können auch Standardwerte für Scoring-Optionen festlegen.

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

Vorhandenes ersetzen, wenn Name bereits verwendet wird. Wenn Sie dieses Kontrollkästchen auswählen, wird ein vorhandenes Modell mit demselben Namen ggf. überschrieben.

Für Scoring bereitstellen. Sie können hier die Standardwerte für die Scoring-Optionen festlegen, die im Dialogfeld für das Modellnugget angezeigt werden. Details zu den Optionen finden Sie im Hilfethema für die Registerkarte "Einstellungen" des betreffenden Nuggets.

Verwalten von Modellen

Bei der Erstellung eines IBM Netezza- oder IBM Data Warehouse-Modells mit SPSS Modeler wird einerseits in SPSS Modeler ein Modell erstellt und andererseits in der IBM Data Warehouse-Datenbank ein Modell erstellt oder ersetzt. Das SPSS Modeler-Modell verweist auf den Inhalt eines in einem Datenbankserver gespeicherten Datenbankmodells. SPSS Modeler kann eine Konsistenzprüfung durchführen, indem sowohl im SPSS Modeler-Modell als auch im Netezza- oder Data Warehouse-Modell eine identische, generierte Modellschlüsselzeichenfolge gespeichert wird.

Der Modellname wird für jedes Netezza- oder Data Warehouse-Modell im Dialogfeld für die Auflistung von Datenbankmodellen in der Spalte *Modellinformationen* angezeigt. Der Modellname eines SPSS Modeler-Modells wird auf der Registerkarte "Server" eines SPSS Modeler-Modells (wenn es sich in einem Stream befindet) als der Modellschlüssel ausgegeben.

Mit der Schaltfläche "Überprüfen" wird ermittelt, ob die Modellschlüssel des SPSS Modeler-Modells und des Netezza- oder Data Warehouse-Modells übereinstimmen. Wenn in Netezza oder Data Warehouse kein Modell mit demselben Namen gefunden wird oder wenn die Modellschlüssel nicht übereinstimmen, bedeutet dies, dass das Netezza- oder Data Warehouse-Modell seit der Erstellung des SPSS Modeler-Modells gelöscht oder neu erstellt wurde.

Auflisten der Datenbankmodelle

SPSS Modeler stellt ein Dialogfeld bereit, über das die in IBM Data Warehouse gespeicherten Modelle aufgelistet und gelöscht werden können. Der Zugriff auf dieses Dialogfeld erfolgt über das Dialogfeld für IBM Hilfsanwendungen und über die Dialogfelder zum Erstellen, Suchen und Anwenden für die mit dem Datenmining von IBM Data Warehouse und IBM Netezza verbundenen Knoten. Zu jedem Modell werden folgende Informationen angezeigt:

- Modellname (Name des Modells - zum Sortieren der Liste verwendet).
- Eignername.
- Der im Modell verwendete Algorithmus.
- Der aktuelle Status des Modells, z. B. Abgeschlossen.
- Das Datum, an dem das Modell erstellt wurde.

IBM Data WH-Regressionsbaum

Ein Regressionsbaum ist ein baumbasierter Algorithmus, der eine Stichprobe von Fällen wiederholt aufteilt, um anhand von Werten eines numerischen Ausgabefeldes gleichartige Subsets abzuleiten. Ebenso wie Entscheidungsbäume zerlegen auch Regressionsbäume die Daten in Subsets, in denen die Blätter des Baums hinreichend kleinen bzw. hinreichend einheitlichen Subsets entsprechen. Aufteilungen werden ausgewählt, um die Streuung der Zielattributwerte zu verringern, sodass sie angemessen gut durch ihren Mittelwert an Blättern vorhergesagt werden können.

Erstellungsoptionen für IBM Data WH-Regressionsbaum - Baumerweiterung

Sie können Erstellungsoptionen für die Baumerweiterung und die Baumreduzierung festlegen.

Die folgenden Erstellungsoptionen sind für den Baumaufbau verfügbar:

Maximale Baumtiefe. Die maximale Anzahl der Ebenen, auf die ein Baum unterhalb des Stammknotens erweitert werden kann, d. h. die Anzahl der rekursiven Teilungen einer Stichprobe. Der Standardwert liegt bei 62. Dies ist die maximale Baumtiefe für Modellierungszwecke.

Anmerkung: Wenn der Viewer im Modellnugget das Modell in Texform darstellt, werden maximal 12 Ebenen des Baums angezeigt.

Aufteilungskriterien. Diese Optionen steuern, wann die Aufteilung des Baums aufhört. Wenn Sie nicht die Standardwerte verwenden möchten, klicken Sie auf **Anpassen** und ändern die Werte.

- **Maß zur Aufteilungsevaluierung.** Dieses Evaluierungsmaß für die Klasse ermittelt die beste Position für eine Baumteilung.

  Anmerkung: Derzeit ist "Varianz" die einzig mögliche Option.
- **Mindestverbesserung für Aufteilungen.** Der Mindestwert der Unreinheitsreduzierung, bevor eine neue Aufteilung des Baums erfolgt. Das Ziel der Baumerstellung besteht darin, Untergruppen mit ähnlichen Ausgabewerten zu erstellen, um die Unreinheit in den einzelnen Knoten zu minimieren. Wenn die beste für eine Verzweigung mögliche Aufteilung die Unreinheit um weniger als den durch die Aufteilungskriterien vorgegebenen Betrag reduziert, wird die Verzweigung nicht aufgeteilt.
- **Mindestanzahl der Instanzen für eine Aufteilung.** Die Mindestanzahl von Datensätzen, die aufgeteilt werden können. Wenn weniger nicht aufgeteilte Datensätze verbleiben, werden keine weiteren Aufteilungen durchgeführt. Sie können dieses Feld verwenden, um die Erstellung kleiner Untergruppen im Baum zu verhindern.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Wählen Sie eine der folgenden Optionen aus:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.
- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

Erstellungsoptionen für IBM Data WH-Regressionsbaum - Baumreduzierung

Sie können die Reduzierungsoptionen verwenden, um Reduzierungskriterien für den Regressionsbaum festzulegen. Ziel der Reduzierung ist es, das Risiko der übermäßigen Anpassung zu verringern, indem zu stark erweiterte Untergruppen entfernt werden, welche die erwartete Genauigkeit für neue Daten nicht verbessern.

Reduzierungsmaß. Das Reduzierungsmaß gewährleistet, dass die geschätzte Genauigkeit des Modells nach der Entfernung eines Blatts aus dem Baum innerhalb akzeptabler Grenzen bleibt. Sie können eines der folgenden Maße auswählen.

- **mse.** Mittlerer quadratischer Fehler - (Standard) misst, wie eng eine angepasste Linie an den Datenpunkten liegt.
- **r2.** R-Quadrat - misst den Anteil an Variation in der abhängigen Variablen, der durch das Regressionsmodell erklärt wird.
- **Pearson.** Korrelationskoeffizienten nach Pearson - misst die Stärke der Beziehung zwischen linear abhängigen Variablen, die normal verteilt sind.
- **Spearman.** Korrelationskoeffizient nach Spearman - erkennt nicht lineare Beziehungen, die laut der Korrelation nach Pearson schwach erscheinen, jedoch möglicherweise tatsächlich stark sind.

Daten für die Reduzierung. Sie können einen Teil oder alle Trainingsdaten verwenden, um die erwartete Genauigkeit der neuen Daten abzuschätzen. Alternativ können Sie zu diesem Zweck ein separates Dataset für die Reduzierung aus einer festgelegten Tabelle verwenden.

- **Alle Trainingsdaten verwenden.** Diese (standardmäßige) Option verwendet alle Trainingsdaten, um die Modellgenauigkeit zu schätzen.
- **% der Trainingsdaten für die Reduzierung verwenden.** Teilen Sie mithilfe dieser Option die Daten in zwei Gruppen (eine für das Training und eine für die Reduzierung) unter Verwendung des hier angegebenen Prozentsatzes für die Reduzierungsdaten.

  Wählen Sie das Feld **Ergebnisse replizieren** aus, wenn Sie einen Zufallsstartwert angeben möchten, um sicherzustellen, dass die Daten bei jeder Ausführung des Streams auf dieselbe Weise partitioniert werden. Sie können entweder eine ganze Zahl im Feld **Für Reduzierung verwendeter Startwert** angeben oder auf **Generieren** klicken, wodurch eine pseudozufällige ganze Zahl erstellt wird.
- **Daten aus einer vorhandenen Tabelle verwenden.** Geben Sie den Tabellennamen eines separaten Datasets für die Reduzierung an, anhand dessen die Modellgenauigkeit geschätzt wird. Diese Vorgehensweise wird als zuverlässiger betrachtet als die Nutzung von Trainingsdaten. Wenn Sie diese Option wählen, wird jedoch eventuell ein großes Subset von Daten aus dem Trainingsset entfernt, wodurch die Qualität des Entscheidungsbaum beeinträchtigt wird.

Netezza - Divisives Clustering

Divisives Clustering ist eine Methode der Clusteranalyse, bei der der Algorithmus wiederholt ausgeführt wird, um Cluster in Subcluster aufzuteilen, bis ein angegebener Stopppunkt erreicht wird.

Die Clusterbildung beginnt mit einem einzelnen Cluster, der sämtliche Trainingsinstanzen (Datensätze) enthält. Bei der ersten Iteration des Algorithmus wird das Dataset in zwei Subcluster aufgeteilt, die durch die nachfolgenden Iterationen in weitere Subcluster aufgespaltet werden. Die Stoppkriterien werden angegeben als maximale Anzahl an Iterationen, als maximale Anzahl der Ebenen, in die das Dataset unterteilt wird, und als erforderliche Mindestanzahl an Instanzen für die weitere Partitionierung.

Der sich so ergebende hierarchische Clustering-Baum kann verwendet werden, um Instanzen zu klassifizieren, indem diese aus dem Stammcluster nach unten weitergegeben werden, wie im folgenden Beispiel.

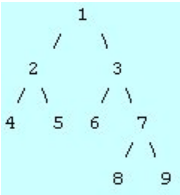

Abbildung 3. Beispiel für einen divisiven Clustering-Baum

Auf jeder Ebene wird der Subcluster mit der besten Übereinstimmung hinsichtlich des Abstands der Instanz von den Subclusterzentren ausgewählt.

Wenn die Instanzen mit einer angewendeten Hierarchieebene von -1 (Standard) gescort werden, gibt das Scoring lediglich einen Blattcluster zurück, da Blätter durch negative Nummern gekennzeichnet sind. In diesem Beispiel wäre dies einer der Cluster 4, 5, 6, 8 oder 9. Wenn jedoch die Hierarchieebene beispielsweise auf 2 gesetzt ist, gibt das Scoring einen der Cluster auf der zweiten Ebene unterhalb des Stammclusters aus, also 4, 5, 6 oder 7.

Feldoptionen für "Netezza - Divisives Clustering"

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

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

Erstellungsoptionen für "Netezza - Divisives Clustering"

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

Distanzmaß. Die zur Messung des Abstands zwischen Datenpunkten zu verwendende Methode. Größere Abstände deuten auf größere Unähnlichkeiten hin. Folgende Optionen stehen zur Auswahl:

- **Euklidisch**. (Standardvorgabe) Der Abstand zwischen zwei Punkten wird anhand einer geradlinigen Verbindung zwischen den Punkten berechnet.
- **Manhattan**. Der Abstand zwischen zwei Punkten wird als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet.
- **Canberra**. Ähnlich wie die Manhattan-Distanz, jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.
- **Maximum**. Der Abstand zwischen zwei Punkten wird als der größte Unterschied in einer beliebigen Koordinatendimension berechnet.

Maximale Anzahl an Iterationen. Bei diesem Algorithmus werden mehrere Iterationen desselben Vorgangs durchgeführt. Mit dieser Option können Sie das Modelltraining nach der angegebenen Anzahl von Iterationen beenden.

Maximale Tiefe der Clusterbäume. Die maximale Anzahl an Ebenen, in die das Dataset unterteilt werden kann.

Ergebnisse replizieren. Aktivieren Sie dieses Kontrollkästchen, wenn Sie einen Zufallsstartwert festlegen möchten, mit dem Sie Analysen reproduzieren können. Sie können entweder eine ganze Zahl angeben oder auf **Generieren** klicken, wodurch eine pseudozufällige ganze Zahl erstellt wird.

Mindestanzahl der Instanzen für eine Aufteilung. Die Mindestanzahl von Datensätzen, die aufgeteilt werden können. Wenn weniger nicht aufgeteilte Datensätze verbleiben, werden keine weiteren Aufteilungen durchgeführt. Sie können dieses Feld verwenden, um die Erstellung sehr kleiner Untergruppen im Clusterbaum zu verhindern.

IBM Data WH - Verallgemeinert linear

Die lineare Regression ist ein etabliertes statistisches Verfahren zur Klassifizierung von Datensätzen auf der Grundlage der Werte von numerischen Eingabefeldern. Die lineare Regression entspricht einer geraden Linie oder Fläche, die die Diskrepanzen zwischen den vorhergesagten und den tatsächlichen Werten minimiert. Lineare Modelle sind aufgrund ihrer Einfachheit sowohl beim Training als auch bei der Modellanwendung nützlich für die Modellierung einer breiten Palette von Phänomenen aus der Praxis. Lineare Modelle setzen jedoch eine Normalverteilung in der abhängigen (Ziel-)Variablen und eine lineare Auswirkung der unabhängigen (Prädiktor-)Variablen auf die abhängige Variable voraus.

Es gibt viele Situationen, in denen eine lineare Regression nützlich ist, in denen die oben stehenden Annahmen jedoch nicht gelten. Beispielsweise weist die abhängige Variable bei der Modellierung der Verbraucherentscheidung zwischen einer diskreten Anzahl an Produkten vermutlich eine Multinomialvertei-

lung auf. Und bei der Modellierung des Einkommens in Abhängigkeit vom Alter steigt das Einkommen zwar typischerweise mit zunehmendem Alter, die Verknüpfung zwischen den beiden Elementen ist jedoch kaum so einfach, als dass sie durch eine Gerade ausgedrückt werden könnte.

In diesen Situationen kann ein verallgemeinertes lineares Modell verwendet werden. Verallgemeinerte lineare Modelle erweitern das lineare Regressionsmodell, sodass die abhängige Variable über eine angegebene Verknüpfungsfunktion (für die eine Reihe geeigneter Funktionen zur Auswahl steht) mit den Prädiktorvariablen in Bezug gesetzt wird. Außerdem ist es mit diesem Modell möglich, dass die abhängige Variable eine von der Normalverteilung abweichende Verteilung (beispielsweise Poisson-Verteilung) aufweist.

Der Algorithmus sucht iterativ nach dem am besten angepassten Modell, wobei die maximale Anzahl an Iterationen festgelegt wird. Bei der Berechnung der besten Anpassung wird der Fehler durch die Quadratsumme der Differenzen zwischen dem vorhergesagten und dem tatsächlichen Wert der abhängigen Variablen dargestellt.

Optionen von Feldern für verallgemeinerte lineare IBM Data WH-Modelle

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

Vordefinierte Rollen verwenden. Bei dieser Option werden die Rolleneinstellungen, z. B. Ziele oder Prädiktoren, aus einem vorgeordneten Typknoten oder die Registerkarte "Typen" eines vorgeordneten Quellenknotens verwendet.

Benutzerdefinierte Feldzuweisungen verwenden. Wählen Sie diese Option aus, wenn Sie Ziele, Prädiktoren und andere Rollen in dieser Anzeige manuell zuweisen möchten.

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

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

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird. Die Werte in diesem Feld müssen für jeden Datensatz eindeutig sein, z. B. Kundennummern.

Instanzgewichtung. Geben Sie ein Feld an, das Instanzgewichtungen verwenden soll. Eine Instanzgewichtung ist eine Gewichtung pro Zeile von Eingabedaten. Standardmäßig wird angenommen, dass alle Eingabedatensätze die gleiche relative Wichtigkeit haben. Sie können die Wichtigkeit ändern, indem Sie den Eingabedatensätzen unterschiedliche Gewichtungen zuweisen. Das Feld, das Sie angeben, muss eine numerische Gewichtung für jede Zeile der Eingabedaten enthalten.

Prädiktoren (Eingaben). Wählen Sie das Eingabefeld bzw. die Eingabefelder aus. Diese Aktion ist ähnlich, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.

Optionen für verallgemeinertes lineares IBM Data WH-Modell - Allgemein

Auf der Registerkarte "Modelloptionen" können Sie bestimmen, ob Sie einen Namen für das Modell angeben oder automatisch erstellen lassen möchten. Sie können auch verschiedene Einstellungen bezüglich des Modells der Verknüpfungsfunktion und der Eingabefeld-Interaktionen (sofern vorhanden) vornehmen und Standardwerte für Scoring-Optionen festlegen.

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

Feldoptionen. Sie können die Rollen der Eingabefelder für die Modellerstellung angeben.

Allgemeine Einstellungen. Diese Einstellungen beziehen sich auf die Stoppkriterien für den Algorithmus.

- **Maximale Anzahl an Iterationen.** Dies ist die maximale Anzahl der Iterationen, die im Algorithmus vorgenommen werden; der Minimalwert ist 1 und der Standardwert ist 20.
- **Maximaler Fehler (1e).** Der maximale Fehlerwert (in wissenschaftlicher Notation), bei dem der Algorithmus die Suche nach dem am besten angepassten Modell beenden soll. Der Minimalwert ist 0 und der Standardwert ist -3, d. h. 1E-3 bzw. 0,001.
- **Vernachlässigbarer Fehlerschwellenwert (1e).** Der Wert (in wissenschaftlicher Notation), unterhalb dessen Fehler so behandelt werden, als hätten sie den Wert 0. Der Minimalwert ist -1 und der Standardwert ist -7, es werden also Fehlerwerte unter 1E-7 (bzw. 0,0000001) als nicht signifikant gewertet.

Verteilungseinstellungen. Diese Einstellungen beziehen sich auf die Verteilung der abhängigen (Ziel-)Variablen.

- **Verteilung der Antwortvariablen.** Der Verteilungstyp. Zur Auswahl stehen **Bernoulli** (Standardwert), **Gauß**, **Poisson**, **Binomial**, **Negativ-Binomial**, **Wald** (invers normal) und **Gamma**.
- **Parameter.** (Nur Poisson- oder Binomialverteilung) Sie müssen im Feld **Parameter angeben** eine der folgenden Optionen angeben:
  - Wählen Sie **Standard** aus, damit die Parameter automatisch aufgrund der Daten geschätzt werden.
  - Wählen Sie **Quasi** aus, um die Optimierung der Verteilung-Quasi-Likelihood zu ermöglichen.
  - Wählen Sie **Explizit** aus, um den Parameterwert explizit anzugeben.

  (Nur Binomialverteilung) Sie müssen die Eingabetabellenspalte, die für das Testfeld verwendet werden soll, den Anforderungen der Binomialverteilung entsprechend angeben. Diese Spalte enthält die Anzahl der Tests für die Binomialverteilung.

  (Nur negative Binomialverteilung) Sie können den Standardwert -1 verwenden oder einen anderen Parameterwert angeben.

Einstellungen der Verknüpfungsfunktion. Diese Einstellungen beziehen sich auf die Verknüpfungsfunktion, die die abhängige Variable mit den Prädiktorvariablen in Bezug setzt.

- **Verknüpfungsfunktion.** Die zu verwendende Funktion. Zur Auswahl stehen: **Identität**, **Kehrwert**, **Kehrwert negativ**, **Kehrwert Quadrat**, **Wurzel**, **Potenz**, **Oddspower**, **Log**, **Clog**, **Loglog**, **Cloglog**, **Logit** (Standardvorgabe), **Probit**, **Gaussit**, **Cauchit**, **Canbinom**, **Cangeom**, **Cannegbinom**.
- **Parameter.** (Nur bei den Verknüpfungsfunktionen "Potenz" und "Oddspower") Sie können einen Parameterwert angeben, wenn die Verknüpfungsfunktion **Potenz** oder **Oddspower** verwendet wird. Sie können entweder einen Wert angeben oder die Standardeinstellung 1 verwenden.

Optionen für verallgemeinertes lineares IBM Data WH-Modell - Interaktionen

Der Bereich "Interaktionen" enthält die Optionen zur Angabe von Interaktionen (d. h. von multiplikativen Effekten zwischen Eingabefeldern).

Spalteninteraktion. Aktivieren Sie dieses Kontrollkästchen, um Interaktionen zwischen Eingabefeldern anzugeben. Lassen Sie dieses Feld leer, wenn keine Interaktionen bestehen.

Geben Sie Interaktionen in das Modell ein, indem Sie ein oder mehrere Felder in der Quellenliste auswählen und sie in die Interaktionsliste ziehen. Welche Art von Interaktion erstellt wird, hängt davon ab, auf welchem Hotspot Sie die Auswahl ablegen.

- **Haupt.** Die abgelegten Felder werden unten in der Interaktionsliste als separate Hauptinteraktionen angezeigt.
- **Zweiweg.** Alle möglichen Paare der abgelegten Felder werden unten in der Interaktionsliste als Zweiwegeinteraktionen angezeigt.
- **Dreiweg.** Alle möglichen Dreiergruppen der abgelegten Felder werden unten in der Interaktionsliste als Dreiwegeinteraktionen angezeigt.
- ***.** Die Kombination aller abgelegten Felder wird unten in der Interaktionsliste als Einzelinteraktion angezeigt.

Konstanten Term einschließen. Der konstante Term wird normalerweise in das Modell eingeschlossen. Wenn anzunehmen ist, dass die Daten durch den Koordinatenursprung verlaufen, können Sie den konstanten Term ausschließen.

Dialogfeldschaltflächen

Mit den Schaltflächen rechts neben der Anzeige können Sie an den im Modell verwendeten Termen Änderungen vornehmen.

Abbildung 4. Löschschaltfläche

Löschen von Termen aus dem Modell durch Auswahl der zu löschenden Terme und anschließendes Klicken auf die Schaltfläche zum Löschen.

Abbildung 5. Umordnungsschaltflächen

Umsortieren der Terme innerhalb des Modells durch Auswahl der Terme, die umsortiert werden sollen, und anschließendes Klicken auf den Aufwärts- bzw. Abwärtspfeil.

Abbildung 6. Schaltfläche für benutzerdefinierte Interaktion

Hinzufügen von benutzerdefinierten Termen

Sie können benutzerdefinierte Interaktionen im Format *n1*x1*x1*x1..* festlegen. Wählen Sie in der Liste **Felder** ein Feld aus, klicken Sie auf die Schaltfläche mit dem Rechtspfeil, um **Benutzerdefinierter Term** das Feld hinzuzufügen, und klicken Sie auf **Nach***. Wählen Sie dann das nächste Feld aus, klicken Sie auf die Schaltfläche mit dem Rechtspfeil und so weiter. Wenn Sie die benutzerdefinierte Interaktion fertiggestellt haben, klicken Sie auf **Term hinzufügen**, um ihn an den Bereich "Interaktionen" zurückzugeben.

Verallgemeinertes lineares IBM Data WH-Modell - Scoring-Optionen der Modelloptionen

Für Scoring bereitstellen. Sie können hier die Standardwerte für die Scoring-Optionen festlegen, die im Dialogfeld für das Modellnugget angezeigt werden. Weitere Informationen finden Sie im Thema „Nugget für verallgemeinertes lineares IBM Data WH-Modell - Registerkarte "Einstellungen"" auf Seite 89.

- **Eingabefelder einschließen.** Aktivieren Sie dieses Kontrollkästchen, wenn Sie die Eingabefelder in die Modellausgabe und die Vorhersage mit aufnehmen möchten.

IBM Data WH-Entscheidungsbäume

Ein Entscheidungsbaum ist eine hierarchische Struktur, die ein Klassifizierungsmodell darstellt. Mit einem Entscheidungsbaummodell können Sie ein Klassifizierungssystem entwickeln, die zukünftige Beobachtungen auf der Grundlage eines Satzes von Trainingsdaten vorhersagen oder klassifizieren. Die Klassifizierung hat die Form einer Baumstruktur, in der die Verzweigungen Teilungspunkte innerhalb der Klassifizierung darstellen. Die Teilungspunkte teilen die Daten rekursiv in Untergruppen auf, bis ein Endpunkt erreicht wird. Die Baumknoten an den Endpunkten werden als **Blätter** bezeichnet. Jedes Blatt weist den Mitgliedern seiner Untergruppe oder Klasse eine Beschriftung zu, die als **Klassenbeschriftung** bezeichnet wird.

Instanzgewichtungen und Klassengewichtungen

Standardmäßig wird davon ausgegangen, dass alle Datensätze und Klassen die gleiche relative Wichtigkeit aufweisen. Sie können dies Ändern, indem Sie den Mitgliedern eines dieser Elemente bzw. beider Elemente individuelle Gewichtungen zuweisen. Dies kann beispielsweise dann sinnvoll sein, wenn die Datenpunkte in den Trainingsdaten nicht realistisch auf die verschiedenen Kategorien verteilt sind. Mit Gewichtungen können Sie das Modell verzerren, um einen Ausgleich für diejenigen Kategorien zu bewirken, die in den Daten unterrepräsentiert sind. Durch die Erhöhung der Gewichtung für einen Zielwert sollte der Prozentsatz der richtigen Vorhersagen für die betreffende Kategorie erhöht werden.

Im Entscheidungsbaummodellierungsknoten können Sie zwei Arten von Gewichtungen angeben. **Instanzgewichtungen** weisen jeder Zeile von Eingabedaten eine Gewichtung zu. Die Gewichtungen werden in der Regel für die meisten Fälle als 1,0 angegeben. Höhere oder niedrigere Werte erhalten nur diejenigen Fälle, die wichtiger oder weniger wichtig sind als die Mehrheit (siehe folgende Tabelle).

Tabelle 5. Beispiel für Instanzgewichtung

Datensatz-ID	**Ziel**	**Instanzgewichtung**
E	MedikamentA	1,1
Z	MedikamentB	1,0
3	MedikamentA	1,0
4	MedikamentB	0,3

Klassengewichtungen weisen jeder Kategorie des Zielfelds eine Gewichtung zu. Dies ist in der folgenden Tabelle dargestellt.

Tabelle 6. Beispiel für Klassengewichtung

Klasse	**Klassengewichtung**
MedikamentA	1,0
MedikamentB	1,5

Beide Gewichtungstypen können gleichzeitig verwendet werden. In diesem Fall werden sie miteinander multipliziert und als Instanzgewichtungen verwendet. Wenn also die beiden vorigen Beispiele zusammen verwendet werden, führt dies zu den in der folgenden Tabelle dargestellten Instanzgewichtungen beim Algorithmus.

Tabelle 7. Beispiel für Berechnung der Instanzgewichtung

Datensatz-ID	**Berechnung**	**Instanzgewichtung**
E	1,1*1,0	1,1
Z	1,0*1,5	1,5
3	1,0*1,0	1,0
4	0,3*1,5	0,45

Netezza-Entscheidungsbaum - Feldoptionen

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

Vordefinierte Rollen verwenden. Bei dieser Option werden die Rolleneinstellungen (Ziele, Prädiktoren usw.) aus einem vorgeordneten Typknoten verwendet (oder die Registerkarte "Typen" eines vorgeordneten Quellenknotens).

Benutzerdefinierte Feldzuweisungen verwenden. Wählen Sie diese Option aus, um Ziele, Prädiktoren und andere Rollen manuell zuzuweisen.

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

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

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird. Die Werte in diesem Feld müssten für jeden Datensatz eindeutig sein (z. B. Kundennummern).

Instanzgewichtung. Indem Sie hier ein Feld angeben, können Sie anstelle der Klassengewichtungen oder zusätzlich zu den KIassengewichtungen (eine Gewichtung pro Kategorie für das Zielfeld) Instanzgewichtungen (eine Gewichtung pro Zeile an Eingabedaten) verwenden. Das hier angegebene Feld muss eine numerische Gewichtung für jede Zeile an Eingabedaten enthalten. Weitere Informationen finden Sie im Thema „Instanzgewichtungen und Klassengewichtungen” auf Seite 64.

Prädiktoren (Eingaben). Wählen Sie das/die Eingabefeld(er) aus. Dies gleicht in etwa der Einstellung der Feldrolle auf *Eingabe* in einem Typknoten.

IBM Data WH-Entscheidungsbaum - Erstellungsoptionen

Die folgenden Erstellungsoptionen sind für den Baumaufbau verfügbar:

Erweiterungsmaß. Diese Optionen steuern, wie die Baumerweiterung gemessen wird.

- **Unreinheitsmaß.** Dieses Maß ermittelt die beste Position für eine Baumteilung. Es handelt sich um ein Maß für die Variabilität in einer Untergruppe oder einem Datensegment. Ein niedriges Unreinheitsmaß gibt eine Gruppe an, in der die meisten Mitglieder ähnliche Werte für das Kriterium oder Zielfeld aufweisen.

  Die Maße **Entropie** und **Gini** werden unterstützt. Diese Maße basieren auf Wahrscheinlichkeiten der Zugehörigkeit zu einer Kategorie einer Verzweigung.
- **Maximale Baumtiefe.** Die maximale Anzahl der Ebenen, auf die ein Baum unterhalb des Stammknotens erweitert werden kann, d. h. die Anzahl der rekursiven Teilungen einer Stichprobe. Der Standardwert dieser Eigenschaft ist 10. Der Maximalwert, den Sie für diese Eigenschaft festlegen können, ist 62.

  Anmerkung: Wenn der Viewer im Modellnugget das Modell in Texform darstellt, werden maximal 12 Ebenen des Baums angezeigt.

Aufteilungskriterien. Diese Optionen steuern, wann die Aufteilung des Baums aufhört.

- **Mindestverbesserung für Aufteilungen.** Der Mindestwert der Unreinheitsreduzierung, bevor eine neue Aufteilung des Baums erfolgt. Das Ziel der Baumerstellung besteht darin, Untergruppen mit ähnlichen Ausgabewerten zu erstellen, um die Unreinheit in den einzelnen Knoten zu minimieren. Wenn die beste für eine Verzweigung mögliche Aufteilung die Unreinheit um weniger als den durch die Aufteilungskriterien vorgegebenen Betrag reduziert, wird die Verzweigung nicht aufgeteilt.
- **Mindestanzahl der Instanzen für eine Aufteilung.** Die Mindestanzahl von Datensätzen, die aufgeteilt werden können. Wenn weniger nicht aufgeteilte Datensätze verbleiben, werden keine weiteren Aufteilungen durchgeführt. Sie können dieses Feld verwenden, um die Erstellung kleiner Untergruppen im Baum zu verhindern.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Wählen Sie eine der folgenden Optionen aus:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.

- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

IBM Data WH-Entscheidungsbaumknoten - Klassengewichtungen

Hier können Sie den einzelnen Klassen Gewichtungen zuweisen. Standardmäßig wird allen Klassen der Wert 1 zugewiesen, sodass sie gleich gewichtet sind. Durch die Angabe von unterschiedlichen numerischen Gewichtungen für verschiedene Klassenbeschriftungen weisen Sie den Algorithmus an, die Trainingssets bestimmter Klassen entsprechend zu gewichten.

Doppelklicken Sie zum Ändern einer Gewichtung in die Spalte **Gewichtung** und nehmen Sie die gewünschten Änderungen vor.

Wert. Die Menge der Klassenbeschriftungen, abgeleitet aus den möglichen Werten des Zielfelds.

Gewichtung. Die Gewichtung, die einer bestimmten Klasse zugewiesen werden soll. Durch Zuweisen einer höheren Gewichtung zu einer Klasse reagiert das Modell im Vergleich zu den anderen Klassen sensibler auf diese Klasse.

Klassengewichtungen können in Verbindung mit Instanzgewichtungen verwendet werden. Weitere Informationen finden Sie im Thema „Instanzgewichtungen und Klassengewichtungen" auf Seite 64.

IBM Data WH-Entscheidungsbaumknoten - Baumreduzierung

Sie können die Reduzierungsoptionen verwenden, um Reduzierungskriterien für den Entscheidungsbaum festzulegen. Ziel der Reduzierung ist es, das Risiko der übermäßigen Anpassung zu verringern, indem zu stark erweiterte Untergruppen entfernt werden, welche die erwartete Genauigkeit für neue Daten nicht verbessern.

Reduzierungsmaß. Das standardmäßige Reduzierungsmaß, **Genauigkeit**, gewährleistet, dass die geschätzte Genauigkeit des Modells nach der Entfernung eines Blatts aus dem Baum innerhalb akzeptabler Grenzen bleibt. Nutzen Sie das Alternativmaß, **Gewichtete Genauigkeit**, wenn Sie die Klassengewichtungen in die Reduzierung mit einbeziehen möchten.

Daten für die Reduzierung. Sie können einen Teil oder alle Trainingsdaten verwenden, um die erwartete Genauigkeit der neuen Daten abzuschätzen. Alternativ können Sie zu diesem Zweck ein separates Dataset für die Reduzierung aus einer festgelegten Tabelle verwenden.

- **Alle Trainingsdaten verwenden.** Diese (standardmäßige) Option verwendet alle Trainingsdaten, um die Modellgenauigkeit zu schätzen.
- **% der Trainingsdaten für die Reduzierung verwenden.** Teilen Sie mithilfe dieser Option die Daten in zwei Gruppen (eine für das Training und eine für die Reduzierung) unter Verwendung des hier angegebenen Prozentsatzes für die Reduzierungsdaten.

  Wählen Sie das Feld **Ergebnisse replizieren** aus, wenn Sie einen Zufallsstartwert angeben möchten, um sicherzustellen, dass die Daten bei jeder Ausführung des Streams auf dieselbe Weise partitioniert werden. Sie können entweder eine ganze Zahl im Feld **Für Reduzierung verwendeter Startwert** angeben oder auf **Generieren** klicken, wodurch eine pseudozufällige ganze Zahl erstellt wird.
- **Daten aus einer vorhandenen Tabelle verwenden.** Geben Sie den Tabellennamen eines separaten Datasets für die Reduzierung an, anhand dessen die Modellgenauigkeit geschätzt wird. Diese Vorgehensweise wird als zuverlässiger betrachtet als die Nutzung von Trainingsdaten. Wenn Sie diese Option wählen, wird jedoch eventuell ein großes Subset von Daten aus dem Trainingsset entfernt, wodurch die Qualität des Entscheidungsbaum beeinträchtigt wird.

IBM Data WH - Lineare Regression

Bei linearen Modellen wird ein stetiges Ziel auf der Basis linearer Beziehungen zwischen dem Ziel und einem oder mehreren Prädiktoren vorhergesagt. Lineare Regressionsmodelle sind zwar auf die direkte Modellierung linearer Beziehungen beschränkt, sind jedoch relativ einfach und ergeben eine einfach zu interpretierende mathematische Formel für das Scoring. Lineare Modelle sind schnell, effizient und benutzer-

freundlich, auch wenn ihre Anwendbarkeit im Vergleich zu den durch stärker verfeinerte Regressionsalgorithmen produzierten eingeschränkt ist.

IBM Data WH - Erstellungsoptionen der linearen Regression

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

Einzelwertzerlegung zum Lösen von Gleichungen verwenden. Die Verwendung der Matrix zur Einzelwertzerlegung anstelle der ursprünglichen Matrix bietet den Vorteil einer größeren Robustheit gegenüber numerischen Fehlern und kann außerdem die Berechnung beschleunigen.

Konstanten Term in Modell einschließen. Durch Einschließen des konstanten Terms wird die Gesamtgenauigkeit der Lösung erhöht.

Modelldiagnose berechnen. Durch diese Option wird eine Anzahl von Diagnosen für das Modell berechnet. Die Ergebnisse werden in Matrizen oder Tabellen gespeichert , damit sie später überprüft werden können. Zu den Diagnosen gehören R-Quadrat, Residuenquadratsumme, Schätzung der Varianz, Standardabweichung, *p*-Wert und *t*-Wert.

Diese Diagnosen beziehen sich auf Validität und Brauchbarkeit des Modells. Sie sollten separate Diagnosen an den zugrunde liegenden Daten ausführen, um sicherzustellen, dass diese Linearitätsannahmen erfüllen.

IBM Data WH - KNN

Die Nächste-Nachbarn-Analyse ist eine Methode zur Klassifizierung von Fällen anhand ihrer Ähnlichkeit zu anderen Fällen. Im Maschinenlernen wurde es entwickelt, um Datenmuster zu erkennen, ohne dass eine exakte Übereinstimmung mit gespeicherten Mustern oder Fällen benötigt wird. Ähnliche Fälle 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 Zahl der zu untersuchenden nächsten Nachbarn festlegen; dieser Wert wird *k* genannt. Die Abbildungen zeigen, wie ein neuer Fall mithilfe von zwei verschiedenen Werten von *k* klassifiziert würde. Ist *k* = 5, wird der neue Fall der Kategorie *1* zugeordnet, da der Großteil der nächsten Nachbarn der Kategorie *1* angehört. Ist jedoch *k* = 9, wird der neue Fall der Kategorie *0* zugeordnet, da der Großteil der nächsten Nachbarn der Kategorie *0* angehört.

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.

IBM Data WH - allgemeine KNN-Modelloptionen

Auf der Registerkarte "Modelloptionen - Allgemein" können Sie bestimmen, ob Sie einen Namen für das Modell angeben oder automatisch erstellen lassen möchten. Sie können auch Optionen festlegen, die steuern, wie die Anzahl der nächsten Nachbarn berechnet wird, und Optionen für eine verbesserte Leistung und Genauigkeit des Modells angeben.

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

Nachbarn

Distanzmaß. Die zur Messung des Abstands zwischen Datenpunkten zu verwendende Methode. Größere Abstände deuten auf größere Unähnlichkeiten hin. Folgende Optionen stehen zur Auswahl:

- **Euklidisch**. (Standardvorgabe) Der Abstand zwischen zwei Punkten wird anhand einer geradlinigen Verbindung zwischen den Punkten berechnet.
- **Manhattan**. Der Abstand zwischen zwei Punkten wird als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet.
- **Canberra**. Ähnlich wie die Manhattan-Distanz, jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.
- **Maximum**. Der Abstand zwischen zwei Punkten wird als der größte Unterschied in einer beliebigen Koordinatendimension berechnet.

Anzahl der nächstgelegenen Nachbarn (k). Die Anzahl der nächsten Nachbarn für einen bestimmten Fall. Beachten Sie dabei, dass eine höhere Anzahl an Nachbarn nicht unbedingt ein präziseres Modell hervorbringt.

Der für *k* ausgewählte Wert legt die Balance zwischen der Vermeidung der Überanpassung (kann wichtig sein, insbesondere für "verrauschte" Daten) und der Auflösung (Ausgabe unterschiedlicher Vorhersagen für ähnliche Instanzen) fest. Normalerweise müssen Sie den Wert von *k* für jedes Dataset anpassen. Die typischen Werte liegen im Bereich von 1 bis zu mehreren Dutzend.

Leistung und Genauigkeit verbessern

Messungen vor Berechnung des Abstands standardisieren. Bei Aktivierung dieser Option werden die Messungen für stetige Eingabefelder standardisiert, bevor die Abstandswerte berechnet werden.

Core-Sets zur Leistungsverbesserung bei großen Datasets verwenden. Bei Aktivierung dieser Option wird Stichprobennahme mit Core-Sets verwendet, um die Berechnung zu beschleunigen, wenn große Datasets involviert sind.

IBM Data WH-KNN-Modelloptionen - Scoring-Optionen

Auf der Registerkarte "Modelloptionen - Scoring-Optionen" können Sie den Standardwert für eine Scoring-Option festlegen und den einzelnen Klassen relative Gewichtungen zuweisen.

Bereitstellen für Scoring

Eingabefelder einschließen. Gibt an, ob die Eingabefelder standardmäßig in das Scoring eingeschlossen werden.

Klassengewichtungen

Verwenden Sie diese Option, wenn Sie die relative Bedeutung einzelner Klassen bei der Modellerstellung ändern möchten.

Hinweis: Diese Option ist nur aktiviert, wenn Sie KNN für die Klassifizierung verwenden. Wenn Sie eine Regression durchführen (d. h., wenn der Typ des Zielfelds "Stetig" lautet), ist die Option inaktiviert.

Standardmäßig wird allen Klassen der Wert 1 zugewiesen, sodass sie gleich gewichtet sind. Durch die Angabe von unterschiedlichen numerischen Gewichtungen für verschiedene Klassenbeschriftungen weisen Sie den Algorithmus an, die Trainingssets bestimmter Klassen entsprechend zu gewichten.

Doppelklicken Sie zum Ändern einer Gewichtung in die Spalte **Gewichtung** und nehmen Sie die gewünschten Änderungen vor.

Wert. Die Menge der Klassenbeschriftungen, abgeleitet aus den möglichen Werten des Zielfelds.

Gewichtung. Die Gewichtung, die einer bestimmten Klasse zugewiesen werden soll. Durch Zuweisen einer höheren Gewichtung zu einer Klasse reagiert das Modell im Vergleich zu den anderen Klassen sensibler auf diese Klasse.

IBM Data WH - K-Means

Der K-Means-Knoten implementiert den *k*-Means-Algorithmus, der eine Methode der Clusteranalyse bietet. Mit diesem Knoten können Sie ein Clustering der Datasets in einzelne Gruppen vornehmen.

Bei dem Algorithmus handelt es sich um einen distanzbasierten Clusteralgorithmus, der auf einer Distanzmetrik (Funktion) zur Messung der Ähnlichkeit zwischen Datenpunkten beruht. Die Datenpunkte werden dem nächsten Cluster gemäß der verwendeten Distanzmetrik zugewiesen.

Bei diesem Algorithmus werden mehrere Iterationen desselben Grundverfahren durchgeführt. Dabei wird jede Trainingsinstanz dem nächstgelegenen Cluster zugewiesen (in Bezug auf die angegebene Distanzfunktion, angewendet auf Instanz und Clusterzentrum). Anschließend werden alle Clusterzentren als Attribut-Mittelwertvektoren der Instanzen neu berechnet, die den jeweiligen Clustern zugewiesen wurden.

IBM Data WH - K-Means-Feldoptionen

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

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

Registerkarte mit K-Means-Erstellungsoptionen von IBM Data WH

Durch Festlegen der Erstellungsoptionen können Sie die Erstellung des Modells Ihren Anforderungen entsprechend anpassen.

Wenn Sie ein Modell mit den Standardoptionen erstellen wollen, klicken Sie auf **Ausführen**.

Distanzmaß. Dieser Parameter definiert die Methode für die Messung der Distanz zwischen Datenpunkten. Größere Distanzen geben größere Unähnlichkeiten an. Wählen Sie eine der folgenden Optionen aus:

- **Euklidisch.** Das euklidische Maß ist die geradlinige Distanz zwischen zwei Datenpunkten.
- **Euklidisch normalisiert.** Das euklidisch normalisierte Maß ähnelt dem euklidischen Maß, wird jedoch durch das Quadrat der Standardabweichung normalisiert. Im Gegensatz zum euklidischen Maß ist das euklidisch normalisierte Maß zudem skaleninvariant.
- **Mahalanobis.** Das Mahalanobis-Maß ist ein verallgemeinertes euklidisches Maß, das Korrelationen von Eingabedaten berücksichtigt. Wie das euklidisch normalisierte Maß ist das Mahalanobis-Maß skaleninvariant.
- **Manhattan.** Das Manhattan-Maß ist die Distanz zwischen zwei Datenpunkten, die als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet wird.
- **Canberra.** Das Canberra-Maß ähnelt dem Manhattan-Maß, ist jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.
- **Maximum.** Das Maximum-Maß ist die Distanz zwischen zwei Datenpunkten, die als größter Unterschied in einer beliebigen Koordinatendimension berechnet wird.

Anzahl der Cluster. Dieser Parameter definiert die Anzahl der zu erstellenden Cluster.

Maximale Anzahl an Iterationen. Bei diesem Algorithmus werden mehrere Iterationen desselben Prozesses durchgeführt. Dieser Parameter definiert die Anzahl von Iterationen, nach denen das Modelltraining beendet wird.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Wählen Sie eine der folgenden Optionen aus:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.
- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

Ergebnisse replizieren. Aktivieren Sie dieses Kontrollkästchen, wenn Sie einen Startwert für Zufallszahlen festlegen wollen, um Analysen zu replizieren. Sie können eine ganze Zahl angeben oder durch Klicken auf **Generieren** eine pseudozufällige ganze Zahl erstellen.

IBM Data WH - Naive Bayes

Naive Bayes ist ein bekannter Algorithmus für Klassifizierungsprobleme. Das Modell wird als *naive* bezeichnet, weil es alle vorgeschlagenen Vorhersagevariablen als voneinander unabhängig behandelt. Naive Bayes ist ein schneller, skalierbarer Algorithmus, der für Kombinationen von Attributen und für das Zielattribut bedingte Wahrscheinlichkeiten berechnet. Aus den Trainingsdaten wird eine unabhängige Wahrscheinlichkeit ermittelt. Diese liefert die Wahrscheinlichkeit jeder Zielklasse anhand des Vorkommens der einzelnen Wertekategorien aus jeder einzelnen Eingabevariablen.

Netezza-Bayes-Netz

Ein Bayes-Netz ist ein Modell, das Variablen in einem Dataset und die probabilistischen bzw. bedingten Unabhängigkeiten zwischen diesen Variablen anzeigt. Mithilfe des Knotens "Netezza-Bayes-Netz" können Sie ein Wahrscheinlichkeitsmodell erstellen, indem Sie beobachtete und aufgezeichnete Hinweise mit gesundem Menschenverstand kombinieren, um die Wahrscheinlichkeit des Vorkommens unter Verwendung scheinbar nicht miteinander verknüpfter Attribute zu ermitteln.

Netezza-Bayes-Netz - Feldoptionen

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

Bei diesem Knoten wird das Zielfeld lediglich für das Scoring benötigt, weshalb es nicht auf dieser Registerkarte angezeigt wird. Sie können das Ziel auf einem Typknoten, auf der Registerkarte "Modelloptionen" dieses Knotens oder auf der Registerkarte "Einstellungen" des Modellnuggets festlegen bzw. ändern. Weitere Informationen finden Sie im Thema „Nugget für "Netezza-Bayes-Netz" - Registerkarte "Einstellungen"” auf Seite 82.

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

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

Netezza-Bayes-Netz - Erstellungsoptionen

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

Basisindex. Die numerische Kennung, die dem ersten Attribut (Eingabefeld) zur einfacheren internen Verwaltung zugewiesen werden soll.

Stichprobenumfang. Der Umfang der zu ziehenden Stichprobe, wenn die Anzahl der Attribute so groß ist, dass sie zu einer unakzeptabel langen Verarbeitungsdauer führen würde.

Zusätzliche Informationen während der Ausführung anzeigen. Wenn dieses Kontrollkästchen aktiviert ist (Standard), werden in einem Nachrichtendialogfeld zusätzliche Fortschrittsinformationen angezeigt.

Netezza-Zeitreihe

Eine **Zeitreihe** ist eine Folge numerischer Datenwerte, die zu aufeinander folgenden Zeitpunkten (wenn auch nicht unbedingt in regelmäßigen Abständen) gemessen werden, beispielsweise die täglichen Aktienkurse oder wöchentliche Umsatzdaten. Die Analyse solcher Daten kann beispielsweise dafür nützlich sein, um bestimmte Verhaltensmuster, wie Trends oder Saisonalität (ein sich wiederholendes Muster), hervorzuheben oder das zukünftige Verhalten aus vergangenen Ereignissen vorherzusagen.

"Netezza-Zeitreihe" unterstützt folgende Zeitreihenalgorithmen.

- Spektralanalyse
- Exponentielle Glättung
- Autoregressiver, integrierter gleitender Durchschnitt (AutoRegressive Integrated Moving Average, ARIMA)
- Saisonale Zerlegung in Trends

Diese Algorithmen zerlegen eine Zeitreihe in eine Trend-Komponente und eine saisonale Komponente. Diese Komponenten werden dann analysiert, um ein Modell zu erstellen, das für die Vorhersage verwendet werden kann.

Spektralanalyse wird zur Identifizierung von periodischem Verhalten bei Zeitreihen verwendet. Bei Zeitreihen, die aus mehreren zugrunde liegenden Periodizitäten bestehen, oder wenn die Daten Zufallsrauschen in erheblichem Umfang aufweisen, bietet die Spektralanalyse die beste Methode zur Identifizierung periodischer Komponenten. Mit dieser Methode werden die Häufigkeiten von periodischem Verhalten durch die Transformation der Reihe aus der Zeitdomäne in eine Reihe der Häufigkeitsdomäne ermittelt.

Exponentielles Glätten ist eine Vorhersagemethode, bei der gewichtete Werte aus früheren Beobachtungen der Zeitreihe verwendet werden, um zukünftige Werte vorherzusagen. Beim exponentiellen Glätten nimmt der Einfluss von Beobachtungen im Laufe der Zeit exponentiell ab. Bei dieser Methode wird jeweils ein Punkt vorhergesagt und diese Vorhersagen werden angepasst, wenn neue Daten eingehen, wobei Addition, Trend und Saisonalität berücksichtigt werden.

ARIMA-Modelle stellen ausgereifte Methoden zur Modellierung von Trendkomponenten und saisonalen Komponenten als Modelle mit exponentiellem Glätten bereit. Bei dieser Methode müssen die Ordnung der Autoregression, die Ordnung des gleitenden Durchschnitts und der Grad der Differenzenbildung angegeben werden.

Hinweis: In der Praxis bedeutet dies, dass ARIMA-Modelle besonders nützlich sind, wenn Sie Prädiktoren einschließen möchten, die zur Erklärung des Verhaltens der vorhergesagten Reihe beitragen können, wie beispielsweise die Anzahl der versendeten Kataloge oder die Anzahl der Aufrufe einer Unternehmens-

webseite. Modelle mit exponentiellem Glätten beschreiben das Verhalten der Zeitreihen, ohne dass versucht wird zu erklären, warum sich die Zeitreihe so verhält.

Saisonale Zerlegung in Trends entfernt periodisches Verhalten aus der Zeitreihe, um eine Trendanalyse durchzuführen, und wählt dann eine Grundform für den Trend aus, beispielsweise eine quadratische Funktion. Diese Grundformen weisen eine Reihe von Parametern auf, deren Werte bestimmt werden, um den mittleren quadratischen Fehler der Residuen (d. h. die Differenzen zwischen den angepassten und den beobachteten Werten der Zeitreihe) zu minimieren.

Interpolation von Werten in Netezza-Zeitreihen

Interpolation ist das Schätzen und Einfügen fehlender Werte in Zeitreihendaten.

Wenn die Intervalle der Zeitreihe regelmäßig sind, einige Werte jedoch fehlen, können die fehlenden Werte mittels linearer Interpolation geschätzt werden. Betrachten Sie die folgende Reihe der monatlichen Passagierankunftszahlen an einem Flughafenterminal.

Tabelle 8. Monatliche Ankünfte am Passagierterminal

Monat	**Passagiere**
3	3.500.000
4	3.900.000
5	-
6	3.400.000
7	4.500.000
8	3.900.000
9	5.800.000
10	6.000.000

In diesem Fall würde die lineare Interpolation den fehlenden Wert für Monat 5 auf 3.650.000 schätzen (Mitte zwischen 4 und 6).

Unregelmäßige Intervalle werden anders gehandhabt. Betrachten Sie folgende Reihe von Temperaturmessungen.

Tabelle 9. Temperaturmessungen

Datum	**Zeit**	**Temperatur**
24.7.2011	7:00	57
24.7.2011	14:00	75
24.7.2011	21:00	72
25.7.2011	7:15	59
25.7.2011	14:00	77
25.7.2011	20:55	74
27.7.2011	7:00	60
27.7.2011	14:00	78
27.7.2011	22:00	74

Diese Messungen wurden während drei Tagen an jeweils drei Zeitpunkten vorgenommen, jedoch zu unterschiedlichen Uhrzeiten, die nicht alle zwischen den verschiedenen Tagen übereinstimmen. Außerdem folgen nur zwei der Tage unmittelbar aufeinander.

Mit dieser Situation kann auf zwei verschiedenen Weisen umgegangen werden: Berechnung von Aggregatwerten oder Bestimmen einer Schrittweite.

Bei den Aggregatwerten könnte es sich um tägliche Aggregate handeln, die anhand einer Formel auf der Grundlage semantischer Kenntnisse über die Daten berechnet werden. Dadurch könnte sich folgendes Dataset ergeben.

Tabelle 10. Temperaturmessungen (aggregiert)

Datum	**Zeit**	**Temperatur**
24.7.2011	24:00	69
25.7.2011	24:00	71
26.7.2011	24:00	null
27.7.2011	24:00	72

Alternativ kann der Algorithmus die Reihe als eindeutige Reihe behandeln und eine geeignete Schrittweite bestimmen. In diesem Fall könnte vom Algorithmus eine Schrittweite von 8 Stunden festgelegt werden, was zu folgenden Daten führt.

Tabelle 11. Temperaturmessungen mit berechneter Schrittweite

Datum	**Zeit**	**Temperatur**
24.7.2011	6:00	
24.7.2011	14:00	75
24.7.2011	22:00	
25.7.2011	6:00	
25.7.2011	14:00	77
25.7.2011	22:00	
26.7.2011	6:00	
26.7.2011	14:00	
26.7.2011	22:00	
27.7.2011	6:00	
27.7.2011	14:00	78
27.7.2011	22:00	74

Hier entsprechen nur vier Messwerte den ursprünglichen Messungen, mithilfe der anderen bekannten Werte aus der ursprünglichen Reihe können die fehlenden Werte jedoch wiederum mittels Interpolation berechnet werden.

Netezza-Zeitreihen - Feldoptionen

Auf der Registerkarte "Felder" geben Sie Rollen für die Eingabefelder in den Quellendaten an.

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

Ziel. Wählen Sie ein Feld als Ziel für die Vorhersage aus. Dieses Feld muss das Messniveau "Stetig" aufweisen.

(Prädiktor-)Zeitpunkte. (erforderlich) Das Eingabefeld, das die Datums- bzw. Zeitwerte für die Zeitreihe enthält. Dabei muss es sich um ein Feld mit dem Messniveau "Stetig" oder "Kategorial" und dem Datenspeichertyp "Datum", "Zeit", "Zeitmarke" oder "Numerisch" handeln. Durch den Datenspeichertyp des hier angegebenen Felds wird auch der Eingabetyp für einige Felder auf anderen Registerkarten dieses Modellierungsknotens definiert.

(Prädiktor-)Zeitreihen-IDs (nach). Ein Feld mit Zeitreihen-IDs; verwenden Sie dies, wenn die Eingabe mehrere Zeitreihen enthält.

Netezza-Zeitreihen - Erstellungsoptionen

Es gibt zwei Ebenen von Erstellungsoptionen:

- Einfach - Einstellungen für Algorithmusauswahl, Interpolation und Zeitbereich
- Erweitert - Einstellungen für Vorhersagen

In diesem Abschnitt werden die einfachen Optionen beschrieben.

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

Algorithmus

Dies sind die Einstellungen, die sich auf den zu verwendenden Zeitreihenalgorithmus beziehen.

Algorithmusname. Wählen Sie den zu verwendenden Zeitreihenalgorithmus aus. Die verfügbaren Algorithmen sind **Spektralanalyse**, **Exponentielles Glätten** (Standardeinstellung), **ARIMA** und **Saisonale Zerlegung in Trends**. Weitere Informationen finden Sie im Thema „Netezza-Zeitreihe” auf Seite 71.

Trend. (Nur bei "Exponentielles Glätten") Einfaches exponentielles Glätten funktioniert nicht gut, wenn die Zeitreihe einen Trend aufweist. Geben Sie in diesem Feld den Trend an, sofern vorhanden, damit er vom Algorithmus berücksichtigt werden kann.

- **Systembestimmt.** (Standardvorgabe) Das System versucht, den optimalen Wert für diesen Parameter zu finden.
- **Keine (N).** Die Zeitreihe weist keinen Trend auf.
- **Additiv (A).** Ein Trend, der im Laufe der Zeit stetig zunimmt.
- **Gedämpft additiv (DA).** Ein additiver Trend, der schließlich verschwindet.
- **Multiplikativ (M).** Ein Trend, der im Laufe der Zeit zunimmt, typischerweise rascher als ein stetiger additiver Trend.
- **Gedämpft multiplikativ (DM).** Ein multiplikativer Trend, der schließlich verschwindet.

Saisonalität. (Nur bei "Exponentielles Glätten") Geben Sie in diesem Feld an, ob die Zeitreihe saisonale Muster in den Daten aufweist.

- **Systembestimmt.** (Standardvorgabe) Das System versucht, den optimalen Wert für diesen Parameter zu finden.
- **Keine (N).** Die Zeitreihe weist keine saisonalen Muster auf.
- **Additiv (A).** Das Muster der saisonalen Schwankungen weist einen stetigen Aufwärtstrend im Zeitverlauf auf.
- **Multiplikativ (M).** Wie additive Saisonalität, jedoch erhöht sich zusätzlich die Amplitude (Abstand zwischen Minima und Maxima) der saisonalen Schwankungen relativ zum allgemeinen Aufwärtstrend der Schwankungen.

Systembestimmte Einstellungen für ARIMA verwenden. (Nur bei "ARIMA") Wählen Sie diese Option aus, wenn das System die Einstellungen für den ARIMA-Algorithmus festlegen soll.

Angeben. (Nur bei "ARIMA") Wählen Sie diese Option aus und klicken Sie auf die Schaltfläche, um die ARIMA-Einstellungen manuell anzugeben.

Interpolation

Wenn in den Quellendaten der Zeitreihe Werte fehlen, können Sie hier eine Methode auswählen, nach der die Lücken in den Daten durch Schätzwerte aufgefüllt werden. Weitere Informationen finden Sie im Thema „Interpolation von Werten in Netezza-Zeitreihen" auf Seite 72.

- **Linear.** Wählen Sie diese Methode aus, wenn die Intervalle in der Zeitreihe regelmäßig sind, jedoch bestimmte Werte fehlen.
- **Exponentielle Splines.** Passt eine glatte Kurve an die Stellen an, an denen die Werte der bekannten Datenpunkte stark steigen oder sinken.
- **Kubische Splines.** Passt eine glatte Kurve an die bekannten Datenpunkte an, um die fehlenden Werte zu schätzen.

Zeitbereich

Hier können Sie auswählen, ob Sie zur Erstellung des Modells den vollständigen Bereich der Daten in der Zeitreihe oder ein zusammenhängendes Subset dieser Daten verwenden möchten. Die gültigen Eingaben für diese Felder werden durch den Datenspeichertyp des Felds festgelegt, das auf der Registerkarte "Felder" für "Zeitpunkte" angegeben wurde. Weitere Informationen finden Sie im Thema „Netezza-Zeitreihen - Feldoptionen" auf Seite 73.

- **Frühesten und spätesten Zeitpunkt in Daten verwenden.** Verwenden Sie diese Option, wenn Sie den vollständigen Bereich der Zeitreihendaten verwenden möchten.
- **Zeitfenster angeben.** Verwenden Sie diese Option, wenn Sie nur einen Teilbereich der Zeitreihe verwenden möchten. Geben Sie die Grenzen des Bereichs in den Feldern **Frühester Zeitpunkt (von)** und **Spätester Zeitpunkt (bis)** an.

ARIMA-Struktur

Sie können die Werte der verschiedenen nicht saisonalen und saisonalen Komponenten des ARIMA-Modells angeben. Setzen Sie dabei jeweils den Operator auf **=** (gleich) oder **<=** (kleiner-gleich) und geben Sie dann den Wert in das angrenzende Feld ein. Die Werte müssen nicht negative ganze Zahlen sein und die Maße angeben.

Nicht saisonal. Die Werte für die verschiedenen nicht saisonalen Komponenten des Modells.

- **Autokorrelationsmaße (p).** Die Anzahl autoregressiver Ordnungen im Modell. Autoregressive Ordnungen geben die zurückliegenden Werte der Zeitreihe an, die für die Vorhersage der aktuellen Werte verwendet werden. Eine autoregressive Ordnung von 2 gibt beispielsweise an, dass die Werte der Zeitreihe, die zwei Zeitperioden zurückliegt, für die Vorhersage der aktuellen Werte verwendet wird.
- **Ableitung (d).** Gibt die Ordnung der Differenzierung an, die vor dem Schätzen der Modelle auf die Zeitreihe angewendet wurde. Differenzierung ist erforderlich, wenn Trends vorhanden sind. (Zeitreihen mit Trends sind normalerweise nicht stationär und bei der ARIMA-Modellierung wird Stationarität angenommen.) Mithilfe der Differenzierung werden die Effekte der Trends entfernt. Die Ordnung der Differenzierung entspricht dem Grad des Trends der Zeitreihe: Differenzierung erster Ordnung erklärt lineare Trends, Differenzierung zweiter Ordnung erklärt quadratische Trends usw.
- **Gleitender Durchschnitt (q).** Die Anzahl von Ordnungen des gleitenden Durchschnitts im Modell. Ordnungen des gleitenden Durchschnitts geben an, wie Abweichungen vom Mittelwert der Zeitreihe für zurückliegende Werte zum Vorhersagen der aktuellen Werte verwendet werden. Moving-Average-Ordnungen von 1 und 2 geben beispielsweise an, dass beim Vorhersagen der aktuellen Werte der Zeitreihe Abweichungen vom Mittelwert der Zeitreihe von den beiden letzten Zeitperioden berücksichtigt werden sollen.

Saisonal. Saisonale Komponenten von Autokorrelation (SP), Ableitung (SD) und gleitendem Durchschnitt haben jeweils dieselbe Rolle wie ihr nicht saisonales Gegenstück. Bei saisonalen Ordnungen werden die Werte der aktuellen Zeitreihe jedoch von Werten zurückliegender Zeitreihen beeinflusst, die durch eine oder mehrere saisonalen Perioden getrennt sind. Bei monatlichen Daten (saisonale Periode von 12) beispielsweise bedeutet eine saisonale Ordnung von 1, dass der Wert der aktuellen Zeitreihe durch den Zeit-

reihenwert beeinflusst wird, der 12 Perioden vor dem aktuellen liegt. Eine saisonale Ordnung von 1 entspricht bei monatlichen Daten einer nicht saisonalen Ordnung von 12.

Die saisonalen Einstellungen werden nur berücksichtigt, wenn Saisonalität in den Daten ermittelt wurde oder wenn Sie auf der Registerkarte "Erweitert" Einstellungen für die Periode angeben.

Erstellungsoptionen für Netezza-Zeitreihen - Erweitert

Mit den erweiterten Einstellungen können Sie Optionen für Vorhersagen angeben.

Systembestimmte Einstellungen für Modellerstellungsoptionen verwenden. Wählen Sie diese Option aus, wenn die erweiterten Einstellungen vom System festgelegt werden sollen.

Angeben. Wählen Sie diese Option aus, wenn Sie die erweiterten Optionen manuell angeben möchten. (Die Option ist beim Algorithmus "Spektralanalyse" inaktiviert).

- **Periode/Periodeneinheiten.** Die Zeitperiode, nach der sich ein bestimmtes charakteristisches Verhalten der Zeitreihe wiederholt. Bei einer Zeitreihe aus wöchentlichen Verkaufszahlen würden Sie beispielsweise 1 für die Periode und `Wochen` für die Einheiten angeben. **Periode** muss eine nicht negative ganze Zahl sein; für **Periodeneinheiten** stehen die Optionen **Millisekunden**, **Sekunden**, **Minuten**, **Stunden**, **Tage**, **Wochen**, **Quartale** und **Jahre** zur Auswahl. Legen Sie keine **Periodeneinheiten** fest, wenn die Option **Periode** nicht gesetzt wurde oder wenn der Zeittyp nicht numerisch ist. Wenn Sie jedoch eine **Periode** angeben, müssen Sie auch **Periodeneinheiten** festlegen.

Einstellungen für die Vorhersage. Sie können auswählen, ob Vorhersagen bis einschließlich eines bestimmten Zeitpunkts oder zu konkreten Zeitpunkten erstellt werden sollen. Die gültigen Eingaben für diese Felder werden durch den Datenspeichertyp des Felds festgelegt, das auf der Registerkarte "Felder" für "Zeitpunkte" angegeben wurde. Weitere Informationen finden Sie im Thema „Netezza-Zeitreihen - Feldoptionen” auf Seite 73.

- **Vorhersagehorizont.** Wählen Sie diese Option aus, wenn Sie ausschließlich einen Endpunkt für die Vorhersageerstellung angeben möchten. Vorhersagen werden bis zu diesem Zeitpunkt erstellt.
- **Vorhersagezeiten.** Wählen Sie diese Option aus, um einen oder mehrere Zeitpunkte anzugeben, zu denen Vorhersagen erstellt werden sollen. Klicken Sie auf **Hinzufügen**, um eine neue Zeile zu der Tabelle der Zeitpunkte hinzuzufügen. Um eine Zeile zu löschen, wählen Sie die Zeile aus und klicken Sie dann auf **Löschen**.

Netezza-Zeitreihenmodell - Optionen

Auf der Registerkarte "Modelloptionen" können Sie bestimmen, ob Sie einen Namen für das Modell angeben oder automatisch erstellen lassen möchten. Sie können auch Standardwerte für die Modellausgabeoptionen festlegen.

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

Für Scoring bereitstellen. Sie können hier die Standardwerte für die Scoring-Optionen festlegen, die im Dialogfeld für das Modellnugget angezeigt werden.

- **Historische Werte in Ergebnis einschließen.** Standardmäßig enthält die Modellausgabe nicht die historischen Datenwerte (diejenigen, die für die Vorhersage verwendet wurden). Aktivieren Sie dieses Kontrollkästchen, um diese Werte mit einzuschließen.
- **Interpolierte Werte in Ergebnis einschließen.** Wenn Sie historische Werte in das Ergebnis mit einschließen, können Sie durch Aktivieren dieses Kontrollkästchens auch die interpolierten Werte, sofern vorhanden, mit einschließen. Beachten Sie, dass die Interpolation nur für historische Daten ausgelegt ist. Daher ist dieses Kontrollkästchen nur verfügbar, wenn die Option **Historische Werte in Ergebnis einschließen** ausgewählt wurde. Weitere Informationen finden Sie im Thema „Interpolation von Werten in Netezza-Zeitreihen” auf Seite 72.

IBM Data WH - TwoStep

Der TwoStep-Knoten implementiert den Algorithmus TwoStep, der eine Methode zum Bilden von Datenclustern für große Datasets bereitstellt.

Mit diesem Knoten können sie einen Datencluster unter Berücksichtigung der verfügbaren Ressourcen wie Speicher und Zeitvorgaben bilden.

Der Algorithmus TwoStep ist ein Algorithmus für das Datenbankmining, der auf folgenden Weise Datencluster bildet:

1. Ein CF-Baum (Clustering Feature) wird erstellt. Dieser hochgradig ausgewogene Baum speichert Clustering-Funktionen für das hierarchische Clustering, bei denen ähnliche Eingabedatensätze Teil derselben Baumknoten werden.
2. Die Blätter des CF-Baums werden speicherintern hierarchisch in Gruppen zusammengefasst, um das endgültige Clustering-Ergebnis zu generieren. Die beste Anzahl Cluster wird automatisch bestimmt. Wenn Sie eine maximale Anzahl Cluster angeben, wird die beste Anzahl Cluster innerhalb des angegebenen Grenzwerts bestimmt.
3. Das Clustering-Ergebnis wird in einem zweiten Schritt optimiert, in dem ein dem K-Means-Algorithmus ähnlicher Algorithmus auf die Daten angewendet wird.

IBM Data WH - TwoStep-Feldoptionen

Durch Festlegen der Feldoptionen können Sie angeben, dass die in vorgeordneten Knoten definierten Feldrolleneinstellungen verwendet werden. Sie können die Feldzuweisungen auch manuell vornehmen.

Element auswählen. Wählen Sie diese Option aus, um die Rolleneinstellungen eines vorgeordneten Typknotens oder von der Registerkarte **Typen** eines vorgeordneten Quellenknoten zu verwenden. Rolleneinstellungen sind beispielsweise Ziele und Prädiktoren.

Benutzerdefinierte Feldzuweisungen verwenden. Wählen Sie diese Option aus, wenn Sie Ziele, Prädiktoren und andere Rollen manuell zuweisen möchten.

Felder. Verwenden Sie die Pfeile, um den Rollenfeldern rechts Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

IBM Data WH - TwoStep-Erstellungsoptionen

Durch Festlegen der Erstellungsoptionen können Sie die Erstellung des Modells Ihren Anforderungen entsprechend anpassen.

Wenn Sie ein Modell mit den Standardoptionen erstellen wollen, klicken Sie auf **Ausführen**.

Distanzmaß. Dieser Parameter definiert die Methode für die Messung der Distanz zwischen Datenpunkten. Größere Distanzen geben größere Unähnlichkeiten an. Folgende Optionen stehen zur Auswahl:

- **Log-Likelihood.** Mit dem Likelihood-Maß wird eine Wahrscheinlichkeitsverteilung für die Variablen vorgenommen. Bei stetigen Variablen wird von einer Normalverteilung, bei kategorialen Variablen von einer multinomialen Verteilung ausgegangen. Bei allen Variablen wird davon ausgegangen, dass sie unabhängig sind.
- **Euklidisch.** Das euklidische Maß ist die geradlinige Distanz zwischen zwei Datenpunkten.
- **Euklidisch normalisiert.** Das euklidisch normalisierte Maß ähnelt dem euklidischen Maß, wird jedoch durch das Quadrat der Standardabweichung normalisiert. Im Gegensatz zum euklidischen Maß ist das euklidisch normalisierte Maß zudem skaleninvariant.

Clusteranzahl. Dieser Parameter definiert die Anzahl der zu erstellenden Cluster. Folgende Optionen stehen zur Auswahl:

- **Anzahl der Cluster automatisch berechnen.** Die Anzahl der Cluster wird automatisch berechnet. Sie können die maximale Anzahl der Cluster im Feld **Maximum** angeben.
- **Anzahl der Cluster angeben.** Geben Sie an, wie viele Cluster erstellt werden sollen.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Folgende Optionen stehen zur Auswahl:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.
- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

Ergebnisse replizieren. Aktivieren Sie dieses Kontrollkästchen, wenn Sie einen Startwert für Zufallszahlen festlegen wollen, um Analysen zu replizieren. Sie können eine ganze Zahl angeben oder durch Klicken auf **Generieren** eine pseudozufällige ganze Zahl erstellen.

IBM Data WH - PCA

Die Hauptkomponentenanalyse (PCA) ist ein leistungsstarkes Verfahren zur Datenreduktion, mit dem die Komplexität der Daten verringert werden soll. PCA findet lineare Kombinationen der Eingabefelder, die die Varianz im gesamten Set der Felder am besten erfassen, wenn die Komponenten orthogonal zueinander (nicht miteinander korreliert) sind. Das Ziel besteht darin, eine kleinere Zahl abgeleiteter Felder (die Hauptkomponenten) zu finden, mit denen die Informationen in der ursprünglichen Menge der Eingabefelder effektiv zusammengefasst werden können.

Anmerkung: Beim Scoring des Modells kann ein Fehler auftreten, wenn Feldnamen in Kleinbuchstaben verwendet werden. Dies ist ein bekannter Fehler in Db2 Data Warehouse, der umgangen werden kann, indem vor dem Scoring alle Feldnamen in Großbuchstaben geändert werden.

IBM Data WH - PCA-Feldoptionen

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

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

IBM Data WH - PCA-Erstellungsoptionen

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

Daten vor PCA-Berechnung zentrieren. Wenn diese Option aktiviert ist (Standard), wird vor der Analyse Datenzentrierung (auch als Mittelwertsubtraktion bezeichnet) durchgeführt. Die Datenzentrierung ist notwendig, um sicherzustellen, dass die erste Hauptkomponente die Richtung der Maximalvarianz beschreibt. Andernfalls korrespondiert die Komponente möglicherweise enger mit dem Mittelwert der Daten. Diese Option wird normalerweise nur zur Leistungsverbesserung inaktiviert, sofern die Daten bereits auf diese Weise vorbereitet wurden.

Vor PCA-Berechnung Datenskalierung durchführen. Mit dieser Option wird vor der Analyse eine Datenskalierung durchgeführt. Auf diese Weise wird die Analyse eventuell weniger arbiträr, wenn verschiedene Variablen in verschiedenen Einheiten gemessen werden. In ihrer einfachsten Form kann Datenskalierung erreicht werden, indem jede Variable durch ihre Standardvariation dividiert wird.

Weniger genaue, jedoch schnellere Methode zur PCA-Berechnung verwenden. Bei dieser Option verwendet der Algorithmus eine genauere, aber schnellere Methode (forceEigensolve) zur Ermittlung der Hauptkomponenten.

Verwalten von IBM Data WH- und Netezza-Modellen

IBM Data Warehouse- und IBM Netezza Analytics-Modelle werden auf dieselbe Weise zum Erstellungsbereich und zur Modellpalette hinzugefügt wie andere IBM SPSS Modeler-Modelle und können auf annähernd dieselbe Weise verwendet werden. Es gibt jedoch einige wichtige Unterschiede, die sich daraus ergeben, dass zurzeit jedes in IBM SPSS Modeler erstellte IBM Data Warehouse- oder IBM Netezza Analytics-Modell auf ein in einem Datenbankserver gespeichertes Modell verweist. Damit ein Stream ordnungsgemäß funktioniert, muss eine Verbindung mit der Datenbank hergestellt werden, in der das Modell erstellt wurde, und die Modelltabelle darf nicht von einem externen Prozess geändert worden sein.

Scoring von IBM Data Warehouse- und IBM Netezza Analytics-Modellen

Modelle werden im Erstellungsbereich durch ein goldenes Modellnugget-Symbol repräsentiert. Der Hauptzweck eines Nuggets ist das Scoring von Daten, um Vorhersagen zu generieren oder eine weitere Analyse der Modelleigenschaften zu erlauben. Scores werden in Form eines oder mehrerer zusätzlicher Datenfelder hinzugefügt, die durch Verknüpfen eines Tabellenknotens mit dem Nugget und Ausführen des betreffenden Zweigs des Streams sichtbar gemacht werden können, wie weiter unten in diesem Abschnitt beschrieben. Einige Nugget-Dialogfelder, beispielsweise diejenigen für Entscheidungsbaum oder Regressionsbaum, enthalten zusätzlich die Registerkarte "Modell", die eine visuelle Darstellung des Modells bietet.

Die zusätzlichen Felder sind durch das Präfix $<ID>- gekennzeichnet, das dem Namen des Zielfelds hinzugefügt wird. Dabei hängt <ID> vom Modell ab und gibt den Typ der hinzugefügten Informationen an. Die unterschiedlichen Kennzeichner werden in den Themen für die einzelnen Modellnuggets beschrieben.

Führen Sie zur Anzeige der Scores folgende Schritte aus:

1. Verbinden Sie einen Tabellenknoten mit dem Modellnugget.
2. Öffnen Sie den Tabellenknoten.
3. Klicken Sie auf **Ausführen**.
4. Blättern Sie im Tabellenausgabefenster nach rechts, um die zusätzlichen Felder und ihre Scores anzuzeigen.

Registerkarte "Server" für IBM Data WH- und Netezza-Modellnuggets

Auf der Registerkarte "Server" können Sie Serveroptionen zum Scoren des Modells festlegen. Sie können entweder eine Serververbindung weiterverwenden, die weiter oben im Stream angegeben wurde, oder Sie können die Daten in eine andere Datenbank verschieben, die Sie hier angeben.

IBM Data Warehouse-Serverdetails. Hier geben Sie die Verbindungsdetails für die für das Modell zu verwendende Datenbank an.

- **Vorgeordnete Verbindung verwenden.** (Standardeinstellung) Verwendet die Verbindungsdetails, die in einem vorgeordneten Knoten, beispielsweise dem Datenbankquellenknoten, angegeben sind. Diese Op-

tion funktioniert nur, wenn alle vorgeordneten Knoten SQL-Pushback verwenden können. In diesem Fall müssen die Daten nicht aus der Datenbank verschoben werden, da die SQL alle vorgeordneten Knoten vollständig implementiert.

- **Daten in Verbindung verschieben**. Dient zum Verschieben der Daten in die hier angegebene Datenbank. Dadurch kann die Modellierung funktionieren, wenn sich die Daten in einer anderen IBM Data Warehouse-Datenbank, einer Datenbank eines anderen Anbieters oder in einer Flatfile befinden. Darüber hinaus werden die Daten in die hier angegebene Datenbank zurückverschoben, wenn die Daten extrahiert wurden, da ein Knoten kein SQL-Pushback durchgeführt hat. Klicken Sie auf die Schaltfläche **Bearbeiten**, um eine Verbindung zu suchen und auszuwählen.

Vorsicht: IBM Netezza Analytics und IBM Data Warehouse werden in der Regel mit sehr großen Datasets verwendet. Das Übertragen großer Datenmengen zwischen Datenbanken bzw. aus einer Datenbank und wieder zurück kann sehr zeitaufwendig sein und sollte nach Möglichkeit vermieden werden.

Modellname. Der Name des Modells. Die Name wird nur zu Ihrer Information angezeigt. Sie können ihn hier nicht ändern.

IBM Data WH-Entscheidungsbaum - Modelnuggets

Das Entscheidungsbaummodellnugget zeigt die Ausgabe des Modellierungsvorgangs an und ermöglicht es Ihnen außerdem, einige Optionen für das Scoren des Modells festzulegen.

Wenn Sie einen Stream ausführen, der ein Entscheidungsbaummodellnugget enthält, fügt der Knoten standardmäßig ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 12. Modellscoring-Feld für Entscheidungsbaum

Name des hinzugefügten Felds	**Bedeutung**
$I-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Wenn Sie die Option **Wahrscheinlichkeiten zugewiesener Klassen für Bewertungsdatensätze berechnen** entweder beim Modellierungsknoten oder beim Modellnugget auswählen und den Stream ausführen, wird ein weiteres Feld hinzugefügt.

Tabelle 13. Modellscoring-Feld für Entscheidungsbaum - zusätzlich

Name des hinzugefügten Felds	**Bedeutung**
$IP-*Zielname*	Konfidenzwert (von 0,0 bis 1,0) für die Vorhersage.

IBM Data WH-Entscheidungsbaumnugget - Registerkarte "Modell"

Auf der Registerkarte **Modell** wird der Prädiktoreinfluss des Entscheidungsbaummodells im grafischen Format angezeigt. Die Länge des Balkens gibt den Einfluss des Prädiktors an.

Anmerkung: Wenn Sie mit IBM Netezza Analytics Version 2.x oder niedriger arbeiten, wird der Inhalt des Entscheidungsbaummodells nur in Textformat angezeigt.

Für diese Versionen werden die folgenden Informationen angezeigt:

- Jede Zeile des Textes entspricht einem Knoten oder Blatt.
- Die Einrückung steht für die Baumebene.
- Für einen Knoten wird die Aufteilungsbedingung angezeigt.
- Für ein Blatt wird die zugewiesene Klassenbeschriftung angezeigt.

IBM Data WH-Entscheidungsbaumnugget - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie einige Optionen für das Scoren des Modells festlegen.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Berechnen Sie Wahrscheinlichkeiten zugewiesener Klassen für Bewertungsdatensätze. (Nur Entscheidungsbaum (Decision Tree) und Naive Bayes) Wenn diese Option ausgewählt ist (Standardeinstellung), enthalten die zusätzlichen Modellierungsfelder neben dem Vorhersagefeld auch ein Konfidenzfeld (also ein Wahrscheinlichkeitsfeld). Wenn Sie dieses Kontrollkästchen inaktivieren, wird nur das Vorhersagefeld erstellt.

Deterministische Eingabedaten verwenden. Wenn diese Option ausgewählt ist, stellt sie sicher, dass jeder Netezza-Algorithmus, der mehrere Durchläufe derselben Ansicht ausführt, denselben Satz von Daten für jeden Durchlauf verwendet. Wenn Sie dieses Kontrollkästchen löschen, um zu zeigen, dass nicht deterministische Daten verwendet werden, wird eine temporäre Tabelle erstellt, um die Datenausgabe, beispielsweise die von einem Partitionsknoten erstellte Ausgabe, für die Verarbeitung zu speichern. Diese Tabelle wird gelöscht, nachdem das Modell erstellt wurde.

IBM Data WH-Entscheidungsbaumnugget - Registerkarte "Viewer"

Die Registerkarte **Viewer** stellt den Baum eines Baummodells in derselben Weise dar, wie SPSS Modeler es für das Entscheidungsbaummodell tut.

Anmerkung: Wird das Modell mit IBM Netezza Analytics Version 2.x oder einer früheren Version erstellt, ist die Registerkarte **Viewer** leer.

IBM Data WH - K-Means-Modellnugget

Nuggets für K-Means-Modelle enthalten alle Informationen, die vom Clustermodell erfasst wurden, sowie Informationen zu den Trainingsdaten und dem Schätzvorgang.

Wenn Sie einen Stream ausführen, der ein Modellnugget vom Typ "K-Means" enthält, fügt der Knoten zwei neue Felder hinzu, die die Clusterzugehörigkeit und die Entfernung vom zugewiesenen Clusterzentrum für den betreffenden Datensatz enthalten. Das neue Feld mit dem Namen '$KM-K-Means' ist für die Clusterzugehörigkeit und das neue Feld mit dem Namen '$KMD-K-Means' ist für die Entfernung vom Clusterzentrum.

IBM Data WH-K-Means-Nugget - Registerkarte "Modell"

Die Registerkarte **Modell** enthält verschiedene grafische Ansichten, die Auswertungsstatistiken und Verteilungen für Felder von Clustern zeigen. Sie können die Daten aus dem Modell exportieren oder die Ansicht als Grafik exportieren.

Wenn Sie mit IBM Netezza Analytics Version 2.x oder niedriger arbeiten oder das Modell mit dem Distanzmaß Mahalanobis erstellen, wird der Inhalt des K-Means-Modells nur in Textformat angezeigt.

Für diese Versionen werden die folgenden Informationen angezeigt:

- **Auswertungsstatistik.** Für den kleinsten und den größten Cluster zeigt die Auswertungsstatistik die Anzahl der Datensätze an. Die Auswertungsstatistik zeigt zudem den Prozentsatz des Datasets an, der auf diese Cluster entfällt. In der Liste wird auch das Größenverhältnis zwischen dem größten und dem kleinsten Cluster angegeben.
- **Clusterübersicht.** In der Clusterübersicht werden die Cluster aufgelistet, die vom Algorithmus erstellt werden. Für jeden Cluster wird in der Tabelle die Anzahl der Datensätze in diesem Cluster angezeigt, sowie die mittlere Entfernung vom Clusterzentrum für diese Datensätze.

IBM Data WH-K-Means-Nugget - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie einige Optionen für das Scoren des Modells festlegen.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Distanzmaß. Die zur Messung des Abstands zwischen Datenpunkten zu verwendende Methode. Größere Abstände deuten auf größere Unähnlichkeiten hin. Folgende Optionen stehen zur Auswahl:

- **Euklidisch**. (Standardvorgabe) Der Abstand zwischen zwei Punkten wird anhand einer geradlinigen Verbindung zwischen den Punkten berechnet.
- **Manhattan**. Der Abstand zwischen zwei Punkten wird als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet.
- **Canberra**. Ähnlich wie die Manhattan-Distanz, jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.
- **Maximum**. Der Abstand zwischen zwei Punkten wird als der größte Unterschied in einer beliebigen Koordinatendimension berechnet.

Modellnugget für "Netezza-Bayes-Netz"

Das Modellnugget für ein Bayes-Netz bietet eine Möglichkeit zur Festlegung von Optionen zum Scoren des Modells.

Wenn Sie einen Stream ausführen, der ein Bayes-Netzmodellnugget enthält, fügt der Knoten ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 14. Modellscoring-Feld für Bayes-Netz

Name des hinzugefügten Felds	Bedeutung
$BN-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Sie können das zusätzliche Feld anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

Nugget für "Netezza-Bayes-Netz" - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modells festlegen.

Ziel. Wenn Sie ein Zielfeld scoren möchten, das vom aktuellen Ziel abweicht, wählen Sie hier das neue Ziel aus.

Datensatz-ID. Wenn kein Feld für die Datensatz-ID angegeben ist, wählen Sie hier das zu verwendende Feld aus.

Vorhersagetyp. Die Variation des zu verwendenden Vorhersagealgorithmus:

- **Beste (Nachbar mit höchster Korrelation).** (Standard) Verwendet den Nachbarknoten mit der höchsten Korrelation.
- **Nachbarn (gewichtete Vorhersage von Nachbarn).** Verwendet eine gewichtete Vorhersage aller Nachbarknoten.
- **NN-Nachbarn (Nicht-NULL-Nachbarn).** Wie bei der vorangegangenen Option, mit der Ausnahme, dass Knoten mit Nullwerten (also Knoten, die Attributen entsprechen, die fehlende Werte für die Instanz aufweisen, für die die Vorhersage berechnet wird) ignoriert werden.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

IBM Data WH - Modellnuggets für Naive Bayes

Das Modellnugget für Naive Bayes bietet eine Möglichkeit zur Festlegung von Optionen zum Scoren des Modells.

Wenn Sie einen Stream ausführen, der ein Naive Bayes-Modellnugget enthält, fügt der Knoten standardmäßig ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 15. Modellscoring-Feld für Naive Bayes - Standard

Name des hinzugefügten Felds	**Bedeutung**
$I-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Wenn Sie die Option **Wahrscheinlichkeiten zugewiesener Klassen für Bewertungsdatensätze berechnen** entweder beim Modellierungsknoten oder beim Modellnugget auswählen und den Stream ausführen, werden zwei weitere Felder hinzugefügt.

Tabelle 16. Modellscoring-Felder für Naive Bayes - Zusatz

Name des hinzugefügten Felds	**Bedeutung**
$IP-*Zielname*	Der Bayes-Zähler der Klasse für die betreffende Instanz (d. h. das Produkt aus der vorherigen Klassenwahrscheinlichkeit und den bedingten Wahrscheinlichkeiten der Instanzattributwerte).
$ILP-*Zielname*	Der natürliche Logarithmus des letzteren.

Sie können die zusätzlichen Felder anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

IBM Data WH - Nugget für Naive Bayes - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modells festlegen.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Berechnen Sie Wahrscheinlichkeiten zugewiesener Klassen für Bewertungsdatensätze. (Nur Entscheidungsbaum (Decision Tree) und Naive Bayes) Wenn diese Option ausgewählt ist (Standardeinstellung), enthalten die zusätzlichen Modellierungsfelder neben dem Vorhersagefeld auch ein Konfidenzfeld (also ein Wahrscheinlichkeitsfeld). Wenn Sie dieses Kontrollkästchen inaktivieren, wird nur das Vorhersagefeld erstellt.

Wahrscheinlichkeitsgenauigkeit für kleine oder extrem unausgewogene Datasets verbessern. Bei der Berechnung von Wahrscheinlichkeiten ruft diese Option das *m*-Schätzverfahren zur Vermeidung der Wahrscheinlichkeit null während der Schätzung auf. Diese Art der Wahrscheinlichkeitsschätzung mag langsamer sein, kann jedoch bei kleinen oder extrem unausgewogenen Datasets zu besseren Ergebnissen führen.

IBM Data WH - KNN-Modellnuggets

Das Modellnugget für KNN bietet eine Möglichkeit zur Festlegung von Optionen zum Scoren des Modells.

Wenn Sie einen Stream ausführen, der ein Modellnugget für KNN enthält, fügt der Knoten ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 17. Modellscoring-Feld für KNN	
Name des hinzugefügten Felds	**Bedeutung**
$KNN-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Sie können das zusätzliche Feld anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

IBM Data WH-KNN-Nugget - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modells festlegen.

Distanzmaß. Die zur Messung des Abstands zwischen Datenpunkten zu verwendende Methode. Größere Abstände deuten auf größere Unähnlichkeiten hin. Folgende Optionen stehen zur Auswahl:

- **Euklidisch**. (Standardvorgabe) Der Abstand zwischen zwei Punkten wird anhand einer geradlinigen Verbindung zwischen den Punkten berechnet.
- **Manhattan**. Der Abstand zwischen zwei Punkten wird als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet.
- **Canberra**. Ähnlich wie die Manhattan-Distanz, jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.
- **Maximum**. Der Abstand zwischen zwei Punkten wird als der größte Unterschied in einer beliebigen Koordinatendimension berechnet.

Anzahl der nächstgelegenen Nachbarn (k). Die Anzahl der nächsten Nachbarn für einen bestimmten Fall. Beachten Sie dabei, dass eine höhere Anzahl an Nachbarn nicht unbedingt ein präziseres Modell hervorbringt.

Der für *k* ausgewählte Wert legt die Balance zwischen der Vermeidung der Überanpassung (kann wichtig sein, insbesondere für "verrauschte" Daten) und der Auflösung (Ausgabe unterschiedlicher Vorhersagen für ähnliche Instanzen) fest. Normalerweise müssen Sie den Wert von *k* für jedes Dataset anpassen. Die typischen Werte liegen im Bereich von 1 bis zu mehreren Dutzend.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Messungen vor Berechnung des Abstands standardisieren. Bei Aktivierung dieser Option werden die Messungen für stetige Eingabefelder standardisiert, bevor die Abstandswerte berechnet werden.

Core-Sets zur Leistungsverbesserung bei großen Datasets verwenden. Bei Aktivierung dieser Option wird Stichprobennahme mit Core-Sets verwendet, um die Berechnung zu beschleunigen, wenn große Datasets involviert sind.

Modellnuggets für "Netezza - Divisives Clustering"

Das Modellnugget für divisives Clustering bietet eine Möglichkeit zur Festlegung von Optionen zum Scoren des Modells.

Wenn Sie einen Stream ausführen, der ein Modellnugget für divisives Clustering enthält, fügt der Knoten zwei neue Felder hinzu, deren Namen aus dem Zielnamen abgeleitet werden.

Tabelle 18. Modellscoring-Felder für divisives Clustering	
Name des hinzugefügten Felds	**Bedeutung**
$DC-*Zielname*	Kennung des Subclusters, dem der aktuelle Datensatz zugewiesen ist.

Tabelle 18. Modellscoring-Felder für divisives Clustering (Forts.)	
Name des hinzugefügten Felds	**Bedeutung**
$DCD-*Zielname*	Entfernung vom Zentrum des Subclusters für aktuellen Datensatz.

Sie können die zusätzlichen Felder anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

Nugget für "Netezza - Divisives Clustering" - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modells festlegen.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Distanzmaß. Die zur Messung des Abstands zwischen Datenpunkten zu verwendende Methode. Größere Abstände deuten auf größere Unähnlichkeiten hin. Folgende Optionen stehen zur Auswahl:

- **Euklidisch**. (Standardvorgabe) Der Abstand zwischen zwei Punkten wird anhand einer geradlinigen Verbindung zwischen den Punkten berechnet.
- **Manhattan**. Der Abstand zwischen zwei Punkten wird als Summe der absoluten Unterschiede zwischen ihren Koordinaten berechnet.
- **Canberra**. Ähnlich wie die Manhattan-Distanz, jedoch empfindlicher für Datenpunkte, die näher am Ursprung liegen.
- **Maximum**. Der Abstand zwischen zwei Punkten wird als der größte Unterschied in einer beliebigen Koordinatendimension berechnet.

Angewendete Hierarchieebene. Die auf die Daten anzuwendende Hierarchieebene.

IBM Data WH - PCA-Modellnuggets

Das Modellnugget für PCA bietet eine Möglichkeit zur Festlegung von Optionen zum Scoren des Modells.

Wenn Sie einen Stream ausführen, der ein Modellnugget für PCA enthält, fügt der Knoten standardmäßig ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 19. Modellscoring-Feld für PCA	
Name des hinzugefügten Felds	**Bedeutung**
$F-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Wenn Sie im Feld **Anzahl der ... Hauptkomponenten** im Modellierungsknoten oder im Modellnugget einen Wert größer 1 angeben und den Stream ausführen, fügt der Knoten ein neues Feld für jede Komponente hinzu. In diesem Fall erhalten die Feldnamen das Suffix *-n*. Dabei steht *n* für die Anzahl an Komponenten. Wenn Ihr Modell beispielsweise den Namen *pca* aufweist und drei Komponenten enthält, werden die neuen Felder wie folgt benannt: *$F-pca-1*, *$F-pca-2* und *$F-pca-3*.

Sie können die zusätzlichen Felder anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

Anmerkung: Beim Scoring des Modells kann ein Fehler auftreten, wenn Feldnamen in Kleinbuchstaben verwendet werden. Dies ist ein bekannter Fehler in Db2 Data Warehouse, der umgangen werden kann, indem vor dem Scoring alle Feldnamen in Großbuchstaben geändert werden.

IBM Data WH-PCA-Nugget - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modells festlegen.

Anzahl der in der Projektion zu verwendenden Hauptkomponenten. Die Anzahl an Hauptkomponenten, auf die das Dataset reduziert werden soll. Dieser Wert darf nicht die Anzahl an Attributen (Eingabefeldern) übersteigen.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Modellnuggets für "Netezza-Regressionsbaum"

Das Modellnugget für den Regressionsbaum bietet eine Möglichkeit zur Festlegung von Optionen zum Scoren des Modells.

Wenn Sie einen Stream ausführen, der ein Regressionsbaummodellnugget enthält, fügt der Knoten standardmäßig ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 20. Model-Scoring-Feld für Regressionsbaum

Name des hinzugefügten Felds	**Bedeutung**
$I-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Wenn Sie die Option **Geschätzte Varianz berechnen** entweder beim Modellierungsknoten oder beim Modellnugget auswählen und den Stream ausführen, wird ein weiteres Feld hinzugefügt.

Tabelle 21. Modellscoring-Feld für Regressionsbaum - zusätzlich

Name des hinzugefügten Felds	**Bedeutung**
$IV-*Zielname*	Geschätzte Varianzen des vorhergesagten Werts.

Sie können die zusätzlichen Felder anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

Nugget für "Netezza-Regressionsbaum" - Registerkarte "Modell"

Auf der Registerkarte **Modell** wird der Prädiktoreinfluss des Regressionsbaummodells im grafischen Format angezeigt. Die Länge des Balkens gibt den Einfluss des Prädiktors an.

Anmerkung: Wenn Sie mit IBM Netezza Analytics Version 2.x oder niedriger arbeiten, wird der Inhalt des Regressionsbaummodells nur in Textformat angezeigt.

Für diese Versionen werden die folgenden Informationen angezeigt:

- Jede Zeile des Textes entspricht einem Knoten oder Blatt.
- Die Einrückung steht für die Baumebene.
- Für einen Knoten wird die Aufteilungsbedingung angezeigt.
- Für ein Blatt wird die zugewiesene Klassenbeschriftung angezeigt.

Nugget für "Netezza-Regressionsbaum" - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modells festlegen.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivie-

ren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Geschätzte Varianz berechnen. Gibt an, ob die Varianz zugewiesener Klassen in die Ausgabe aufgenommen werden soll.

Netezza-Regressionsbaumnugget - Registerkarte "Viewer"

Die Registerkarte **Viewer** stellt den Baum eines Baummodells in derselben Weise dar, wie SPSS Modeler es für das Regressionsbaummodell tut.

Anmerkung: Wird das Modell mit IBM Netezza Analytics Version 2.x oder einer früheren Version erstellt, ist die Registerkarte **Viewer** leer.

IBM Data WH - Modelnuggets der linearen Regression

Das Modellnugget für die lineare Regression bietet eine Möglichkeit zur Festlegung von Optionen zum Scoren des Modells.

Wenn Sie einen Stream ausführen, der ein Modellnugget für die lineare Regression enthält, fügt der Knoten ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 22. Model-Scoring-Feld für lineare Regression

Name des hinzugefügten Felds	**Bedeutung**
$LR-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Nugget für lineare IBM Data WH-Regression - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen zum Scoren des Modells festlegen.

Eingabefelder einschließen. Wenn diese Option ausgewählt ist, werden alle ursprünglichen Eingabefelder nach unten weitergegeben, wobei das zusätzliche Modellierungsfeld (bzw. die zusätzlichen Modellierungsfelder) an die einzelnen Datenzeilen angehängt werden. Wenn Sie dieses Kontrollkästchen inaktivieren, werden nur das Feld "Datensatz-ID" und die zusätzlichen Modellierungsfelder weitergeleitet, sodass der Stream schneller abläuft.

Netezza-Zeitreihenmodellnugget

Das Modellnugget bietet Zugriff auf die Ausgabe der Zeitreihenmodellierung. Die Ausgabe besteht aus den folgenden Feldern.

Tabelle 23. Ausgabefelder des Zeitreihenmodells

Feld	**Beschreibung**
TSID	Die ID der Zeitreihe. Der Inhalt des Felds, das auf der Registerkarte "Felder" des Modellierungsknotens unter "Zeitreihen-IDs" angegeben wurde. Weitere Informationen finden Sie im Thema „Netezza-Zeitreihen - Feldoptionen” auf Seite 73.
ZEIT	Der Zeitraum innerhalb der aktuellen Zeitreihe.
HISTORY	Die historischen Datenwerte (diejenigen, die für die Vorhersage verwendet wurden). Dieses Feld wird nur eingeschlossen, wenn die Option **Historische Werte in Ergebnis einschließen** auf der Registerkarte "Einstellungen" des Modellnuggets ausgewählt wurde.

Tabelle 23. Ausgabefelder des Zeitreihenmodells (Forts.)	
Feld	**Beschreibung**
$TS-INTERPOLATED	Die interpolierten Werte, sofern verwendet. Dieses Feld wird nur eingeschlossen, wenn die Option **Interpolierte Werte in Ergebnis einschließen** auf der Registerkarte "Einstellungen" des Modellnuggets ausgewählt wurde. "Interpolation" ist eine Option auf der Registerkarte "Erstellungsoptionen" des Modellierungsknotens.
$TS-FORECAST	Die Vorhersagewerte für die Zeitreihe.

Fügen Sie zum nAnzeige der Modellausgabe einen Tabellenknoten (von der Registerkarte "Ausgabe" der Knotenpalette) zum Modellnugget hinzu und führen Sie den Tabellenknoten aus.

Nugget für "Netezza-Zeitreihe" - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie Optionen für die Anpassung der Modellausgabe angeben.

Modellname. Der Name des Modells laut Angabe auf der Registerkarte "Modelloptionen" des Modellierungsknotens.

Die anderen Optionen sind mit denen auf der Registerkarte "Modelloptionen" des Modellierungsknotens identisch.

Nugget für verallgemeinertes lineares IBM Data WH-Modell

Das Modellnugget bietet Zugriff auf die Ausgabe der Modellierung.

Wenn Sie einen Stream ausführen, der ein Modellnugget für verallgemeinerte lineare Modelle enthält, fügt der Knoten ein neues Feld hinzu, dessen Name aus dem Zielnamen abgeleitet wird.

Tabelle 24. Modellscoring-Feld für verallgemeinerte lineare Modelle	
Name des hinzugefügten Felds	**Bedeutung**
$GLM-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.

Auf der Registerkarte "Modell" werden verschiedene Statistiken zu dem Modell angezeigt.

Die Ausgabe besteht aus den folgenden Feldern.

Tabelle 25. Ausgabefelder für das verallgemeinerte lineare Modell	
Ausgabefeld	**Beschreibung**
Parameter	Die vom Modell verwendeten Parameter (d. h. die Prädiktorvariablen). Dies sind die numerischen und nominalen Spalten sowie der konstante Term (im Regressionsmodell).
Beta	Der Korrelationskoeffizient (d. h. die lineare Komponente des Modells).
Std-Fehler	Die Standardabweichung für Beta.
Test	Die Teststatistiken zum Evaluieren der Gültigkeit der Parameter.
p-Wert	Die Wahrscheinlichkeit für einen Fehler, wenn angenommen wird, dass der Parameter signifikant ist.
Residuenübersicht	
Residuentyp	Der Residuentyp der Vorhersage, für die Übersichtswerte angezeigt werden.
RSS	Der Wert des Residuums.

Tabelle 25. Ausgabefelder für das verallgemeinerte lineare Modell (Forts.)	
Ausgabefeld	**Beschreibung**
df	Die Freiheitsgrade des Residuums.
p-Wert	Die Wahrscheinlichkeit für einen Fehler. Ein hoher Wert signalisiert ein in geringem Maß passendes Modell. Ein niedriger Wert signalisiert ein in hohem Maß passendes Modell.

Nugget für verallgemeinertes lineares IBM Data WH-Modell - Registerkarte "Einstellungen"

Auf der Registerkarte "Einstellungen" können Sie die Modellausgabe anpassen.

Die Option ist mit der für "Scoring-Optionen" auf dem Modellierungsknoten angezeigten identisch. Weitere Informationen finden Sie im Thema „Verallgemeinertes lineares IBM Data WH-Modell - Scoring-Optionen der Modelloptionen” auf Seite 63.

IBM Data WH - TwoStep-Modellnugget

Wenn Sie einen Stream ausführen, der ein TwoStep-Modellnugget enthält, fügt der Knoten zwei neue Felder hinzu, die die Clusterzugehörigkeit und die Entfernung vom zugewiesenen Clusterzentrum für den betreffenden Datensatz enthalten. Das neue Feld mit dem Namen '$TS-Twostep' ist für die Clusterzugehörigkeit und das neue Feld mit dem Namen '$TSP-Twostep' ist für die Entfernung vom Clusterzentrum.

IBM Data WH-TwoStep-Nugget - Registerkarte "Modell"

Die Registerkarte **Modell** enthält verschiedene grafische Ansichten, die Auswertungsstatistiken und Verteilungen für Felder von Clustern zeigen. Sie können die Daten aus dem Modell exportieren oder die Ansicht als Grafik exportieren.

Kapitel 6. Datenbankmodellierung mit IBM Db2 for z/OS

IBM SPSS Modeler und IBM Db2 for z/OS

SPSS Modeler unterstützt die Integration in Db2 for z/OS, was Ihnen die Möglichkeit gibt, erweiterte Analysefunktionen auf Db2 for z/OS-Servern auszuführen. Sie können über die grafische Benutzerschnittstelle und die am Workflow orientierte Entwicklungsumgebung von SPSS Modeler auf diese Funktionen zugreifen. Auf diese Weise können Sie die Data-Mining-Algorithmen direkt in der Db2 for z/OS-Umgebung ausführen und dabei die Leistungsfähigkeit von IBM Db2 Analytics Accelerator nutzen.

SPSS Modeler unterstützt die Integration der folgenden Algorithmen aus Db2 for z/OS.

- Entscheidungsbäume
- K-Means
- Naive Bayes
- Regressionsbaum
- Two Step

Anforderungen für die Integration in IBM Db2 for z/OS

Die folgenden Bedingungen sind Voraussetzungen für die Modellierung innerhalb der Datenbank über Db2 for z/OS und IBM Db2 Analytics Accelerator for z/OS. Sie müssen möglicherweise Ihren Datenbankadministrator zu Rate ziehen, um sicherzustellen, dass diese Bedingungen erfüllt sind. Detaillierte Informationen zu den Anforderungen, einschließlich unterstützter Versionen, finden Sie in den Software Product Compatibility Reports.

- IBM SPSS Modeler im lokalen Modus oder im Rahmen einer SPSS Modeler Server-Installation unter Windows oder UNIX
- Db2 for z/OS zusammen mit Db2 Analytics Accelerator for z/OS
- IBM SPSS Data Access Pack
- Auf dem Server, auf dem SPSS Modeler Server aktiv ist, eines der folgenden Systeme:
  - IBM Db2 Data Server Driver for ODBC and CLI
  - Eine beliebige Version von Db2 for Linux®, UNIX, and Windows mit einer ODBC-Datenquelle, die für Db2 for z/OS konfiguriert ist
- Lizenz für Db2 Connect for System z
- Aktivierte SQL-Generierung und -Optimierung in SPSS Modeler
- Für das datenbankinterne Mining von Db2 z/OS sind INZA-Unterstützung sowie reine Akzeleratortabellen (AOT - Accelerator-only Tables) oder beschleunigte Tabellen erforderlich. IDAA INZA wurde in IDAA 5.1 eingeführt. Dies bedeutet, dass die Db2 z/OS-Knoten für das datenbankinterne Mining nicht mit vorherigen Versionen von IDAA verwendet werden können.

  Wenn Sie in Modeler einen IDAA-fähigen DSN verwenden, werden in der Liste der Tabellen, die über diesen DSN im Datenquellenknoten zurückgegeben werden, nur reine Akzeleratortabellen oder beschleunigte Tabellen angezeigt.

Aktivieren der Integration in IBM Db2 Analytics Accelerator for z/OS

Das Aktivieren der Integration in Db2 Analytics Accelerator for z/OS umfasst die folgenden Schritte:

- Konfigurieren von Db2 for z/OS und Db2 Analytics Accelerator for z/OS
- Erstellen einer ODBC-Datenquelle
- Aktivieren der Integration von IBM Db2 for z/OS in IBM SPSS Modeler
- Aktivieren der SQL-Generierung und -Optimierung in SPSS Modeler
- Aktivieren von IBM SPSS Modeler Server Scoring Adapter für Db2 for z/OS
- Konfigurieren von DSN mit IBM Db2-Client in IBM SPSS Modeler

Konfigurieren von IBM Db2 for z/OS und IBM Analytics Accelerator for z/OS

Die Vorgehensweise beim Konfigurieren von Db2 for z/OS und Analytics Accelerator for z/OS wird auf der folgenden Website beschrieben:

Db2 Analytics Accelerator for z/OS.

Erstellen einer ODBC-Quelle für IBM Db2 for z/OS und IBM Db2 Analytics Accelerator

Informationen zum Aktivieren einer Verbindung zwischen Db2 for z/OS und IBM Db2 Analytics Accelerator finden Sie auf den folgenden Websites:

- Für Version 4: Db2 Analytics Accelerator for z/OS 4.1.0
- Für Version 3: Db2 Analytics Accelerator for z/OS 3.1.0
- Aktivieren der Abfragebeschleunigung mit IBM Db2 Analytics Accelerator für ODBC- und JDBC-Anwendungen, ohne die Anwendungen zu ändern
- SQL-Fehler von ODBC-Treiber beim Ausführen einer Abfrage in Db2 Analytics Accelerator for z/OS

Aktivieren der Integration von IBM Db2 for z/OS in IBM SPSS Modeler

Mit den folgenden Schritten können Sie die Integration von Db2 for z/OS in SPSS Modeler ermöglichen:

1. Öffnen Sie die Datei `odbc-db2-accelerator-names.cfg` im SPSS Modeler-Verzeichnis `config`.

   Wenn die Datei nicht vorhanden ist, müssen Sie sie erstellen.
2. Fügen Sie die Namen aller Datenquellen und aller Akzeleratoren hinzu. Beispiel:

   ```
   dsn1, akzeleratorname1
   dsn2, akzeleratorname2
   ```
3. Die Standard-CCSID für reine Akzeleratortabellen (AOT) ist Unicode. Wenn Sie diese Einstellung außer Kraft setzen wollen, ändern Sie die Einträge, indem Sie den Akzeleratornamen Codierungszeichenfolgen hinzufügen. Beispiel:

   ```
   dsn1, akzeleratorname1, EBCDIC
   dsn2, akzeleratorname2, UNICODE
   ```
4. Speichern Sie die Datei `odbc-db2-accelerator-names.cfg` und schließen Sie sie. Öffnen Sie dann die Datei `odbc-db2-custom-properties.cfg` in demselben Verzeichnis.
5. SPSS Modeler legt die IDAA-Register über SQL fest. Sie können diese Einträge, falls erforderlich, außer Kraft setzen, indem Sie das SQL in die erforderlichen Werte ändern. Beispiel:

   ```
   current_query_sql_acc, "SET CURRENT QUERY ACCELERATION = ELIGIBLE"
   current_get_archive_acc, "SET CURRENT GET_ACCEL_ARCHIVE = NO"
   ```
6. Standardmäßig verwendet SPSS Modeler SQL zum Erstellen temporärer Tabellen für einen Datenbankcache. Sie können diese Einstellung, falls erforderlich, außer Kraft setzen, indem Sie den erwarteten Datenbanknamen angeben. Beispiel:

```
[OSZ]
table_create_temp_sql_acc, 'CREATE TABLE <Tabellenname> <(Tabellenspalten)> IN DATABASE
NAME_OF_DATABASE_FOR_AOT'
```

7. Standardmäßig geht SPSS Modeler davon aus, dass in einem ODBC-Quellenknoten geschriebene SQL-Abfragen nicht wiederholt werden können. Dies bedeutet, dass angenommen wird, dass die Abfrage, wenn sie mehrmals ausgeführt wird, unterschiedliche Ergebnisse zurückgibt. In einigen Szenarios kann dies jedoch verhindern, dass Modeler SQL für nachgeordnete Knoten generiert. In diesem Fall kann die Einstellung außer Kraft gesetzt werden, indem der entsprechende Wert in Y geändert wird. Beispiel:

```
assume_custom_sql_replayable, Y
```

8. Klicken Sie im SPSS Modeler-Hauptmenü auf **Tools** > **Optionen** > **Hilfsanwendungen**.
9. Klicken Sie auf die Registerkarte **IBM Db2 for z/OS**.
10. Wählen Sie **IBM Db2 for z/OS-Data-Mining-Integration aktivieren** aus und klicken Sie dann auf **OK**.

Anmerkung: IDAA- und Nicht-IDAA-Tabellen können in Modeler nicht gleichzeitig angezeigt werden.

Aktivieren der SQL-Generierung und -Optimierung

Da mit hoher Wahrscheinlichkeit mit sehr großen Datasets gearbeitet wird, sollten Sie aus Leistungsgründen die Optionen zur SQL-Generierung und -Optimierung in IBM SPSS Modeler aktivieren.

Führen Sie die folgenden Schritte aus, um SPSS Modeler zu konfigurieren:

1. Wählen Sie in den IBM SPSS Modeler-Menüs **Tools** > **Streameigenschaften** > **Optionen** aus.
2. Klicken Sie im Navigationsbereich auf die Option **Optimierung**.
3. Überzeugen Sie sich, dass die Option **SQL generieren** aktiviert ist. Diese Einstellung ist für die Datenbankmodellierung erforderlich.
4. Wählen Sie **SQL-Generierung optimieren** und **Andere Ausführung optimieren** aus. (Diese Einstellungen sind nicht unbedingt erforderlich, werden aber zur Leistungsoptimierung empfohlen.)

Konfigurieren von DSN mit IBM Db2-Client in IBM SPSS Modeler

Führen Sie die folgenden Schritte aus, um bei Bedarf einen Datenquellennamen (DSN) mit dem Db2-Client für Db2 in SPSS Modeler zu konfigurieren:

1. Falls er nicht bereits installiert ist, installieren Sie den Db2-Client unter dem Betriebssystem, unter dem Modeler Server installiert ist.
2. Verwenden Sie den Befehl **`db2 catalog`**, um die Datenbank zu katalogisieren und der Datei `db2cli.ini` im Db2-Client eine neue Datenquelle hinzuzufügen. Achten Sie darauf, auf den definierten Datenbankalias zu verweisen.
3. Konfigurieren Sie den Datenzugriff. Detaillierte Schritte sind in der Dokumentation zu Modeler verfügbar.

   Weitere Informationen finden Sie in **Architektur- und Hardwareempfehlungen** > **Datenzugriff** im *Modeler Server Verwaltungs- und Leistungshandbuch* (ModelerServerAdminPerformance.pdf).
4. Erstellen Sie eine neue ODBC-Datenquelle in `odbc.ini`, indem Sie auf den in Schritt 2 definierten Datenbankalias verweisen.
5. Für Linux- oder UNIX-Benutzer gilt Folgendes:
   a. Stellen Sie sicher, dass die Treiberbibliothek `libdb2o.so` verwendet wird (statt `libdb2.so`), und stellen Sie sicher, dass `'DriverUnicodeType=1'` für die neue Datenquelle definiert ist.
   b. Stellen Sie in der IBM SPSS Data Access Pack-Installation sicher, dass der Bibliothekspfad des Db2-Clients der Datei `odbc.sh` hinzugefügt wird.
   c. Stellen Sie sicher, dass Modeler Server eine ODBC-Treiber-Wrapperbibliothek mit UTF-16-Codierung verwendet (sie heißt `'libspssodbc_datadirect_utf16.so'`).

6. Stellen Sie sicher, dass der Benutzer, der eine Verbindung zu Db2 herstellt, die erforderlichen Berechtigungen zum Ausführen der folgenden Abfrage hat:

```
SELECT ACCELERATORNAME FROM SYSACCEL.SYSACCELERATORS
```

Erstellen von Modellen mit IBM Db2 for z/OS

Zu jedem der unterstützten Algorithmen gehört ein Modellierungsknoten. Über die Registerkarte "Datenbankmodellierung" in der Knotenpalette können Sie auf die Db2 for z/OS-Modellierungsknoten zugreifen.

Erläuterung der Daten

Felder in der Datenquelle können, je nach Modellierungsknoten, Variablen verschiedener Datentypen enthalten. In SPSS Modeler werden Datentypen als *Messniveaus* bezeichnet. Auf der Registerkarte "Felder" des Modellierungsknotens werden Symbole verwendet, die die zulässigen Messniveautypen für die Eingabe- und Zielfelder angeben.

Zielfeld. Das Zielfeld ist das Feld, dessen Wert Sie vorherzusagen versuchen. Wenn ein Ziel angegeben werden kann, kann nur eines der Quellendatenfelder als Zielfeld ausgewählt werden.

Datensatz-ID-Feld. Gibt das Feld an, über das jeder Fall eindeutig identifiziert wird. Dies kann beispielsweise ein ID-Feld sein, wie *CustomerID*. Wenn die Quellendaten kein ID-Feld enthalten, können Sie dieses Feld mithilfe eines Ableitungsknotens erstellen, wie in der folgenden Prozedur gezeigt.

1. Wählen Sie den Quellenknoten aus.
2. Doppelklicken Sie auf der Registerkarte "Feldoperationen" auf den Ableitungsknoten.
3. Öffnen Sie den Ableitungsknoten, indem Sie im Erstellungsbereich auf das zugehörige Symbol doppelklicken.
4. Geben Sie im **Ableitungsfeld** (beispielsweise) `ID` ein.
5. Geben Sie `@INDEX` im Feld **Formel** ein und klicken Sie auf **OK**.
6. Verbinden Sie den Ableitungsknoten mit dem Rest des Streams.

Umgang mit Nullwerten

Wenn die Eingabedaten Nullwerte enthalten, kann die Verwendung einiger Db2 for z/OS-Knoten zu Fehlernachrichten oder Streams mit sehr langer Ausführungsdauer führen. Deshalb wird empfohlen, Datensätze mit Nullwerten zu entfernen. Verwenden Sie die folgende Methode.

1. Verbinden Sie einen Auswahlknoten mit dem Quellenknoten.
2. Setzen Sie die Option **Modus** des Auswahlknotens auf **Verwerfen**.
3. Geben Sie Folgendes in das Feld **Bedingung** ein:

```
@NULL(Feld1) [oder @NULL(Feld2)[... oder @NULL(FeldN]])
```

   Achten Sie darauf, alle Eingabefelder mit aufzunehmen.
4. Verbinden Sie den Auswahlknoten mit dem Rest des Streams.

Modellausgabe

Es ist möglich, dass ein Stream, der einen Db2 for z/OS-Modellierungsknoten enthält, bei jeder Ausführung etwas andere Ergebnisse ausgibt. Der Grund hierfür ist, dass die Reihenfolge, in der der Knoten die Quellendaten liest, nicht immer gleich ist, da die Daten vor der Modellerstellung in temporäre Tabellen eingelesen werden. Die durch diesen Effekt erzeugten Unterschiede sind jedoch vernachlässigbar.

Allgemeine Kommentare

- In SPSS Collaboration and Deployment Services können Scoring-Konfigurationen nicht mithilfe von Streams erstellt werden, die Db2 for z/OS-Modellierungsknoten enthalten.
- PMML-Export bzw. -Import ist für Modelle nicht möglich, die von den Db2 for z/OS-Knoten erstellt wurden.

IBM Db2 for z/OS-Modelle - Feldoptionen

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

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

Ziel. Wählen Sie ein Feld als Ziel für die Vorhersage aus. Beachten Sie bei verallgemeinerten linearen Modellen auch das Feld **Tests** in dieser Anzeige.

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

IBM Db2 for z/OS-Modelle - Serveroptionen

Auf der Registerkarte "Server" geben Sie das Db2 for z/OS-System an, in dem das Modell erstellt werden soll.

- **Vorgeordnete Verbindung verwenden.** (Standardeinstellung) Verwendet die Verbindungsdetails, die in einem vorgeordneten Knoten, beispielsweise dem Datenbankquellenknoten, angegeben sind. *Hinweis:* Diese Option funktioniert nur, wenn alle vorgeordneten Knoten SQL-Pushback verwenden können. In diesem Fall müssen die Daten nicht aus der Datenbank verschoben werden, da die SQL alle vorgeordneten Knoten vollständig implementiert.
- **Daten in Verbindung verschieben**. Dient zum Verschieben der Daten in die hier angegebene Datenbank. Dadurch kann die Modellierung funktionieren, wenn sich die Daten in einer anderen IBM Datenbank, einer Datenbank eines anderen Anbieters oder in einer Flatfile befinden. Darüber hinaus werden die Daten in die hier angegebene Datenbank zurückverschoben, wenn die Daten extrahiert wurden, da ein Knoten kein SQL-Pushback durchgeführt hat. Klicken Sie auf die Schaltfläche **Bearbeiten**, um eine Verbindung zu suchen und auszuwählen.

Anmerkung: Der Name der ODBC-Datenquelle wird in jeden SPSS Modeler-Stream eingebettet. Wenn ein auf einem Host erstellter Stream auf einem anderen Host ausgeführt wird, muss der Name der Datenquelle auf beiden Hosts identisch sein. Alternativ kann für jeden Quellen- oder Modellierungsknoten auf der Registerkarte "Server" eine andere Datenquelle ausgewählt werden.

IBM Db2 for z/OS-Modelle - Modelloptionen

Auf der Registerkarte "Modelloptionen" können Sie bestimmen, ob Sie einen Namen für das Modell angeben oder automatisch erstellen lassen möchten.

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

Vorhandenes ersetzen, wenn Name bereits verwendet wird. Wenn Sie dieses Kontrollkästchen auswählen, wird ein vorhandenes Modell mit demselben Namen ggf. überschrieben.

IBM Db2 for z/OS-Modelle - K-Means

Der K-Means-Knoten implementiert den *k*-Means-Algorithmus, der eine Methode der Clusteranalyse bietet. Mit diesem Knoten können Sie ein Clustering der Datasets in einzelne Gruppen vornehmen.

Bei dem Algorithmus handelt es sich um einen distanzbasierten Clusteralgorithmus, der auf einer Distanzmetrik (Funktion) zur Messung der Ähnlichkeit zwischen Datenpunkten beruht. Die Datenpunkte werden dem nächsten Cluster gemäß der verwendeten Distanzmetrik zugewiesen.

Bei diesem Algorithmus werden mehrere Iterationen desselben Grundverfahren durchgeführt. Dabei wird jede Trainingsinstanz dem nächstgelegenen Cluster zugewiesen (in Bezug auf die angegebene Distanzfunktion, angewendet auf Instanz und Clusterzentrum). Anschließend werden alle Clusterzentren als Attribut-Mittelwertvektoren der Instanzen neu berechnet, die den jeweiligen Clustern zugewiesen wurden.

IBM Db2 for z/OS-Modelle - K-Means-Feldoptionen

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

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

IBM Db2 for z/OS-Modelle - K-Means-Erstellungsoptionen

Durch Festlegen der Erstellungsoptionen können Sie die Erstellung des Modells Ihren Anforderungen entsprechend anpassen.

Wenn Sie ein Modell mit den Standardoptionen erstellen wollen, klicken Sie auf **Ausführen**.

Distanzmaß. Dieser Parameter definiert die Methode für die Messung der Distanz zwischen Datenpunkten. Größere Distanzen geben größere Unähnlichkeiten an. Wählen Sie eine der folgenden Optionen aus:

- **Euklidisch.** Das euklidische Maß ist die geradlinige Distanz zwischen zwei Datenpunkten.
- **Euklidisch normalisiert.** Das euklidisch normalisierte Maß ähnelt dem euklidischen Maß, wird jedoch durch das Quadrat der Standardabweichung normalisiert. Im Gegensatz zum euklidischen Maß ist das euklidisch normalisierte Maß zudem skaleninvariant.

Anzahl der Cluster. Dieser Parameter definiert die Anzahl der zu erstellenden Cluster.

Maximale Anzahl an Iterationen. Bei diesem Algorithmus werden mehrere Iterationen desselben Prozesses durchgeführt. Dieser Parameter definiert die Anzahl von Iterationen, nach denen das Modelltraining beendet wird.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Wählen Sie eine der folgenden Optionen aus:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.
- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

Ergebnisse reproduzieren. Aktivieren Sie dieses Kontrollkästchen, wenn Sie einen Startwert für Zufallszahlen festlegen wollen, um Analysen zu replizieren. Sie können eine ganze Zahl angeben oder durch Klicken auf **Generieren** eine pseudozufällige ganze Zahl erstellen.

IBM Db2 for z/OS-Modelle - Naive Bayes

Naive Bayes ist ein bekannter Algorithmus für Klassifizierungsprobleme. Das Modell wird als "naïve" bezeichnet, weil es alle vorgeschlagenen Vorhersagevariablen als voneinander unabhängig behandelt. Naive Bayes ist ein schneller, skalierbarer Algorithmus, der für Kombinationen von Attributen und für das Zielattribut bedingte Wahrscheinlichkeiten berechnet. Aus den Trainingsdaten wird eine unabhängige Wahrscheinlichkeit ermittelt. Diese liefert die Wahrscheinlichkeit jeder Zielklasse anhand des Vorkommens der einzelnen Wertekategorien aus jeder einzelnen Eingabevariablen.

IBM Db2 for z/OS-Modelle - Entscheidungsbäume

Ein Entscheidungsbaum ist eine hierarchische Struktur, die ein Klassifizierungsmodell darstellt. Mit einem Entscheidungsbaummodell können Sie ein Klassifizierungssystem entwickeln, die zukünftige Beobachtungen auf der Grundlage eines Satzes von Trainingsdaten vorhersagen oder klassifizieren. Die Klassifizierung hat die Form einer Baumstruktur, in der die Verzweigungen Teilungspunkte innerhalb der Klassifizierung darstellen. Die Teilungspunkte teilen die Daten rekursiv in Untergruppen auf, bis ein Endpunkt erreicht wird. Die Baumknoten an den Endpunkten werden als *Blätter* bezeichnet. Jedes Blatt weist den Mitgliedern seiner Untergruppe oder Klasse eine Beschriftung zu, die als *Klassenbeschriftung* bezeichnet wird.

IBM Db2 for z/OS-Modelle - Erstellungsoptionen für Entscheidungsbäume

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

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

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

Felder. Verwenden Sie die Pfeiltasten, um den verschiedenen Rollenfeldern rechts in der Anzeige Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

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

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

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird. Die Werte in diesem Feld müssten für jeden Datensatz eindeutig sein (z. B. Kundennummern).

Instanzgewichtung. Indem Sie hier ein Feld angeben, können Sie anstelle der Klassengewichtungen oder zusätzlich zu den KIassengewichtungen (eine Gewichtung pro Kategorie für das Zielfeld) Instanzgewichtungen (eine Gewichtung pro Zeile an Eingabedaten) verwenden. Das hier angegebene Feld muss eine numerische Gewichtung für jede Zeile an Eingabedaten enthalten.

Prädiktoren (Eingaben). Wählen Sie das/die Eingabefeld(er) aus. Dies ist so, als würden Sie in einem Typknoten für die Rolle eines Felds den Wert *Eingabe* festlegen.

IBM Db2 for z/OS-Modelle - Erstellungsoptionen für Entscheidungsbäume

Die folgenden Erstellungsoptionen sind für den Baumaufbau verfügbar:

Erweiterungsmaß. Diese Optionen steuern, wie die Baumerweiterung gemessen wird.

- **Unreinheitsmaß.** Dieses Maß ermittelt die beste Position für eine Baumteilung. Es handelt sich um ein Maß für die Variabilität in einer Untergruppe oder einem Datensegment. Ein niedriges Unreinheitsmaß gibt eine Gruppe an, in der die meisten Mitglieder ähnliche Werte für das Kriterium oder Zielfeld aufweisen.

  Die Maße **Entropie** und **Gini** werden unterstützt. Diese Maße basieren auf Wahrscheinlichkeiten der Zugehörigkeit zu einer Kategorie einer Verzweigung.
- **Maximale Baumtiefe.** Die maximale Anzahl der Ebenen, auf die ein Baum unterhalb des Stammknotens erweitert werden kann, d. h. die Anzahl der rekursiven Teilungen einer Stichprobe. Der Standardwert dieser Eigenschaft ist 10. Der Maximalwert, den Sie für diese Eigenschaft festlegen können, ist 62.

  Anmerkung: Wenn der Viewer im Modellnugget das Modell in Texform darstellt, werden maximal 12 Ebenen des Baums angezeigt.

Aufteilungskriterien. Diese Optionen steuern, wann die Aufteilung des Baums aufhört.

- **Mindestverbesserung für Aufteilungen.** Der Mindestwert der Unreinheitsreduzierung, bevor eine neue Aufteilung des Baums erfolgt. Das Ziel der Baumerstellung besteht darin, Untergruppen mit ähnlichen Ausgabewerten zu erstellen, um die Unreinheit in den einzelnen Knoten zu minimieren. Wenn die beste für eine Verzweigung mögliche Aufteilung die Unreinheit um weniger als den durch die Aufteilungskriterien vorgegebenen Betrag reduziert, wird die Verzweigung nicht aufgeteilt.
- **Mindestanzahl der Instanzen für eine Aufteilung.** Die Mindestanzahl von Datensätzen, die aufgeteilt werden können. Wenn weniger nicht aufgeteilte Datensätze verbleiben, werden keine weiteren Aufteilungen durchgeführt. Sie können dieses Feld verwenden, um die Erstellung kleiner Untergruppen im Baum zu verhindern.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Wählen Sie eine der folgenden Optionen aus:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.
- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

IBM Db2 for z/OS-Modelle - Entscheidungsbaumknoten - Klassengewichtungen

Hier können Sie den einzelnen Klassen Gewichtungen zuweisen. Standardmäßig wird allen Klassen der Wert 1 zugewiesen, sodass sie gleich gewichtet sind. Durch die Angabe von unterschiedlichen numerischen Gewichtungen für verschiedene Klassenbeschriftungen weisen Sie den Algorithmus an, die Trainingssets bestimmter Klassen entsprechend zu gewichten.

Doppelklicken Sie zum Ändern einer Gewichtung in die Spalte **Gewichtung** und nehmen Sie die gewünschten Änderungen vor.

Wert. Die Menge der Klassenbeschriftungen, abgeleitet aus den möglichen Werten des Zielfelds.

Gewichtung. Die Gewichtung, die einer bestimmten Klasse zugewiesen werden soll. Durch Zuweisen einer höheren Gewichtung zu einer Klasse reagiert das Modell im Vergleich zu den anderen Klassen sensibler auf diese Klasse.

Klassengewichtungen können in Verbindung mit Instanzgewichtungen verwendet werden.

IBM Db2 for z/OS-Modelle - Entscheidungsbaumknoten - Baumreduzierung

Sie können die Reduzierungsoptionen verwenden, um Reduzierungskriterien für den Entscheidungsbaum festzulegen. Ziel der Reduzierung ist es, das Risiko der übermäßigen Anpassung zu verringern, indem zu stark erweiterte Untergruppen entfernt werden, welche die erwartete Genauigkeit für neue Daten nicht verbessern.

Reduzierungsmaß. Das standardmäßige Reduzierungsmaß, **Genauigkeit**, gewährleistet, dass die geschätzte Genauigkeit des Modells nach der Entfernung eines Blatts aus dem Baum innerhalb akzeptabler Grenzen bleibt. Verwenden Sie das Alternativmaß, **Gewichtete Genauigkeit**, wenn Sie die Klassengewichtungen in die Reduzierung mit einbeziehen möchten.

Daten für die Reduzierung. Sie können einen Teil oder alle Trainingsdaten verwenden, um die erwartete Genauigkeit der neuen Daten abzuschätzen. Alternativ können Sie zu diesem Zweck ein separates Dataset für die Reduzierung aus einer festgelegten Tabelle verwenden.

- **Alle Trainingsdaten verwenden.** Diese (standardmäßige) Option verwendet alle Trainingsdaten, um die Modellgenauigkeit zu schätzen.
- **% der Trainingsdaten für die Reduzierung verwenden.** Teilen Sie mithilfe dieser Option die Daten in zwei Gruppen (eine für das Training und eine für die Reduzierung) unter Verwendung des hier angegebenen Prozentsatzes für die Reduzierungsdaten.
- Wählen Sie das Feld **Ergebnisse replizieren** aus, wenn Sie einen Zufallsstartwert angeben möchten, um sicherzustellen, dass die Daten bei jeder Ausführung des Streams auf dieselbe Weise partitioniert werden. Sie können entweder eine ganze Zahl im Feld **Für Reduzierung verwendeter Startwert** angeben oder auf **Generieren** klicken, wodurch eine pseudozufällige ganze Zahl erstellt wird.
- **Daten aus einer vorhandenen Tabelle verwenden.** Geben Sie den Tabellennamen eines separaten Datasets für die Reduzierung an, anhand dessen die Modellgenauigkeit geschätzt wird. Diese Vorgehensweise wird als zuverlässiger betrachtet als die Nutzung von Trainingsdaten.

IBM Db2 for z/OS-Modelle - Regressionsbaum

Ein Regressionsbaum ist ein baumbasierter Algorithmus, der eine Stichprobe von Fällen wiederholt aufteilt, um anhand von Werten eines numerischen Ausgabefeldes gleichartige Subsets abzuleiten. Ebenso wie Entscheidungsbäume zerlegen auch Regressionsbäume die Daten in Subsets, in denen die Blätter des Baums hinreichend kleinen bzw. hinreichend einheitlichen Subsets entsprechen. Aufteilungen werden ausgewählt, um die Streuung der Zielattributwerte zu verringern, sodass sie angemessen gut durch ihren Mittelwert an Blättern vorhergesagt werden können.

IBM Db2 for z/OS-Modelle - Erstellungsoptionen für Regressionsbaum - Baumerweiterung

Sie können Erstellungsoptionen für die Baumerweiterung und die Baumreduzierung festlegen.

Die folgenden Erstellungsoptionen sind für den Baumaufbau verfügbar:

Maximale Baumtiefe. Die maximale Anzahl der Ebenen, auf die ein Baum unterhalb des Stammknotens erweitert werden kann, d. h. die Anzahl der rekursiven Teilungen einer Stichprobe. Der Standardwert liegt bei 62. Dies ist die maximale Baumtiefe für Modellierungszwecke.

Anmerkung: Wenn der Viewer im Modellnugget das Modell in Texform darstellt, werden maximal 12 Ebenen des Baums angezeigt.

Aufteilungskriterien. Diese Optionen steuern, wann die Aufteilung des Baums aufhört.

- **Maß zur Aufteilungsevaluierung.** Dieses Evaluierungsmaß für die Klasse ermittelt die beste Position für eine Baumteilung.

  Anmerkung: Derzeit ist "Varianz" die einzig mögliche Option.
- **Mindestverbesserung für Aufteilungen.** Der Mindestwert der Unreinheitsreduzierung, bevor eine neue Aufteilung des Baums erfolgt. Das Ziel der Baumerstellung besteht darin, Untergruppen mit ähnlichen Ausgabewerten zu erstellen, um die Unreinheit in den einzelnen Knoten zu minimieren. Wenn die beste für eine Verzweigung mögliche Aufteilung die Unreinheit um weniger als den durch die Aufteilungskriterien vorgegebenen Betrag reduziert, wird die Verzweigung nicht aufgeteilt.
- **Mindestanzahl der Instanzen für eine Aufteilung.** Die Mindestanzahl von Datensätzen, die aufgeteilt werden können. Wenn weniger nicht aufgeteilte Datensätze verbleiben, werden keine weiteren Aufteilungen durchgeführt. Sie können dieses Feld verwenden, um die Erstellung kleiner Untergruppen im Baum zu verhindern.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Wählen Sie eine der folgenden Optionen aus:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.
- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

IBM Db2 for z/OS-Modelle - Erstellungsoptionen für Regressionsbaum - Baumreduzierung

Sie können die Reduzierungsoptionen verwenden, um Reduzierungskriterien für den Regressionsbaum festzulegen. Ziel der Reduzierung ist es, das Risiko der übermäßigen Anpassung zu verringern, indem zu stark erweiterte Untergruppen entfernt werden, welche die erwartete Genauigkeit für neue Daten nicht verbessern.

Reduzierungsmaß. Das Reduzierungsmaß gewährleistet, dass die geschätzte Genauigkeit des Modells nach der Entfernung eines Blatts aus dem Baum innerhalb akzeptabler Grenzen bleibt. Sie können eines der folgenden Maße auswählen.

- **mse.** Mittlerer quadratischer Fehler - (Standard) misst, wie eng eine angepasste Linie an den Datenpunkten liegt.
- **r2.** R-Quadrat - misst den Anteil an Variation in der abhängigen Variablen, der durch das Regressionsmodell erklärt wird.
- **Pearson.** Korrelationskoeffizienten nach Pearson - misst die Stärke der Beziehung zwischen linear abhängigen Variablen, die normal verteilt sind.
- **Spearman.** Korrelationskoeffizient nach Spearman - erkennt nicht lineare Beziehungen, die laut der Korrelation nach Pearson schwach erscheinen, jedoch möglicherweise stark sind.

Daten für die Reduzierung. Sie können einen Teil oder alle Trainingsdaten verwenden, um die erwartete Genauigkeit der neuen Daten abzuschätzen. Alternativ können Sie zu diesem Zweck ein separates Dataset für die Reduzierung aus einer festgelegten Tabelle verwenden.

- **Alle Trainingsdaten verwenden.** Diese (standardmäßige) Option verwendet alle Trainingsdaten, um die Modellgenauigkeit zu schätzen.

- **% der Trainingsdaten für die Reduzierung verwenden.** Teilen Sie mithilfe dieser Option die Daten in zwei Gruppen (eine für das Training und eine für die Reduzierung) unter Verwendung des hier angegebenen Prozentsatzes für die Reduzierungsdaten.

  Wählen Sie das Feld **Ergebnisse replizieren** aus, wenn Sie einen Zufallsstartwert angeben möchten, um sicherzustellen, dass die Daten bei jeder Ausführung des Streams auf dieselbe Weise partitioniert werden. Sie können entweder eine ganze Zahl im Feld **Für Reduzierung verwendeter Startwert** angeben oder auf **Generieren** klicken, wodurch eine pseudozufällige ganze Zahl erstellt wird.
- **Daten aus einer vorhandenen Tabelle verwenden.** Geben Sie den Tabellennamen eines separaten Datasets für die Reduzierung an, anhand dessen die Modellgenauigkeit geschätzt wird. Diese Vorgehensweise wird als zuverlässiger betrachtet als die Nutzung von Trainingsdaten.

IBM Db2 for z/OS-Modelle - TwoStep

Der TwoStep-Knoten implementiert den Algorithmus TwoStep, der eine Methode zum Bilden von Datenclustern für große Datasets bereitstellt.

Mit diesem Knoten können sie einen Datencluster unter Berücksichtigung der verfügbaren Ressourcen wie Speicher und Zeitvorgaben bilden.

Der Algorithmus TwoStep ist ein Algorithmus für das Datenbankmining, der auf folgenden Weise Datencluster bildet:

1. Ein CF-Baum (Clustering Feature) wird erstellt. Dieser hochgradig ausgewogene Baum speichert Clustering-Funktionen für das hierarchische Clustering, bei denen ähnliche Eingabedatensätze Teil derselben Baumknoten werden.
2. Die Blätter des CF-Baums werden speicherintern hierarchisch in Gruppen zusammengefasst, um das endgültige Clustering-Ergebnis zu generieren. Die beste Anzahl Cluster wird automatisch bestimmt. Wenn Sie eine maximale Anzahl Cluster angeben, wird die beste Anzahl Cluster innerhalb des angegebenen Grenzwerts bestimmt.
3. Das Clustering-Ergebnis wird in einem zweiten Schritt optimiert, in dem ein dem K-Means-Algorithmus ähnlicher Algorithmus auf die Daten angewendet wird.

IBM Db2 for z/OS-Modelle - TwoStep-Feldoptionen

Durch Festlegen der Feldoptionen können Sie angeben, dass die in vorgeordneten Knoten definierten Feldrolleneinstellungen verwendet werden. Sie können die Feldzuweisungen auch manuell vornehmen.

Element auswählen. Wählen Sie diese Option aus, um die Rolleneinstellungen eines vorgeordneten Typknotens oder von der Registerkarte **Typen** eines vorgeordneten Quellenknoten zu verwenden. Rolleneinstellungen sind beispielsweise Ziele und Prädiktoren.

Benutzerdefinierte Feldzuweisungen verwenden. Wählen Sie diese Option aus, wenn Sie Ziele, Prädiktoren und andere Rollen manuell zuweisen möchten.

Felder. Verwenden Sie die Pfeile, um den Rollenfeldern rechts Elemente aus dieser Liste manuell zuzuweisen. Die Symbole zeigen das gültige Messniveau für jedes Rollenfeld an.

Datensatz-ID. Das Feld, das als eindeutige ID für einen Datensatz verwendet wird.

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

IBM Db2 for z/OS-Modelle - TwoStep-Erstellungsoptionen

Durch Festlegen der Erstellungsoptionen können Sie die Erstellung des Modells Ihren Anforderungen entsprechend anpassen.

Wenn Sie ein Modell mit den Standardoptionen erstellen wollen, klicken Sie auf **Ausführen**.

Distanzmaß. Dieser Parameter definiert die Methode für die Messung der Distanz zwischen Datenpunkten. Größere Distanzen geben größere Unähnlichkeiten an. Option:

- **Log-Likelihood.** Mit dem Likelihood-Maß wird eine Wahrscheinlichkeitsverteilung für die Variablen vorgenommen. Bei stetigen Variablen wird von einer Normalverteilung, bei kategorialen Variablen von einer multinomialen Verteilung ausgegangen. Bei allen Variablen wird davon ausgegangen, dass sie unabhängig sind.

Clusteranzahl. Dieser Parameter definiert die Anzahl der zu erstellenden Cluster. Folgende Optionen stehen zur Auswahl:

- **Anzahl der Cluster automatisch berechnen.** Die Anzahl der Cluster wird automatisch berechnet. Sie können die maximale Anzahl der Cluster im Feld **Maximum** angeben.
- **Anzahl der Cluster angeben.** Geben Sie an, wie viele Cluster erstellt werden sollen.

Statistiken. Dieser Parameter definiert, wie viele Statistikdaten in das Modell eingeschlossen werden. Folgende Optionen stehen zur Auswahl:

- **Alle.** Alle spaltenbezogenen und alle wertebezogenen Statistikdaten werden eingeschlossen.

  Anmerkung: Dieser Parameter schließt die maximale Anzahl Statistikdaten ein und kann daher die Systemleistung beeinträchtigen. Wenn Sie das Modell nicht im grafischen Format anzeigen möchten, geben Sie **Keine** an.
- **Spalten.** Spaltenbezogene Statistikdaten werden eingeschlossen.
- **Keine.** Nur Statistikdaten, die für das Scoring des Modells erforderlich sind, werden eingeschlossen.

Ergebnisse reproduzieren. Aktivieren Sie dieses Kontrollkästchen, wenn Sie einen Startwert für Zufallszahlen festlegen wollen, um Analysen zu replizieren. Sie können eine ganze Zahl angeben oder durch Klicken auf **Generieren** eine pseudozufällige ganze Zahl erstellen.

IBM Db2 for z/OS-Modelle - TwoStep-Nugget - Registerkarte "Modell"

Die Registerkarte **Modell** enthält verschiedene grafische Ansichten, die Auswertungsstatistiken und Verteilungen für Felder von Clustern zeigen. Sie können die Daten aus dem Modell exportieren oder die Ansicht als Grafik exportieren.

Verwalten von IBM Db2 for z/OS-Modellen

Db2 for z/OS-Modelle werden dem Erstellungsbereich und der Modellpalette auf dieselbe Weise hinzugefügt wie andere IBM SPSS Modeler-Modelle und können auf annähernd dieselbe Weise verwendet werden.

Führen Sie die folgenden Schritte aus, um für die Daten direkt in Db2 for z/OS ein Scoring durchzuführen:

1. Installieren Sie SPSS Scoring Adapter in der Db2 for z/OS-Datenbank, in der die Daten gespeichert werden.
2. Stellen Sie sicher, dass der Stream eine Verbindung zu der Db2 for z/OS-Datenbank herstellt, in der die Daten gespeichert werden.

Durchführen eines Scorings für IBM Db2 for z/OS-Modelle

Modelle werden im Erstellungsbereich durch ein goldenes Modellnugget-Symbol repräsentiert. Der Hauptzweck eines Nuggets ist das Scoring von Daten, um Vorhersagen zu generieren oder eine weitere Analyse der Modelleigenschaften zu erlauben. Scores werden in Form eines oder mehrerer zusätzlicher Datenfelder hinzugefügt, die durch Verknüpfen eines Tabellenknotens mit dem Nugget und Ausführen des betreffenden Zweigs des Streams sichtbar gemacht werden können, wie weiter unten in diesem Abschnitt beschrieben. Einige Nugget-Dialogfelder, beispielsweise diejenigen für Entscheidungsbaum oder Regressionsbaum, enthalten zusätzlich die Registerkarte "Modell", die eine visuelle Darstellung des Modells bietet.

Die zusätzlichen Felder sind durch das Präfix $<ID>- gekennzeichnet, das dem Namen des Zielfelds hinzugefügt wird. Dabei hängt <ID> vom Modell ab und gibt den Typ der hinzugefügten Informationen an. Die unterschiedlichen Kennzeichner werden in den Themen für die einzelnen Modellnuggets beschrieben.

Führen Sie zur Anzeige der Scores folgende Schritte aus:

1. Verbinden Sie einen Tabellenknoten mit dem Modellnugget.
2. Öffnen Sie den Tabellenknoten.
3. Klicken Sie auf **Ausführen**.
4. Blättern Sie im Tabellenausgabefenster nach rechts, um die zusätzlichen Felder und ihre Scores anzuzeigen.

Anmerkung: Der Scoring-Vorgang wird nicht im Akzelerator, sondern in Db2 ausgeführt und setzt folglich voraus, dass die Eingabetabelle für das Scoring sich physisch in Db2 befindet. Daher kann als Scoring-Eingabe nur eine Db2-basierte Tabelle oder eine beschleunigte Tabelle verwendet werden. Wenn der Stream eine reine Akzeleratortabelle verwendet, wird der Fehler ausgegeben, dass die Anweisung von Db2 oder im Akzelerator nicht ausgeführt werden kann.

IBM Db2 for z/OS-Entscheidungsbaummodellnuggets

Das Entscheidungsbaummodellnugget zeigt die Ausgabe des Modellierungsvorgangs an und ermöglicht es Ihnen außerdem, einige Optionen für das Scoren des Modells festzulegen.

Wenn Sie einen Stream ausführen, der ein Entscheidungsbaummodellnugget enthält, fügt der Knoten zwei neue Felder hinzu, deren Namen aus dem Ziel abgeleitet werden.

Tabelle 26. Modellscoring-Feld für Entscheidungsbaum

Name des hinzugefügten Felds	**Bedeutung**
$I-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.
$IP-*Zielname*	Konfidenzwert (von 0,0 bis 1,0) für die Vorhersage.

Anmerkung: Aufgrund von Einschränkungen in Db2 for z/OS sind die Spaltennamen möglicherweise abgeschnitten.

IBM Db2 for z/OS-Entscheidungsbaumnugget - Registerkarte "Modell"

Auf der Registerkarte **Modell** wird der Prädiktoreinfluss des Entscheidungsbaummodells im grafischen Format angezeigt. Die Länge des Balkens gibt den Einfluss des Prädiktors an.

IBM Db2 for z/OS-Entscheidungsbaumnugget - Registerkarte "Viewer"

Die Registerkarte **Viewer** stellt den Baum eines Baummodells in derselben Weise dar, wie SPSS Modeler es für das Entscheidungsbaummodell tut.

IBM Db2 for z/OS-K-Means-Modellnugget

Nuggets für K-Means-Modelle enthalten alle Informationen, die vom Clustermodell erfasst wurden, sowie Informationen zu den Trainingsdaten und dem Schätzvorgang.

Wenn Sie einen Stream ausführen, der ein K-Means-Modellnugget enthält, fügt der Knoten zwei neue Felder hinzu, die die Clusterzugehörigkeit und die Entfernung vom zugewiesenen Clusterzentrum für den betreffenden Datensatz enthalten. Die neuen Feldnamen werden durch Präfigierung von `$KM-` für die Clusterzugehörigkeit und `$KMD-` für die Entfernung vom Clusterzentrum aus dem Modellnamen abgeleitet. Beispiel: Wenn das Modell den Namen `Kmeans` trägt, erhalten die neuen Felder die Namen `$KM-Kmeans` und `$KMD-Kmeans`.

Anmerkung: Aufgrund von Einschränkungen in Db2 for z/OS sind die Spaltennamen möglicherweise abgeschnitten.

IBM Db2 for z/OS-K-Means-Modellnugget - Registerkarte "Modell"

Die Registerkarte **Modell** enthält verschiedene grafische Ansichten, die Auswertungsstatistiken und Verteilungen für Felder von Clustern zeigen. Sie können die Daten aus dem Modell exportieren oder die Ansicht als Grafik exportieren.

IBM Db2 for z/OS-Naive Bayes-Modellnuggets

Wenn Sie einen Stream ausführen, der ein Naive Bayes-Modellnugget enthält, fügt der Knoten zwei neue Felder hinzu, deren Namen aus dem Zielnamen abgeleitet werden.

Tabelle 27. Modellscoring-Feld für Naive Bayes

Name des hinzugefügten Felds	Bedeutung
$I-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.
$IP-*Zielname*	Konfidenzwert (von 0,0 bis 1,0) für die Vorhersage.

Anmerkung: Aufgrund von Einschränkungen in Db2 for z/OS sind die Spaltennamen möglicherweise abgeschnitten.

Sie können die zusätzlichen Felder anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

IBM Db2 for z/OS-Regressionsbaummodellnuggets

Wenn Sie einen Stream ausführen, der ein Regressionsbaummodellnugget enthält, fügt der Knoten zwei neue Felder hinzu, deren Namen aus dem Zielnamen abgeleitet werden.

Tabelle 28. Model-Scoring-Feld für Regressionsbaum

Name des hinzugefügten Felds	Bedeutung
$I-*Zielname*	Vorhergesagter Wert für aktuellen Datensatz.
$IS-*Zielname*	Geschätzte Standardabweichung des vorhergesagten Werts.

Anmerkung: Aufgrund von Einschränkungen in Db2 for z/OS sind die Spaltennamen möglicherweise abgeschnitten.

Sie können die zusätzlichen Felder anzeigen, indem Sie einen Tabellenknoten zum Modellnugget hinzufügen und diesen Tabellenknoten ausführen.

IBM Db2 for z/OS-Regressionsbaumnugget - Registerkarte "Modell"

Auf der Registerkarte **Modell** wird der Prädiktoreinfluss des Regressionsbaummodells im grafischen Format angezeigt. Die Länge des Balkens gibt den Einfluss des Prädiktors an.

IBM Db2 for z/OS-Regressionsbaumnugget - Registerkarte "Viewer"

Die Registerkarte **Viewer** stellt den Baum eines Baummodells in derselben Weise dar, wie SPSS Modeler es für das Regressionsbaummodell tut.

IBM Db2 for z/OS-TwoStep-Modellnugget

Wenn Sie einen Stream ausführen, der ein TwoStep-Modellnugget enthält, fügt der Knoten zwei neue Felder hinzu, die die Clusterzugehörigkeit und die Entfernung vom zugewiesenen Clusterzentrum für den betreffenden Datensatz enthalten. Die neuen Feldnamen werden durch Hinzufügen des Präfix $TS- für die Clusterzugehörigkeit und des Präfix $TSD- für die Entfernung vom Clusterzentrum aus dem Modellnamen

abgeleitet. Beispiel: Wenn das Modell den Namen MDL trägt, erhalten die neuen Felder die Namen $TS-MDL und $TSD-MDL.

Bemerkungen

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden. IBM stellt dieses Material möglicherweise auch in anderen Sprachen zur Verfügung. Für den Zugriff auf das Material in einer anderen Sprache kann eine Kopie des Produkts oder der Produktversion in der jeweiligen Sprache erforderlich sein.

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 Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

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 Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

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 Geschäftsbedingun-

gen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Die angeführten Leistungsdaten und Kundenbeispiele dienen nur zur Illustration. Die tatsächlichen Ergebnisse beim Leistungsverhalten sind abhängig von der jeweiligen Konfiguration und den Betriebsbedingungen.

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.

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind Marken oder eingetragene Marken der Adobe Systems Incorporated in den USA und/oder anderen Ländern.

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

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und/oder anderen Ländern.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken 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.

Bedingungen für Produktdokumentation

Die Berechtigungen zur Nutzung dieser Veröffentlichungen werden Ihnen auf der Basis der folgenden Bedingungen gewährt.

Anwendbarkeit

Diese Bedingungen sind eine Ergänzung der Nutzungsbedingungen auf der IBM Website.

Persönliche Nutzung

Sie dürfen diese Veröffentlichungen für Ihre persönliche, nicht kommerzielle Nutzung unter der Voraussetzung vervielfältigen, dass alle Eigentumsvermerke erhalten bleiben. Sie dürfen diese Veröffentlichungen oder Teile der Veröffentlichungen ohne ausdrückliche Genehmigung von IBM weder weitergeben oder anzeigen noch abgeleitete Werke davon erstellen.

Kommerzielle Nutzung

Sie dürfen diese Veröffentlichungen nur innerhalb Ihres Unternehmens und unter der Voraussetzung, dass alle Eigentumsvermerke erhalten bleiben, vervielfältigen, weitergeben und anzeigen. Sie dürfen diese Veröffentlichungen oder Teile der Veröffentlichungen ohne ausdrückliche Genehmigung von IBMaußerhalb Ihres Unternehmens weder vervielfältigen, weitergeben oder anzeigen noch abgeleitete Werke davon erstellen.

Berechtigungen

Abgesehen von den hier gewährten Berechtigungen werden keine weiteren Berechtigungen, Lizenzen oder Rechte (veröffentlicht oder stillschweigend) in Bezug auf die Veröffentlichungen oder darin enthaltene Informationen, Daten, Software oder geistiges Eigentum gewährt.

IBM behält sich das Recht vor, die hierin gewährten Berechtigungen nach eigenem Ermessen zurückzuziehen, wenn sich die Nutzung der Veröffentlichungen für IBM als nachteilig erweist oder wenn die obigen Nutzungsbestimmungen nicht genau befolgt werden.

Sie dürfen diese Informationen nur in Übereinstimmung mit allen anwendbaren Gesetzen und Vorschriften, einschließlich aller US-amerikanischen Exportgesetze und Verordnungen, herunterladen und exportieren.

IBM übernimmt keine Gewährleistung für den Inhalt dieser Veröffentlichungen. Diese Veröffentlichungen werden auf der Grundlage des gegenwärtigen Zustands (auf "as-is"-Basis) und ohne eine ausdrückliche oder stillschweigende Gewährleistung für die Handelsüblichkeit, die Verwendungsfähigkeit für einen bestimmten Zweck oder die Freiheit von Rechten Dritter zur Verfügung gestellt.

Index

F

G

H

I

K

L

M

N

O

P

R

S

T

U

V

Z