IBM SPSS Data Preparation 24

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 33 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 24, Release 0, Modifikation 0 von IBM® SPSS Statistics und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Data Preparation 24,
herausgegeben von International Business Machines Corporation, USA

Herausgegeben von:
TSC Germany
Kst. 2877
Januar 2016

Inhaltsverzeichnis

Kapitel 1. Einführung in Data Preparation (Datenaufbereitung)

Der Informationsbedarf wächst proportional mit dem Anstieg der Leistungsfähigkeit von Computern. Das führt zu immer größeren Datensammlungen, zu mehr Fällen, mehr Variablen und mehr Fehlern bei der Dateneingabe. Diese Fehler behindern Vorhersagen auf der Grundlage von Vorhersagemodellen, dem wichtigsten Ziel des Data-Warehousing. Deswegen müssen die Daten "sauber" gehalten werden. Die Menge der gespeicherten Daten ist jedoch bereits so weit über die Kapazitäten zur manuellen Prüfung der Daten hinausgewachsen, dass es entscheidend ist, automatisierte Prozesse für die Datenvalidierung zu implementieren.

Mit dem Zusatzmodul "Data Preparation" (Datenaufbereitung) können Sie ungewöhnliche und ungültige Fälle, Variablen und Datenwerte im aktuellen Dataset identifizieren und Daten zur Modellierung vorbereiten.

Prozeduren von "Data Preparation" (Datenaufbereitung) verwenden

Es hängt von Ihren Bedürfnissen ab, welche Prozeduren von "Data Preparation" (Datenaufbereitung) für Sie infrage kommen. Nachdem Sie die Daten geladen haben, könnte eine typische Vorgehensweise folgendermaßen aussehen:

- **Vorbereitung der Metadaten.** Überprüfen Sie die Variablen in der Datendatei und bestimmen Sie die gültigen Werte, Beschriftungen und Messniveaus. Identifizieren Sie die Kombinationen von Variablenwerten, die zwar unmöglich, jedoch häufig falsch codiert sind. Definieren Sie auf der Grundlage dieser Informationen Validierungsregeln. Dies kann zeitraubend sein, ist jedoch den Aufwand wert, wenn Sie regelmäßig Datendateien mit ähnlichen Attributen validieren müssen.
- **Datenvalidierung.** Führen Sie grundlegende Prüfungen und Prüfungen mit definierten Validierungsregeln durch, um ungültige Fälle, Variablen und Datenwerte zu identifizieren. Wenn ungültige Daten gefunden werden, untersuchen und beseitigen Sie die Ursache. Dies macht möglicherweise einen weiteren Durchlauf durch die Vorbereitung der Metadaten erforderlich.
- **Vorbereitung des Modells.** Verwenden Sie die automatisierte Datenvorbereitung, um Transformationen der ursprünglichen Felder zu erhalten, die die Modellerstellung verbessern. Identifizieren Sie potenzielle statistische Ausreißer, die in vielen Vorhersagemodellen Probleme verursachen können. Einige Ausreißer sind das Ergebnis von ungültigen Variablenwerten, die noch nicht identifiziert wurden. Dies macht möglicherweise einen weiteren Durchlauf durch die Vorbereitung der Metadaten erforderlich.

Sobald die Datendatei "sauber" ist, können Sie Modelle in anderen Zusatzmodulen erstellen.

Kapitel 2. Validierungsregeln

Eine Regel wird verwendet, um zu entscheiden, ob ein Fall gültig ist. Es gibt zwei Typen von Validierungsregeln:

- **Regeln für eine Variable.** Regeln für eine Variable bestehen aus einem festen Set von Prüfungen, die auf eine einzige Variable angewendet werden, z. B. Prüfungen auf Werte außerhalb des Bereichs. Bei den Regeln für eine Variable können die gültigen Werte als Wertebereich oder als eine Liste zulässiger Werte ausgedrückt werden.
- **Regeln für mehrere Variablen.** Regeln für mehrere Variablen sind benutzerdefinierte Regeln, die auf eine einzige Variable oder eine Kombination von Variablen angewendet werden können. Regeln für mehrere Variablen bestehen aus einem logischen Ausdruck, der ungültige Werte kennzeichnet.

Die Validierungsregeln werden im Datenwörterbuch Ihrer Datendatei gespeichert. Dies ermöglicht es, die Regeln einmal zu definieren und später wiederzuverwenden.

Vordefinierte Validierungsregeln laden

Sie können schnell auf ein Set gebrauchsfertiger Validierungsregeln zugreifen, indem Sie vordefinierte Validierungsregeln aus einer externen Datendatei laden, die in der Installation enthalten ist.

So laden Sie vordefinierte Validierungsregeln:

1. Wählen Sie in den Menüs Folgendes aus:

   Daten > **Validierung** > **Vordefinierte Regeln laden...**

Sie können auch den Assistenten zum Kopieren von Dateneigenschaften verwenden, um Regeln aus einer beliebigen Datendatei zu laden.

Validierungsregeln definieren

Im Dialogfeld "Validierungsregeln definieren" können Sie Validierungsregeln für eine Variable oder mehrere Variablen erstellen und anzeigen.

So erstellen Sie Validierungsregeln und lassen diese anzeigen:

1. Wählen Sie in den Menüs Folgendes aus:

   Daten > **Validierung** > **Regeln definieren...**

   Das Dialogfeld wird mit Validierungsregeln für eine Variable oder mehrere Variablen ausgefüllt, die aus dem Datenwörterbuch ausgelesen werden. Wenn keine Regeln vorliegen, wird automatisch eine neue Regel als Platzhalter erzeugt, die Sie nach Bedarf anpassen können.
2. Wählen Sie einzelne Regeln auf den Registerkarten "Regeln für eine Variable" und "Regeln für mehrere Variablen" aus, um sie anzuzeigen und ihre Eigenschaften zu ändern.

Regeln für eine Variable definieren

Auf der Registerkarte "Regeln für eine Variable" können Sie Validierungsregeln für eine einzelne Variable erstellen, anzeigen und ändern.

Regeln. Die Liste zeigt Validierungsregeln für eine Variable nach Namen und den Variablentyp, auf den die jeweilige Regel angewendet werden kann. Wenn Sie das Dialogfeld öffnen, werden die im Datenwörterbuch definierten Regeln angezeigt. Falls gegenwärtig keine Regel definiert ist, wird eine Platzhalterregel mit dem Namen "EinVarRegel 1" angezeigt. Unter der Liste "Regeln" werden folgende Schaltflächen angezeigt:

- **Neu.** Fügt einen neuen Eintrag am Ende der Liste "Regeln" hinzu. Die Regel wird ausgewählt und erhält den Namen "EinVarRegel *n*". Hierbei ist *n* eine Ganzzahl, sodass der Name der Regel unter den Regeln für eine oder mehrere Variablen eindeutig ist.
- **Duplizieren.** Fügt eine Kopie der ausgewählten Regel am Ende der Liste "Regeln" hinzu. Der Name der Regel wird so angepasst, dass er unter den Regeln für eine oder mehrere Variablen eindeutig ist. Wenn Sie beispielsweise "EinVarRegel 1" duplizieren, erhält die erste duplizierte Regel den Namen "Kopie von EinVarRegel 1", die zweite den Namen "Kopie (2) von EinVarRegel 1" usw.
- **Löschen.** Löscht die ausgewählte Regel.

Regeldefinition. Mit diesen Steuerelementen können Sie die Eigenschaften für eine ausgewählte Regel anzeigen und festlegen.

- **Name.** Der Name der Regel muss unter den Regeln für eine oder mehrere Variablen eindeutig sein.
- **Typ.** Dies ist der Variablentyp, auf den die Regel angewendet werden kann. Wählen Sie **Numerisch**, **Zeichenfolge** oder **Datum** aus.
- **Format.** Hiermit können Sie das Datumsformat für die Regeln auswählen, die auf Datumsvariablen angewendet werden können.
- **Gültige Werte.** Sie können die gültigen Werte als Bereich oder als Werteliste angeben.

  Bereichsdefinition

  Mit den Steuerelementen zum Festlegen eines Bereichs können Sie einen Bereich gültiger Werte angeben. Werte, die sich außerhalb dieses Bereichs befinden, werden als ungültig gekennzeichnet.

  Um einen Bereich anzugeben, geben Sie den Minimum- oder Maximumwert oder beide Werte ein. Mit dem Kontrollkästchen können Sie festlegen, dass Werte ohne Beschriftung und nicht ganzzahlige Werte im Bereich gekennzeichnet werden.

  Listendefinition

  Mit den Steuerelementen zum Festlegen einer Liste können Sie eine Liste gültiger Werte angeben. Werte, die sich nicht in der Liste befinden, werden als ungültig gekennzeichnet.

  Geben Sie im Raster die Listenwerte ein. Mit dem Kontrollkästchen legen Sie fest, ob die Groß-/Kleinschreibung berücksichtigt wird, wenn Zeichenfolgedatenwerte gegen die Liste der zulässigen Werte geprüft werden.
- **Benutzerdefiniert fehlende Werte zulassen.** Hiermit wird festgelegt, ob benutzerdefiniert fehlende Werte als ungültig gekennzeichnet werden.
- **Systemdefiniert fehlende Werte zulassen.** Hiermit wird festgelegt, ob systemdefiniert fehlende Werte als ungültig gekennzeichnet werden. Dies gilt nicht für Regeln für Zeichenfolgen.
- **Leere Werte zulassen.** Hiermit wird festgelegt, ob leere Zeichenfolgewerte als ungültig gekennzeichnet werden. Dies gilt nur für Regeln für Nicht-Zeichenfolgen.

Regeln für mehrere Variablen definieren

Auf der Registerkarte "Regeln für mehrere Variablen" können Sie Validierungsregeln für mehrere Variablen erstellen, anzeigen und ändern.

Regeln. Die Liste enthält die Validierungsregeln für mehrere Variablen nach Namen. Wenn Sie das Dialogfeld öffnen, wird eine Platzhalterregel mit dem Namen "MehrVarRegel 1" angezeigt. Unter der Liste "Regeln" werden folgende Schaltflächen angezeigt:

- **Neu.** Fügt einen neuen Eintrag am Ende der Liste "Regeln" hinzu. Die Regel wird ausgewählt und erhält den Namen "MehrVarRegel *n*". Hierbei ist *n* eine Ganzzahl, sodass der Name der Regel unter den Regeln für eine oder mehrere Variablen eindeutig ist.
- **Duplizieren.** Fügt eine Kopie der ausgewählten Regel am Ende der Liste "Regeln" hinzu. Der Name der Regel wird so angepasst, dass er unter den Regeln für eine oder mehrere Variablen eindeutig ist. Wenn Sie beispielsweise "MehrVarRegel 1" duplizieren, erhält die erste duplizierte Regel den Namen "Kopie von MehrVarRegel 1", die zweite den Namen "Kopie (2) von MehrVarRegel 1" usw.
- **Löschen.** Löscht die ausgewählte Regel.

Regeldefinition. Mit diesen Steuerelementen können Sie die Eigenschaften für eine ausgewählte Regel anzeigen und festlegen.

- **Name.** Der Name der Regel muss unter den Regeln für eine oder mehrere Variablen eindeutig sein.
- **Logischer Ausdruck.** Im Wesentlichen ist dies die Regeldefinition. Die Auswertung des Ausdrucks für einen ungültigen Fall muss 1 entsprechen.

Erstellen von Ausdrücken

1. Um einen Ausdruck zu erstellen, fügen Sie die Komponenten in das Feld "Ausdruck" ein oder geben den Ausdruck direkt in dieses Feld ein.

- Sie können Funktionen oder häufig verwendete Systemvariablen einfügen, indem Sie eine Gruppe aus der Liste "Funktion" auswählen und in der Liste "Funktionen und Sondervariablen" auf die Funktion bzw. Variable doppelklicken (oder die Funktion bzw. Variable auswählen und auf **Einfügen** klicken). Geben Sie alle durch Fragezeichen gekennzeichneten Parameter ein (gilt nur für Funktionen). Die Funktionsgruppe mit der Beschriftung **Alle** bietet eine Liste aller verfügbaren Funktionen und Systemvariablen. Eine kurze Beschreibung der aktuell ausgewählten Funktion oder Variablen wird in einem speziellen Bereich des Dialogfelds angezeigt.
- Zeichenfolgekonstanten müssen in Anführungszeichen oder Apostrophe eingeschlossen werden.
- Wenn die Werte Dezimalstellen enthalten, muss ein Punkt (.) als Dezimaltrennzeichen verwendet werden.

Kapitel 3. Daten validieren

Im Dialogfeld "Daten validieren" können Sie verdächtige oder ungültige Fälle, Variablen und Datenwerte im aktiven Dataset identifizieren.

Beispiel. Ein Datenanalyst muss für den Auftraggeber einen monatlichen Bericht über die Kundenzufriedenheit zusammenstellen. Die monatlich erhaltenen Daten müssen einer Qualitätsprüfung unterzogen werden. Dabei muss nach ungültigen Kunden-IDs, Variablenwerten außerhalb des Bereichs sowie Kombinationen von Variablenwerten gesucht werden, die häufig fehlerhaft eingegeben werden. Im Dialogfeld "Daten validieren" kann der Analyst die Variablen angeben, durch die Kunden eindeutig identifiziert werden, Regeln für die gültigen Wertebereiche einzelner Variablen definieren und Regeln zum Erkennen unmöglicher Kombinationen für mehrere Variablen definieren. Die Prozedur liefert einen Bericht der Problemfälle und -variablen. Darüber hinaus weisen die Daten in jedem Monat die gleichen Datenelemente auf, sodass der Analyst in der Lage ist, die Regeln im folgenden Monat auf die neue Datendatei anzuwenden.

Statistiken. Die Prozedur erzeugt Listen von Variablen, Fällen und Datenwerten, die verschiedene Prüfungen nicht bestehen, Häufigkeiten der Verletzung von Regeln für einzelne oder mehrere Variablen sowie einfache deskriptive Auswertungen der Analysevariablen.

Gewichtungen. Die Prozedur ignoriert Angaben zur Gewichtungsvariablen und behandelt diese stattdessen wie jede andere Analysevariable.

So validieren Sie Daten:

1. Wählen Sie in den Menüs Folgendes aus:

   Daten > **Validierung** > **Daten validieren...**
2. Wählen Sie mindestens eine Analysevariable aus, die durch grundlegende Variablenprüfungen oder Validierungsregeln für eine Variable validiert werden soll.

   Sie haben außerdem folgende Möglichkeiten:
3. Klicken Sie auf die Registerkarte **Regeln für mehrere Variablen** und wenden Sie mindestens eine Regel für mehrere Variablen an.

Die folgenden Optionen sind verfügbar:

- Wählen Sie mindestens eine Fall-ID-Variable aus, um nach doppelten oder unvollständigen IDs zu suchen. Fall-ID-Variablen werden auch zum Beschriften der fallweisen Ausgabe verwendet. Wenn mehr als eine Fall-ID-Variable angegeben wurde, wird die Kombination der Werte als Fall-ID behandelt.

Felder mit unbekanntem Messniveau

Der Messniveau-Alert wird angezeigt, wenn das Messniveau für mindestens eine Variable (ein Feld) im Dataset unbekannt ist. Da sich das Messniveau auf die Berechnung der Ergebnisse für diese Prozedur auswirkt, müssen alle Variablen ein definiertes Messniveau aufweisen.

Daten durchsuchen. Liest die Daten im aktiven Dataset und weist allen Feldern, deren Messniveau zurzeit nicht bekannt ist, das Standardmessniveau zu. Bei großen Datasets kann dieser Vorgang einige Zeit in Anspruch nehmen.

Manuell zuweisen. Öffnet ein Dialogfeld, in dem alle Felder mit unbekanntem Messniveau aufgeführt werden. Mit diesem Dialogfeld können Sie diesen Feldern ein Messniveau zuweisen. Außerdem können Sie in der Variablenansicht des Dateneditors ein Messniveau zuweisen.

Da das Messniveau für diese Prozedur bedeutsam ist, können Sie erst dann auf das Dialogfeld zur Ausführung dieser Prozedur zugreifen, wenn für alle Felder ein Messniveau definiert wurde.

Daten validieren: Grundlegende Prüfungen

Auf der Registerkarte "Grundlegende Prüfungen" können Sie grundlegende Prüfverfahren für Analysevariablen, Fall-IDs und ganze Fälle auswählen.

Analysevariable. Wenn Sie auf der Registerkarte "Variablen" Analysevariablen ausgewählt haben, können Sie die folgenden Gültigkeitsprüfungen auswählen. Mit den Kontrollkästchen können Sie die einzelnen Prüfungen aktivieren oder inaktivieren.

- **Maximaler Prozentsatz fehlender Werte.** Gibt Analysevariablen aus, bei denen der prozentuale Anteil fehlender Werte den angegebenen Wert übersteigt. Der angegebene Wert muss eine positive Zahl kleiner oder gleich 100 sein.
- **Maximaler Prozentsatz der Fälle in einer einzelnen Kategorie.** Wenn kategoriale Analysevariablen vorhanden sind, werden bei dieser Option kategoriale Analysevariablen ausgegeben, bei denen der prozentuale Anteil der Fälle, die eine einzelne nicht fehlende Kategorie darstellen, den angegebenen Wert übersteigt. Der angegebene Wert muss eine positive Zahl kleiner oder gleich 100 sein. Der Prozentsatz entspricht dem Anteil der Fälle mit nicht fehlenden Werten der Variablen.
- **Maximaler Prozentsatz der Kategorien mit Anzahl 1.** Wenn kategoriale Analysevariablen vorhanden sind, werden bei dieser Option kategoriale Analysevariablen ausgegeben, bei denen der prozentuale Anteil der Kategorien der Variablen, die nur einen Fall enthalten, den angegebenen Wert übersteigt. Der angegebene Wert muss eine positive Zahl kleiner oder gleich 100 sein.
- **Minimaler Variationskoeffizient.** Wenn metrische Analysevariablen vorhanden sind, werden bei dieser Option metrische Analysevariablen ausgegeben, bei denen der absolute Wert des Variationskoeffizienten kleiner als der angegebene Wert ist. Diese Option betrifft nur Variablen mit einem von 0 abweichenden Mittelwert. Der angegebene Wert muss eine nicht negative Zahl sein. Durch Angabe von 0 wird die Prüfung des Variationskoeffizienten inaktiviert.
- **Minimale Standardabweichung.** Wenn metrische Analysevariablen vorhanden sind, werden bei dieser Option metrische Analysevariablen ausgegeben, deren Standardabweichung kleiner als der angegebene Wert ist. Der angegebene Wert muss eine nicht negative Zahl sein. Durch Angabe von 0 wird die Prüfung der Standardabweichung inaktiviert.

Fall-IDs. Wenn Sie auf der Registerkarte "Variablen" Fall-ID-Variablen ausgewählt haben, können Sie die folgenden Gültigkeitsprüfungen auswählen.

- **Unvollständige IDs markieren.** Bei dieser Option werden Fälle mit unvollständigen Fall-IDs ausgegeben. Eine ID wird bei einem gegebenen Fall als unvollständig betrachtet, wenn der Wert einer ID-Variable leer ist oder fehlt.
- **Doppelte IDs markieren.** Bei dieser Option werden Fälle mit doppelten Fall-IDs ausgegeben. Unvollständige Fall-IDs werden aus der Menge der möglichen doppelten Werte ausgeschlossen.

Leere Fälle markieren. Bei dieser Option werden Fälle ausgegeben, bei denen alle Variablen leer sind oder fehlen. Sie können festlegen, ob zum Identifizieren leerer Fälle alle Variablen in der Datei (mit Ausnahme von ID-Variablen) oder nur die auf der Registerkarte "Variablen" ausgewählten Analysevariablen herangezogen werden sollen.

Daten validieren: Regeln für eine Variable

Auf der Registerkarte "Regeln für eine Variable" werden verfügbare Validierungsregeln für eine Variable angezeigt, die Sie auf die Analysevariablen anwenden können. Um weitere Regeln für einzelne Variablen zu definieren, klicken Sie auf **Regeln definieren**. Weitere Informationen finden Sie im Thema „Regeln für eine Variable definieren” auf Seite 3.

Analysevariable. In der Liste werden Analysevariablen aufgeführt, ihre Verteilungen zusammengefasst und die Anzahl der Regeln angezeigt, die auf jede Variable angewendet werden. Beachten Sie, dass benutzerdefiniert und systemdefiniert fehlende Werte nicht in den Zusammenfassungen enthalten sind. Durch die Dropdown-Liste "Anzeige" wird gesteuert, welche Variablen angezeigt werden. Zur Auswahl stehen **Alle Variablen**, **Numerische Variablen**, **Zeichenfolgevariablen** und **Datumsvariablen**.

Regeln. Um Regeln auf Analysevariablen anzuwenden, wählen Sie eine oder mehrere Variablen aus und aktivieren Sie in der Liste "Regeln" alle anzuwendenden Regeln. In der Liste "Regeln" werden nur Regeln aufgeführt, die für die ausgewählten Analysevariablen geeignet sind. Wenn beispielsweise numerische Variablen ausgewählt wurden, werden nur numerische Regeln angezeigt. Wurde eine Zeichenfolgevariable ausgewählt, werden nur Zeichenfolgeregeln angezeigt. Wenn keine Analysevariablen ausgewählt wurden oder die ausgewählten Variablen unterschiedliche Datentypen aufweisen, werden keine Regeln angezeigt.

Variablenverteilungen. Die in der Liste "Analysevariablen" angezeigten Verteilungszusammenfassungen können auf allen Fällen beruhen oder auf einer Durchsuchung der ersten *n* Fälle. Dies wird im Textfeld "Fälle" festgelegt. Wenn Sie auf **Erneut durchsuchen** klicken, werden die Verteilungszusammenfassungen aktualisiert.

Daten validieren: Regeln für mehrere Variablen

Auf der Registerkarte "Regeln für mehrere Variablen" werden verfügbare Regeln für mehrere Variablen angezeigt, die Sie auf die Daten anwenden können. Um weitere Regeln für mehrere Variablen zu definieren, klicken Sie auf **Regeln definieren**. Weitere Informationen finden Sie im Thema „Regeln für mehrere Variablen definieren” auf Seite 4.

Daten validieren: Ausgabe

Fallweiser Bericht. Wenn Sie Validierungsregeln für eine Variable oder mehrere Variablen ausgewählt haben, können Sie einen Bericht anfordern, der die Verletzungen der Validierungsregeln für einzelne Fälle enthält.

- **Mindestanzahl der Verletzungen, damit ein Fall enthalten ist.** Mit dieser Option wird die Mindestanzahl der Verletzungen angegeben, die erforderlich sind, damit ein Fall in den Bericht aufgenommen wird. Geben Sie eine positive Ganzzahl ein.
- **Maximale Anzahl an Fällen.** Mit dieser Option wird die Höchstanzahl der Fälle angegeben, die im Fallbericht enthalten sein soll. Geben Sie eine positive ganze Zahl kleiner oder gleich 1000 ein.

Validierungsregeln für eine Variable. Wenn Sie Validierungsregeln für einzelne Variablen angewendet haben, können Sie auswählen, ob und wie die Ergebnisse angezeigt werden sollen.

- **Verletzungen nach Analysevariable zusammenfassen.** Bei dieser Option werden für jede Analysevariable alle Validierungsregeln für eine Variable aufgeführt, die verletzt wurden, und die Anzahl der Werte angegeben, die eine Verletzung der einzelnen Regeln darstellen. Außerdem wird für jede Variable die Gesamtanzahl der Verletzungen von Regeln für eine Variable ausgegeben.
- **Verletzungen nach Regel zusammenfassen.** Bei dieser Option werden für jede Validierungsregel für eine Variable die Variablen ausgegeben, die die Regeln verletzen, und die Anzahl der ungültigen Werte pro Variable angegeben. Außerdem wird variablenübergreifend die Gesamtanzahl der Werte ausgegeben, die eine Verletzung der einzelnen Regeln darstellen.

Deskriptive Statistik für Analysevariablen anzeigen. Mit dieser Option können Sie deskriptive Statistiken für Analysevariablen anfordern. Für jede kategoriale Variable wird eine Häufigkeitstabelle erzeugt. Für metrische Variablen wird eine Tabelle mit Auswertungsstatistiken erzeugt, darunter der Mittelwert, die Standardabweichung, das Minimum und das Maximum.

Fälle, die Validierungsregeln verletzen, an den Anfang des aktiven Datasets verschieben. Bei dieser Option werden Fälle mit Verletzungen von Regeln für eine oder mehrere Variablen an den Anfang des aktiven Datasets verschoben, damit sie einfacher aufgefunden werden können.

Daten validieren: Speichern

Mithilfe der Registerkarte "Speichern" können Sie Variablen, bei denen Regelverletzungen verzeichnet wurden, im aktiven Dataset speichern.

Auswertungsvariablen. Hierbei handelt es sich um einzelne Variablen, die gespeichert werden können. Aktivieren Sie die Kontrollkästchen der zu speichernden Variablen. Für die Variablen sind Standardnamen vorgegeben, die Sie bearbeiten können.

- **Indikator für leere Fälle.** Leeren Fällen wird der Wert 1 zugeordnet. Alle anderen Fälle werden als 0 codiert. Die Werte der Variablen entsprechen dem Umfang, der auf der Registerkarte "Grundlegende Prüfungen" angegeben wurde.
- **Gruppe mit doppelten IDs.** Fälle, die dieselbe Fall-ID aufweisen (mit Ausnahme von Fällen mit unvollständigen IDs), erhalten dieselbe Gruppennummer. Fälle mit eindeutigen oder unvollständigen IDs werden als 0 codiert.
- **Unvollständiger ID-Indikator.** Fälle mit leeren oder unvollständigen Fall-IDs erhalten den Wert 1. Alle anderen Fälle werden als 0 codiert.
- **Verletzungen von Validierungsregeln.** Dies ist die Gesamtanzahl der Verletzungen von Validierungsregeln für eine Variable oder mehrere Variablen pro Fall.

Vorhandene Auswertungsvariablen ersetzen. In der Datendatei gespeicherte Variablen müssen eindeutige Namen aufweisen. Wenn dies nicht der Fall ist, werden Variablen mit demselben Namen ersetzt.

Indikatorvariablen speichern, die alle Verletzungen von Validierungsregeln aufzeichnen. Bei dieser Option wird ein vollständiger Bericht über die Verletzungen der Validierungsregeln gespeichert. Jede Variable entspricht der Anwendung einer Validierungsregel und weist den Wert 1 auf, wenn der Fall die Regel verletzt, oder den Wert 0, wenn die Regel nicht verletzt wird.

Kapitel 4. Automatisierte Datenaufbereitung

Die Aufbereitung von Daten zur Analyse ist einer der wichtigsten Schritte in jedem Projekt – und gewöhnlich auch einer der zeitaufwendigsten. Die automatisierte Datenaufbereitung (ADP - Automated Data Preparation) übernimmt diese Aufgabe für Sie. Sie analysiert Ihre Daten und identifiziert Problemlösungen, findet problematische oder wahrscheinlich nicht nützliche Felder, leitet zum passenden Zeitpunkt neue Attribute ab und verbessert die Leistungsfähigkeit durch intelligente Screening-Methoden. Sie können den Algorithmus **vollautomatisch** verwenden und so Problemlösungen auswählen und anwenden oder Sie können ihn **interaktiv** verwenden und so die Änderungen in einer Vorschau betrachten, bevor sie vorgenommen werden, und sie gegebenenfalls akzeptieren oder ablehnen.

Mit ADP können Sie Ihre Daten schnell und einfach für die Modellerstellung aufbereiten, ohne über Vorkenntnisse der dazugehörigen statistischen Konzepte verfügen zu müssen. Modelle lassen sich damit schneller erstellen und scoren; zudem verbessert sich mit ADP die Robustheit automatisierter Modellierungsprozesse.

Hinweis: Wenn die ADP ein Feld für die Analyse vorbereitet, erstellt sie ein neues Feld, das die Anpassungen oder Transformationen enthält, anstatt die bestehenden Werte und Eigenschaften des alten Felds zu ersetzen. Das alte Feld wird bei der weiteren Analyse nicht verwendet; seine Rolle wird auf "Keine" gesetzt. Beachten Sie außerdem, dass Informationen zu benutzerdefiniert fehlenden Werten nicht in diese neu erstellten Felder übertragen werden und dass alle fehlenden Werte im neuen Feld systemdefiniert fehlend sind.

Beispiel. Eine Versicherungsgesellschaft mit beschränkten Ressourcen für die Untersuchung der Versicherungsansprüche von Hauseigentümern möchte ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen. Vor Erstellung des Modells werden die Daten für die Modellierung mithilfe der automatisierten Datenaufbereitung vorbereitet. Da die vorgeschlagenen Transformationen zunächst überprüft werden sollen, bevor die Transformationen angewendet werden, nutzt das Unternehmen die automatisierte Datenaufbereitung im interaktiven Modus.

Eine Gruppe in der Kraftfahrzeugindustrie erfasst die Verkaufszahlen verschiedener Personenkraftwagen. Um starke und schwache Modelle identifizieren zu können, soll eine Beziehung zwischen den Fahrzeugverkaufszahlen und den Fahrzeugeigenschaften hergestellt werden. Zur Aufbereitung der Daten für die Analyse wird die automatisierte Datenaufbereitung verwendet. Es werden Modelle mit Daten "vor" und "nach" der Aufbereitung erstellt, um zu sehen, wie sich die Ergebnisse unterscheiden.

Was ist Ihr Ziel? Die automatisierte Datenaufbereitung empfiehlt Schritte zur Datenaufbereitung, die sich auf die Geschwindigkeit auswirken, mit der andere Algorithmen Modelle erstellen können und die Vorhersagekraft dieser Modelle verbessern. Diese können die Transformation, Erstellung und Auswahl von Funktionen beinhalten. Das Ziel kann ebenfalls transformiert werden. Sie können die Prioritäten der Modellerstellung festlegen, auf die sich die Datenaufbereitung konzentrieren soll.

- **Geschwindigkeit und Genauigkeit ausgleichen.** Diese Option bereitet die Daten auf und sorgt dabei für eine ausgeglichene Priorität zwischen der Geschwindigkeit, mit der Daten durch die Modellerstellung verarbeitet werden, und der Genauigkeit der Vorhersagen.
- **Geschwindigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Geschwindigkeit Vorrang, mit der Daten durch Modellerstellungsalgorithmen verarbeitet werden. Wählen Sie diese Option, wenn Sie mit sehr großen Datasets arbeiten oder nach einer schnellen Antwort suchen.
- **Genauigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Genauigkeit der durch Modellerstellungsalgorithmen erzeugten Vorhersagen Vorrang.

- **Analyse anpassen.** Wählen Sie diese Option, wenn Sie den Algorithmus auf der Registerkarte "Einstellungen" manuell ändern wollen. Beachten Sie, dass diese Einstellung automatisch ausgewählt wird, wenn Sie anschließend Änderungen auf der Registerkarte "Einstellungen" vornehmen, die mit einem der anderen Ziele nicht kompatibel sind.

Automatische Datenaufbereitung aktivieren

1. Wählen Sie in den Menüs Folgendes aus:

   Transformieren > **Daten für Modellierung vorbereiten** > **Automatisch...**
2. Klicken Sie auf **Ausführen**.

Die folgenden Optionen sind verfügbar:

- Geben Sie ein Ziel auf der Registerkarte "Ziel" an.
- Geben Sie Feldzuweisungen auf der Registerkarte "Felder" an.
- Geben Sie Experteneinstellungen auf der Registerkarte "Einstellungen" an.

Interaktive Datenaufbereitung aktivieren

1. Wählen Sie in den Menüs Folgendes aus:

   Transformieren > **Daten für Modellierung vorbereiten** > **Interaktiv...**
2. Klicken Sie auf **Analysieren** in der Symbolleiste im oberen Bereich des Dialogfelds.
3. Klicken Sie auf die Registerkarte "Analyse" und überprüfen Sie die folgenden Schritte der Datenaufbereitung.
4. Sind alle Angaben korrekt, klicken Sie auf **Ausführen**. Andernfalls klicken Sie auf **Analyse löschen**, ändern die Einstellungen nach Ihren Wünschen und klicken dann auf **Analysieren**.

Die folgenden Optionen sind verfügbar:

- Geben Sie ein Ziel auf der Registerkarte "Ziel" an.
- Geben Sie Feldzuweisungen auf der Registerkarte "Felder" an.
- Geben Sie Experteneinstellungen auf der Registerkarte "Einstellungen" an.
- Speichern Sie die vorgeschlagenen Schritte der Datenaufbereitung in eine XML-Datei, indem Sie auf **XML speichern** klicken.

Registerkarte "Felder"

Die Registerkarte "Felder" gibt an, welche Felder zur weiteren Analyse aufbereitet werden sollen.

Vordefinierte Rollen verwenden. Diese Option greift auf bestehende Feldinformationen zurück. Wenn ein einzelnes Feld mit einer Rolle als "Ziel" vorhanden ist, wird es als Ziel verwendet; in allen anderen Fällen ist kein Ziel vorhanden. Alle Felder mit der vordefinierten Rolle "Eingabe" werden als Eingaben verwendet. Mindestens ein Eingabefeld ist erforderlich.

Benutzerdefinierte Feldzuweisungen verwenden. Wenn Sie Feldrollen durch Verschieben von Feldern aus ihren Standardlisten überschreiben, springt das Dialogfeld automatisch auf diese Option. Wenn Sie benutzerdefinierte Feldzuweisungen vornehmen, geben Sie die folgenden Felder an:

- **Ziel (optional).** Wählen Sie das Zielfeld aus, wenn Sie Modelle erstellen möchten, für die ein Ziel erforderlich ist. Dies gleicht in etwa der Einstellung der Feldrolle auf "Ziel".
- **Eingaben.** Wählen Sie mindestens ein Eingabefeld aus. Dies gleicht in etwa der Einstellung der Feldrolle auf "Eingabe".

Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" enthält mehrere unterschiedliche Gruppen von Einstellungen, die Sie ändern können, um genau festzulegen, wie der Algorithmus Ihre Daten verarbeiten soll. Wenn Sie an den Standardeinstellungen Änderungen vornehmen, die mit den anderen Zielen nicht kompatibel sind, wird auf der Registerkarte "Ziel" automatisch die Option **Analyse anpassen** ausgewählt.

Datum und Uhrzeit aufbereiten

Viele Modellierungsalgorithmen sind nicht in der Lage, Datums- und Zeitangaben direkt zu behandeln. Mit diesen Einstellungen können Sie neue Laufzeitdaten ableiten, die Sie in Ihren bestehenden Daten als Modelleingaben aus Datums- und Zeitangaben verwenden können. Die Felder mit Datums- und Zeitangaben müssen mit Datums- oder Zeitspeichertypen vordefiniert sein. Die ursprünglichen Datums- und Zeitfelder werden nicht als Modelleingaben nach der automatisierten Datenaufbereitung empfohlen.

Datums- und Zeitangaben für Modellierung aufbereiten. Durch Inaktivieren dieser Option werden alle anderen Datums- und Zeiteingaben inaktiviert und die Auswahl wird beibehalten.

Verstrichene Zeit bis zum Referenzdatum berechnen. Errechnet die Anzahl der Jahre/Monate/Tage seit einem Referenzdatum für jede Variable, die Datumsangaben enthält.

- **Referenzdatum.** Geben Sie das Datum an, ab dem die Dauer bezüglich der Datumsinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von **Heutiges Datum** wird das aktuelle Systemdatum stets verwendet, wenn ADP ausgeführt wird. Um ein bestimmtes Datum zu verwenden, wählen Sie **Festes Datum** und geben Sie das erforderliche Datum ein.
- **Einheiten für Datumsdauer.** Legen Sie fest, ob ADP die Einheit der Datumsdauer automatisch bestimmen soll, oder wählen Sie **Feste Einheiten** für Jahre, Monate oder Tage.

Verstrichene Zeit bis zur Referenzzeit berechnen. Errechnet die Anzahl der Stunden/Minuten/Sekunden seit einer Referenzzeit für jede Variable, die Uhrzeiten enthält.

- **Referenzzeit.** Geben Sie die Zeit an, ab der die Dauer bezüglich der Zeitinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von **Aktuelle Uhrzeit** wird die aktuelle Systemzeit stets verwendet, wenn ADP ausgeführt wird. Um eine bestimmte Uhrzeit zu verwenden, wählen Sie **Feste Uhrzeit** und geben Sie die erforderlichen Daten ein.
- **Einheiten für Zeitdauer** Legen Sie fest, ob ADP die Einheit der Zeitdauer automatisch bestimmen soll, oder wählen Sie **Feste Einheiten** für Stunden, Minuten oder Sekunden.

Zyklische Zeitelemente extrahieren. Verwenden Sie diese Einstellungen, um ein einzelnes Datums- oder Zeitfeld in ein oder mehrere Felder aufzuteilen. Wenn Sie zum Beispiel alle drei Datumskontrollkästchen auswählen, wird das Eingabedatumsfeld "1954-05-23" in drei Felder aufgeteilt: 1954, 5 und 23, wobei jedes das unter **Feldnamen** definierte Suffix verwendet und das ursprüngliche Datumsfeld ignoriert wird.

- **Aus Datumsangaben extrahieren.** Legen Sie für eine beliebige Datumseingabe fest, ob Sie Jahre, Monate, Tage oder eine Kombination daraus extrahieren möchten.
- **Aus Zeitangaben extrahieren.** Legen Sie für eine beliebige Zeiteingabe fest, ob Sie Stunden, Minuten, Sekunden oder eine Kombination daraus extrahieren möchten.

Felder ausschließen

Schlechte Datenqualität kann sich negativ auf die Genauigkeit Ihrer Vorhersagen auswirken. Sie können daher die akzeptable Qualitätsstufe für Eingabemerkmale festlegen. Alle konstanten oder 100 % an fehlenden Werten aufweisenden Felder werden automatisch ausgeschlossen.

Eingabefelder mit niedriger Qualität ausschließen. Durch Inaktivieren dieser Option werden alle anderen Befehle "Felder ausschließen" inaktiviert und die Auswahl wird beibehalten.

Felder mit zu vielen fehlenden Werten ausschließen. Felder mit mehr als dem angegebenen Prozentsatz an fehlenden Werten werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Inaktivieren dieser Option entspricht, und einen Wert kleiner oder gleich 100, sodass Felder mit nur fehlenden Werten automatisch ausgeschlossen werden. Der Standardwert ist 50.

Nominale Felder mit zu vielen eindeutigen Kategorien ausschließen. Nominale Felder mit mehr als der angegebenen Anzahl an Kategorien werden aus der weiteren Analyse ausgeschlossen. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 100. Dies ist nützlich für das automatische Entfernen von Feldern aus der Modellierung, die eine für jeden Datensatz eindeutige Information enthalten, wie zum Beispiel eine ID, eine Adresse oder einen Namen.

Kategoriale Felder mit zu vielen Werten in einer einzelnen Kategorie ausschließen. Ordinale und nominale Felder mit einer Kategorie, die mehr als die angegebene Prozentzahl an Datensätzen enthält, werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Inaktivieren dieser Option entspricht, und einen Wert kleiner oder gleich 100, sodass konstante Felder automatisch ausgeschlossen werden. Der Standardwert ist 95.

Messniveau anpassen

Messniveau anpassen. Durch Inaktivieren dieser Option werden alle anderen Befehle "Messniveau anpassen" inaktiviert und die Auswahl wird beibehalten.

Messniveau. Legen Sie fest, ob das Messniveau von stetigen Feldern mit "zu wenigen" Werten auf ordinal und von ordinalen Feldern mit "zu vielen" Werten auf stetig angepasst werden kann.

- **Maximale Anzahl an Werten für ordinale Felder.** Ordinale Felder mit mehr als der angegebenen Anzahl an Kategorien werden in stetige Felder umgewandelt. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 10. Dieser Wert kann größer oder gleich der Mindestanzahl an Werten für stetige Felder sein.
- **Minimale Anzahl an Werten für stetige Felder.** Stetige Felder mit weniger als der angegebenen Anzahl an eindeutigen Werten werden in ordinale Felder umgewandelt. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 5. Dieser Wert kann kleiner oder gleich der Höchstanzahl an Werten für ordinale Felder sein.

Datenqualität verbessern

Felder zur Verbesserung der Datenqualität aufbereiten. Durch Inaktivieren dieser Option werden alle anderen Einstellungen zu "Datenqualität verbessern" inaktiviert und die Auswahl wird beibehalten.

Ausreißerbehandlung. Legen Sie fest, ob Ausreißer für die Eingaben und Ziele ersetzt werden sollen. Wenn ja, geben Sie ein in Standardabweichungen gemessenes Ausreißertrennwertkriterium und eine Methode zum Ersetzen der Ausreißer an. Ausreißer können entweder durch Trimmen (durch Setzen auf den Trennwert) oder durch Einstufung als fehlende Werte ersetzt werden. Jeder als fehlender Wert eingestufte Ausreißer unterliegt den unten ausgewählten Einstellungen für die Behandlung fehlender Werte.

Fehlende Werte ersetzen. Legen Sie fest, ob fehlende Werte von stetigen, nominalen oder ordinalen Feldern ersetzt werden sollen.

Nominale Felder neu sortieren. Mit dieser Option werden die Werte von nominalen (Set-)Feldern von der kleinsten (am seltensten auftretenden) zur größten (am häufigsten auftretenden) Kategorie umcodiert. Die neuen Feldwerte starten mit 0 als der seltensten Kategorie. Hinweis: Das neue Feld ist numerisch, auch wenn das originale Feld eine Zeichenfolge enthält. Wenn zum Beispiel die Datenwerte eines nominalen Felds "A", "A", "A", "B", "C", "C" sind, codiert die automatisierte Datenaufbereitung "B" zu 0 um, "C" zu 1 und "A" zu 2.

Felder neu skalieren

Felder neu skalieren. Durch Inaktivieren dieser Option werden alle anderen Eingaben zu "Felder neu skalieren" inaktiviert und die Auswahl wird beibehalten.

Analysegewichtung. Diese Variable enthält Analysegewichtungen (Regression oder Stichprobe). Analysegewichtungen werden verwendet, um Differenzen in der Varianz zwischen den Ebenen des Zielfelds zu berücksichtigen. Wählen Sie ein stetiges Feld aus.

Stetige Eingabefelder. Mit dieser Option werden stetige Eingabefelder durch eine **z-Score-Transformation** oder eine **Min./Max. Transformation** normalisiert. Die Neuskalierung von Eingaben ist besonders nützlich, wenn Sie **Merkmalerstellung durchführen** in den Einstellungen "Auswählen und erstellen" auswählen.

- **Z-Score-Transformation.** Die Felder werden mithilfe des beobachteten Mittelwerts und der Standardabweichung als Schätzungen der Populationsparameter standardisiert und die *z*-Scores werden anschließend den entsprechenden Werten einer Normalverteilung mit den Angaben für **Endgültiger Mittelwert** und **Endgültige Standardabweichung** zugeordnet. Geben Sie eine Zahl für **Endgültiger Mittelwert** und eine positive Zahl für **Endgültige Standardabweichung** an. Die Standardwerte sind entsprechend der standardisierten Neuskalierung 0 bzw. 1.
- **Min./Max. Transformation.** Die Felder werden mithilfe der beobachteten Mindest- und Höchstwerte als Schätzungen der Populationsparameter den entsprechenden Werten einer Gleichverteilung mit den Angaben für **Minimum** und **Maximum** zugeordnet. Geben Sie für **Maximum** eine Zahl größer als **Minimum** an.

Stetiges Ziel. Mit dieser Option wird ein stetiges Feld mithilfe der Box-Cox-Transformation in ein Feld transformiert, das eine ungefähre Normalverteilung mit den Angaben für **Endgültiger Mittelwert** und **Endgültige Standardabweichung** aufweist. Geben Sie eine Zahl für **Endgültiger Mittelwert** und eine positive Zahl für **Endgültige Standardabweichung** an. Die Standardwerte sind 0 bzw. 1.

Hinweis: Wenn ein Ziel durch ADP transformiert wurde, bewerten nachfolgend mithilfe des transformierten Zielscores erstellte Modelle die transformierten Einheiten. Um die Ergebnisse interpretieren und verwenden zu können, müssen Sie den vorhergesagten Wert wieder in das ursprüngliche metrische Maß zurückkonvertieren. Weitere Informationen finden Sie im Thema „Scores zurücktransformieren” auf Seite 24.

Felder transformieren

Um die Vorhersagekraft Ihrer Daten zu verbessern, können Sie die Eingabefelder transformieren.

Feld für Modellierung transformieren. Durch Inaktivieren dieser Option werden alle anderen Eingaben zu "Felder transformieren" inaktiviert und die Auswahl wird beibehalten.

Kategoriale Eingabefelder Die folgenden Optionen sind verfügbar:

- **Dünn besetzte Kategorien zur Maximierung des Zielzusammenhangs zusammenführen.** Mit dieser Option erstellen Sie ein sparsameres Modell, indem die Anzahl der zu verarbeitenden Felder in Zusammenhang mit dem Ziel reduziert wird. Ähnliche Kategorien werden anhand der Beziehung zwischen der Eingabe und dem Ziel identifiziert. Kategorien, die sich nicht signifikant unterscheiden (d. h. einen *p*-Wert aufweisen, der größer als der angegebene Wert ist), werden zusammengeführt. Geben Sie einen Wert größer als 0 und kleiner oder gleich 1 an. Wenn alle Kategorien zu einer zusammengeführt werden, werden die Original- und abgeleiteten Versionen des Felds aus der weiteren Analyse ausgeschlossen, da sie keinen Wert als Prädiktor aufweisen.
- **Wenn kein Ziel existiert, dünn besetzte Kategorien auf der Basis folgender Häufigkeiten zusammenführen.** Wenn das Dataset kein Ziel aufweist, können Sie dünn besetzte Kategorien von ordinalen und nominalen Feldern zusammenführen. Die Methode der gleichen Häufigkeiten wird verwendet, um Kategorien mit weniger als dem angegebenen Mindestprozentsatz der Gesamtanzahl an Datensätzen zusammenzuführen. Geben Sie einen Wert größer oder gleich 0 und kleiner als 100 ein. Der Standard-

wert ist 10. Die Zusammenführung wird beendet, wenn keine Kategorien mit weniger als dem angegebenen Mindestprozentsatz an Fällen vorhanden sind oder wenn nur noch zwei Kategorien übrig sind.

Stetige Eingabefelder. Wenn das Dataset ein kategoriales Ziel enthält, können Sie stetige Eingaben mit starkem Zusammenhang einteilen, um die Verarbeitungsleistung zu verbessern. Klassen werden anhand der Eigenschaften "homogener Subsets" erstellt, die durch die Scheffé-Methode mithilfe des angegebenen *p*-Werts als Alpha für den kritischen Wert zur Bestimmung homogener Subsets identifiziert werden. Geben Sie einen Wert größer als 0 und kleiner oder gleich 1 ein. Der Standardwert ist 0,05. Wenn in dem Klassierungsvorgang eine einzelne Klassierung für ein bestimmtes Feld durchgeführt wird, werden die Original- und eingeteilten Versionen des Felds ausgeschlossen, da sie keinen Wert als Prädiktor aufweisen.

Hinweis: Die Klassierung in ADP unterscheidet sich von der optimalen Klassierung. Bei der optimalen Klassierung werden Entropieinformationen verwendet, um ein stetiges Feld in ein kategoriales Feld umzuwandeln. Dazu müssen Daten sortiert und im Arbeitsspeicher abgelegt werden. ADP verwendet homogene Subsets zum Klassieren eines stetigen Felds, das bedeutet, dass die ADP-Klassierung keine Daten sortieren und im Arbeitsspeicher ablegen muss. Der Einsatz homogener Subsets zum Klassieren eines stetigen Felds bedeutet, dass die Anzahl der Kategorien nach der Klassierung immer kleiner oder gleich der Anzahl der Kategorien im Ziel ist.

Auswählen und erstellen

Um die Vorhersagekraft Ihrer Daten zu verbessern, können Sie basierend auf den bestehenden Feldern neue Felder erstellen.

Merkmalauswahl durchführen. Eine stetige Eingabe wird aus der Analyse entfernt, wenn der *p*-Wert für seine Korrelation mit dem Ziel größer ist als der angegebene *p*-Wert.

Merkmalerstellung durchführen. Wählen Sie diese Option aus, um neue Funktionen von einer Kombination aus mehreren bestehenden Funktionen abzuleiten. Die alten Funktionen werden bei der weiteren Analyse nicht verwendet. Diese Option gilt nur für stetige Eingabemerkmale mit stetigem Ziel oder Eingabemerkmale, in denen kein Ziel vorhanden ist.

Feldnamen

Zur einfachen Identifikation neuer und transformierter Merkmale erstellt ADP allgemeine neue Namen, Präfixe oder Suffixe und wendet diese an. Sie können diese Namen ändern und ihnen mehr Aussagekraft für Ihre eigenen Anforderungen und Daten geben.

Transformierte und erstellte Felder. Geben Sie die Namenserweiterungen an, die auf transformierte Ziel- und Eingabefelder angewendet werden sollen.

Geben Sie außerdem über die Einstellungen "Auswählen und erstellen" den Präfixnamen an, der auf erstellte Funktionen angewendet werden soll. Der neue Name wird erstellt, indem ein numerisches Suffix an diesen Präfixstammnamen angehängt wird. Das Zahlenformat hängt davon ab, wie viele neue Merkmale abgeleitet werden, zum Beispiel:

- Es werden 1-9 erstellte Merkmale benannt: Merkmal1 bis Merkmal9.
- Es werden 10-99 erstellte Merkmale benannt: Merkmal01 bis Merkmal99.
- Es werden 100-999 erstellte Merkmale benannt: Merkmal001 bis Merkmal999 und so weiter.

So wird gewährleistet, dass die erstellten Merkmale ungeachtet ihrer Anzahl in einer vernünftigen Reihenfolge sortiert werden.

Aus Datums- und Zeitangaben berechnete Dauerzeiten. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben berechnete Dauer angewendet werden sollen.

Aus Datums- und Zeitangaben extrahierte zyklische Elemente. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben extrahierten zyklischen Elemente angewendet werden sollen.

Transformationen anwenden und speichern

Je nachdem, ob Sie die Dialogfelder für interaktive oder automatische Datenaufbereitung verwenden, weichen die Einstellungen zum Anwenden und Speichern von Transformationen leicht voneinander ab.

Interaktive Datenaufbereitung – Transformationen anwenden – Einstellungen

Transformierte Daten. Diese Einstellungen legen den Speicherort der transformierten Daten fest.

- **Neue Felder zu aktivem Dataset hinzufügen.** Alle durch die automatisierte Datenaufbereitung erstellten Felder werden dem aktiven Dataset als neue Felder hinzugefügt. Mit der Option **Rollen für analysierte Felder aktualisieren** wird die Rolle für alle Felder, die von der weiteren Analyse durch die automatisierte Datenaufbereitung ausgeschlossen werden, auf "Keine" gesetzt.
- **Neues Dataset oder Datei mit transformierten Daten erstellen.** Von der automatisierten Datenaufbereitung empfohlene Felder werden einem neuen Dataset oder einer Datei hinzugefügt. Mit der Option **Nicht analysierte Felder einschließen** werden dem Originaldataset Felder hinzugefügt, die im neuen Dataset auf der Registerkarte "Felder" nicht angegeben wurden. Das ist nützlich beim Übertragen von Feldern, die Informationen enthalten, die bei der Modellierung nicht verwendet werden, wie zum Beispiel eine ID, eine Adresse oder ein Name, in das neue Dataset.

Automatische Datenaufbereitung – Anwenden und speichern – Einstellungen

Die Gruppe "Transformierte Daten" ist dieselbe wie in der interaktiven Datenaufbereitung. Bei der automatischen Datenaufbereitung sind die folgenden zusätzlichen Optionen verfügbar:

Transformationen anwenden. Wird diese Option im Dialogfeld für die automatische Datenaufbereitung inaktiviert, werden alle anderen Befehle "Anwenden und speichern" inaktviert und die Auswahl wird beibehalten.

Transformationen als Syntax speichern. Mit dieser Option werden die empfohlenen Transformationen als Befehlssyntax in eine externe Datei gespeichert. Das Dialogfeld "Interaktive Datenaufbereitung" enthält diese Steuerung nicht, da es die Transformationen als Befehlssyntax in das Syntaxfenster einfügt, wenn Sie auf **Einfügen** klicken.

Transformationen als XML speichern. Mit dieser Option werden die empfohlenen Transformationen als XML in einer externen Datei gespeichert, die mithilfe von `TMS MERGE` mit der Modell-PMML zusammengeführt oder mithilfe von `TMS IMPORT` auf ein anderes Dataset angewendet werden kann. Das Dialogfeld "Interaktive Datenaufbereitung" enthält diese Steuerung nicht, da es die Transformationen als XML speichert, wenn Sie in der Symbolleiste im oberen Bereich des Dialogfelds auf **XML speichern** klicken.

Registerkarte "Analyse"

Hinweis: Die Registerkarte "Analyse" wird in der interaktiven Datenaufbereitung verwendet, damit Sie die empfohlenen Transformationen überprüfen können. Das Dialogfeld "Automatische Datenaufbereitung" enthält diesen Schritt nicht.

1. Wenn Sie mit den ADP-Einstellungen einschließlich aller in den Registerkarten "Ziel", "Felder" und "Einstellungen" vorgenommenen Änderungen zufrieden sind, klicken Sie auf **Daten analysieren**. Der Algorithmus wendet die Eingabedaten an und zeigt die Ergebnisse auf der Registerkarte "Analyse" an.

Die Registerkarte "Analyse" enthält Ausgaben in Grafik- und Tabellenform, die die Verarbeitung Ihrer Daten zusammenfassen, und zeigt Empfehlungen an, wie die Daten möglicherweise bearbeitet oder für das Scoring verbessert werden können. Anschließend können Sie diese Empfehlungen überprüfen und entweder akzeptieren oder ablehnen.

Die Registerkarte "Analyse" besteht aus zwei Bereichen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich. Es gibt drei Hauptansichten:

- Feldverarbeitungsübersicht (Standard). Weitere Informationen finden Sie im Thema „Feldverarbeitungsübersicht".
- Felder. Weitere Informationen finden Sie im Thema „Felder" auf Seite 19.
- Aktionsübersicht. Weitere Informationen finden Sie im Thema „Aktionsübersicht" auf Seite 20.

Es gibt vier verknüpfte/Hilfsansichten:

- Vorhersagekraft (Standard). Weitere Informationen finden Sie im Thema „Vorhersagekraft" auf Seite 20.
- Feldertabelle. Weitere Informationen finden Sie im Thema „Feldertabelle" auf Seite 20.
- Felddetails. Weitere Informationen finden Sie im Thema „Felddetails" auf Seite 21.
- Aktionsdetails. Weitere Informationen finden Sie im Thema „Aktionsdetails" auf Seite 22.

Verknüpfungen zwischen Ansichten

In der Hauptansicht steuert unterstrichener Text in den Tabellen die Anzeige in der verknüpften Ansicht. Wenn Sie auf den Text klicken, erhalten Sie Informationen über ein bestimmtes Feld, ein Set von Feldern oder einen Verarbeitungsschritt. Der zuletzt von Ihnen ausgewählte Link wird in einer dunkleren Farbe angezeigt. Dies hilft Ihnen dabei, die Verbindung zwischen den Inhalten der beiden Ansichtsbereiche zu identifizieren.

Zurücksetzen der Ansichten

Klicken Sie auf **Zurücksetzen** im unteren Bereich der Hauptansicht, um die ursprünglichen Empfehlungen der Analyse erneut anzuzeigen und alle in den Analyseansichten vorgenommenen Änderungen rückgängig zu machen.

Feldverarbeitungsübersicht

Die Tabelle "Feldverarbeitungsübersicht" gibt Ihnen eine Momentaufnahme des projizierten Gesamteinflusses der Verarbeitung, einschließlich Änderungen des Status der Funktionen und der Anzahl der erstellten Funktionen.

Beachten Sie, dass dabei kein Modell erstellt wird und somit kein Maß oder keine Grafik der Veränderung der Gesamtvorhersagekraft vor und nach der Datenaufbereitung vorhanden ist. Sie können stattdessen Grafiken der Vorhersagekraft einzelner empfohlener Prädiktoren anzeigen.

Die Tabelle zeigt folgende Informationen an:

- Die Anzahl der Zielfelder.
- Die Anzahl der ursprünglichen Prädiktoren (Eingabeprädiktoren).
- Die für die Analyse und die Modellierung empfohlenen Prädiktoren. Dazu zählen die Gesamtanzahl der empfohlenen Felder, die Anzahl der empfohlenen ursprünglichen untransformierten Felder, die Anzahl der empfohlenen transformierten Felder (ausgenommen Zwischenversionen von Feldern, aus Prädiktoren für Datum/Zeit abgeleitete Felder und konstruierte Prädiktoren), die Anzahl der empfohlenen Felder, die aus Datums-/Zeitfeldern abgeleitet sind, und die Anzahl der empfohlenen konstruierten Prädiktoren.
- Die Anzahl der Eingabeprädiktoren, die in keiner Form empfohlen werden, sei es in ihrer ursprünglichen Form, als abgeleitetes Feld oder als Eingabe für einen konstruierten Prädiktor.

Klicken Sie auf die unterstrichenen Informationen unter **Felder**, um weitere Informationen in einer verknüpften Ansicht anzuzeigen. In der verknüpften Ansicht "Feldertabelle" erhalten Sie Informationen über **Ziel**, **Eingabemerkmale** und **Nicht verwendete Eingabemerkmale**. Weitere Informationen finden Sie im Thema „Feldertabelle” auf Seite 20. **Empfohlene Merkmale für den Einsatz in Analysen** werden in der verknüpften Ansicht "Vorhersagekraft" angezeigt. Weitere Informationen finden Sie im Thema „Vorhersagekraft” auf Seite 20.

Felder

In der Hauptansicht "Felder" werden die verarbeiteten Felder angezeigt sowie, ob ADP diese zur Verwendung in nachgelagerten Modellen empfiehlt. Sie können die Empfehlung für jedes Feld überschreiben, zum Beispiel, um erstellte Merkmale auszuschließen oder Merkmale einzuschließen, von denen ADP empfiehlt, sie auszuschließen. Wenn ein Feld transformiert wurde, können Sie entscheiden, ob Sie die vorgeschlagene Transformation akzeptieren oder die Originalversion verwenden möchten.

Die Feldansicht besteht aus zwei Tabellen, eine für das Ziel und eine für Prädiktoren, die entweder verarbeitet oder erstellt wurden.

Tabelle "Ziel"

Die Tabelle **Ziel** wird nur angezeigt, wenn in den Daten ein Ziel definiert wurde.

Die Tabelle enthält zwei Spalten:

- **Name.** Dies ist der Name oder die Beschriftung des Zielfelds. Der Originalname wird immer verwendet, auch wenn das Feld transformiert wurde.
- **Messniveau.** Hier wird das Symbol für das entsprechende Messniveau angezeigt. Bewegen Sie die Maus über das Symbol, um eine Beschriftung (kontinuierlich (stetig), ordinal, nominal usw.) anzuzeigen, die die Daten beschreibt.

  Wenn das Ziel transformiert wurde, gibt die Spalte **Messniveau** die endgültige transformierte Version an. *Hinweis*: Transformationen für das Ziel können nicht inaktiviert werden.

Registerkarte "Prädiktoren"

Die Tabelle **Prädiktoren** wird immer angezeigt. Jede Zeile der Tabelle repräsentiert ein Feld. Standardmäßig sind die Zeilen nach absteigender Vorhersagekraft sortiert.

Bei gewöhnlichen Funktionen wird der Originalname immer als Zeilenname verwendet. Sowohl Original- als auch abgeleitete Versionen von Datums-/Zeitfeldern werden in der Tabelle (in getrennten Zeilen) angezeigt. Die Tabelle enthält auch konstruierte Prädiktoren.

Beachten Sie, dass transformierte Versionen von in der Tabelle angezeigten Feldern immer die Endversionen darstellen.

Standardmäßig werden in der Tabelle "Prädiktoren" nur empfohlene Felder angezeigt. Um die restlichen Felder anzuzeigen, wählen Sie das Feld **Nicht empfohlene Felder in Tabelle einschließen** über der Tabelle aus. Diese Felder werden dann am Ende der Tabelle angezeigt.

Die Tabelle enthält folgende Spalten:

- **Zu verwendende Version.** Hier wird eine Dropdown-Liste angezeigt, die festlegt, ob ein Feld nachgelagert verwendet wird oder ob die vorgeschlagenen Transformationen verwendet werden sollen. Standardmäßig werden in der Dropdown-Liste die Empfehlungen wiedergegeben.

  Für gewöhnliche Prädiktoren, die transformiert wurden, stehen in der Dropdown-Liste drei Optionen zur Auswahl: **Transformiert**, **Original** und **Nicht verwenden**.

  Für nicht transformierte gewöhnliche Prädiktoren sind folgende Auswahlmöglichkeiten verfügbar: **Original** und **Nicht verwenden**.

Für abgeleitete Datums-/Zeitfelder und konstruierte Prädiktoren sind folgende Auswahlmöglichkeiten verfügbar: **Transformiert** und **Nicht verwenden**.

Für Originaldatumsfelder ist die Dropdown-Liste inaktiviert und auf **Nicht verwenden** gesetzt.

Hinweis: Für Prädiktoren mit Original- und transformierten Versionen werden bei einem Wechsel zwischen den Versionen **Original** und **Transformiert** automatisch die Einstellungen **Messniveau** und **Vorhersagekraft** für diese Funktionen aktualisiert.

- **Name.** Jeder Feldname ist ein Link. Klicken Sie auf den Namen, um in der verknüpften Ansicht weitere Informationen über das Feld anzuzeigen. Weitere Informationen finden Sie im Thema „Felddetails" auf Seite 21.
- **Messniveau.** Hier erscheint das Symbol für den entsprechenden Datentyp. Bewegen Sie die Maus über das Symbol, um eine Beschriftung (kontinuierlich (stetig), ordinal, nominal usw.) anzuzeigen, die die Daten beschreibt.
- **Vorhersagekraft.** Die Vorhersagekraft wird nur für Felder angezeigt, die von ADP empfohlen werden. Diese Spalte wird nicht angezeigt, wenn kein Ziel definiert wurde. Die Vorhersagekraft reicht von 0 bis 1, wobei größere Werte "bessere" Prädiktoren andeuten. Im Allgemeinen ist die Vorhersagekraft für den Vergleich von Prädiktoren in einer ADP-Analyse nützlich, doch sollten Vorhersagekraftwerte nicht in Analysen verglichen werden.

Aktionsübersicht

Bei jeder von der automatisierten Datenaufbereitung vorgenommenen Aktion werden Eingabeprädiktoren transformiert und/oder herausgefiltert. Felder, die in einer Aktion erhalten bleiben, werden in der nächsten verwendet. Die Felder, die bis zum letzten Schritt erhalten bleiben, werden dann für die Modellierung empfohlen, während Eingaben zu transformierten und konstruierten Prädiktoren durch Filterung ausgeschlossen werden.

Die Aktionsübersicht ist eine einfache Tabelle, in der die von der ADP vorgenommenen Verarbeitungsaktionen aufgelistet sind. Klicken Sie auf den unterstrichenen Link **Aktion**, um in einer verknüpften Ansicht weitere Informationen über die durchgeführten Schritte anzuzeigen. Weitere Informationen finden Sie im Thema „Aktionsdetails" auf Seite 22.

Hinweis: Es werden nur die Originalversionen und die endgültigen transformierten Versionen jedes Felds angezeigt, jedoch keine während der Analyse verwendeten Zwischenversionen.

Vorhersagekraft

Wird standardmäßig bei der ersten Ausführung der Analyse angezeigt. Wenn Sie dagegen **Empfohlene Prädiktoren für den Einsatz in Analysen** in der Hauptansicht "Feldverarbeitungsübersicht" auswählen, zeigt das Diagramm die Vorhersagekraft der empfohlenen Prädiktoren an. Felder werden nach Vorhersagekraft sortiert, wobei das Feld mit dem höchsten Wert zuerst erscheint.

Bei transformierten Versionen gewöhnlicher Prädiktoren gibt der Feldname Ihre Suffixauswahl im Bereich "Feldnamen" auf der Registerkarte "Einstellungen" an, zum Beispiel: *_transformiert*.

Symbole für das Messniveau werden nach den einzelnen Feldnamen angezeigt.

Die Vorhersagekraft jedes empfohlenen Prädiktors wird entweder aus einer linearen Regression oder einem Naïve Bayes-Modell berechnet, abhängig davon, ob das Ziel stetig oder kategorial ist.

Feldertabelle

Die Feldertabelle wird angezeigt, wenn Sie in der Hauptansicht "Feldverarbeitungsübersicht" auf **Ziel**, **Prädiktoren** oder **Nicht verwendete Prädiktoren** klicken, und enthält eine einfache Tabelle, die die wichtigsten Prädiktoren auflistet.

Die Tabelle enthält zwei Spalten:

- **Name.** Der Name des Prädiktors.

  Für Ziele wird der Originalname oder die Originalbeschriftung des Felds verwendet, selbst wenn das Ziel transformiert wurde.

  Bei transformierten Versionen gewöhnlicher Prädiktoren gibt der Name Ihre Suffixauswahl im Bereich "Feldnamen" auf der Registerkarte "Einstellungen" an, zum Beispiel: *_transformiert*.

  Bei aus Datums- und Zeitangaben abgeleiteten Feldern wird der Name der endgültigen transformierten Version verwendet, zum Beispiel: *gebdat_jahre*.

  Bei konstruierten Prädiktoren wird der Name des konstruierten Prädiktors verwendet, zum Beispiel: *Prädiktor1*.
- **Messniveau.** Hier erscheint das Symbol für den entsprechenden Datentyp.

  Für das Ziel gibt das **Messniveau** stets die transformierte Version wieder (wenn das Ziel transformiert wurde), zum Beispiel bei einem Wechsel von ordinal (sortiertes Set) zu stetig (Bereich, Skala) oder umgekehrt.

Felddetails

Die Ansicht "Felddetails" wird angezeigt, wenn Sie auf **Name** in der Hauptansicht "Felder" klicken, und enthält Informationen über Verteilung, fehlende Werte und (falls zutreffend) Vorhersagekraftdiagramme für das ausgewählte Feld. Außerdem wird der Verarbeitungsverlauf für das Feld und der Name des transformierten Felds angezeigt (falls zutreffend).

Für jedes Diagrammset werden nebeneinander zwei Versionen angezeigt, um das Feld mit und ohne angewendete Transformationen zu vergleichen. Wenn keine transformierte Version des Felds vorhanden ist, wird nur ein Diagramm für die Originalversion angezeigt. Bei abgeleiteten Datums- und Zeitfeldern und konstruierten Prädiktoren werden die Diagramme nur für den neuen Prädiktor angezeigt.

Hinweis: Wenn ein Feld wegen zu vieler Kategorien ausgeschlossen wurde, wird nur der Verarbeitungsverlauf angezeigt.

Verteilungsdiagramm

Die Verteilung stetiger Felder wird als Histogramm angezeigt, mit einer überlagerten Normalverteilungskurve und einer vertikalen Bezugslinie für den Mittelwert. Kategoriale Felder werden als Balkendiagramm angezeigt.

Die Histogramme werden nach Standardabweichung und Schiefe beschriftet, allerdings wird Letztere nicht angezeigt, wenn die Anzahl der Werte kleiner gleich 2 oder die Varianz des originalen Felds kleiner als 10-20 ist.

Bewegen Sie die Maus über das Diagramm, um entweder den Mittelwert für Histogramme oder die Zählung und den Prozentsatz der Gesamtzahl der Datensätze für Kategorien in Balkendiagrammen anzuzeigen.

Diagramm fehlender Werte

Kreisdiagramme vergleichen den Prozentsatz fehlender Werte mit und ohne angewendete Transformationen; die Diagrammbeschriftungen zeigen den Prozentsatz an.

Wenn ADP die Behandlung fehlender Werte durchgeführt hat, enthält das Kreisdiagramm nach der Transformation auch den Ersatzwert als Beschriftung, d. h. den anstelle von fehlenden Werten verwendeten Wert.

Bewegen Sie die Maus über das Diagramm, um die Zählung der fehlenden Werte und den Prozentsatz der Gesamtzahl an Datensätzen anzuzeigen.

Vorhersagekraftdiagramm

Für empfohlene Felder zeigen Balkendiagramme die Vorhersagekraft vor und nach der Transformation an. Wenn das Ziel transformiert wurde, steht die berechnete Vorhersagekraft in Beziehung zum transformierten Ziel.

Hinweis: Die Vorhersagekraftdiagramme werden nicht angezeigt, wenn kein Ziel definiert wurde oder wenn Sie in der Hauptansicht auf das Ziel klicken.

Bewegen Sie die Maus über das Diagramm, um den Wert der Vorhersagekraft anzuzeigen.

Tabelle "Verarbeitungsverlauf"

Die Tabelle zeigt, wie die transformierte Version eines Felds abgeleitet wurde. Von ADP durchgeführte Aktionen werden in der Reihenfolge ihrer Ausführung aufgelistet. Bei bestimmten Schritten wurden jedoch unter Umständen mehrere Aktionen für ein spezielles Feld durchgeführt.

Hinweis: Die Tabelle wird nur für transformierte Felder angezeigt.

Die Informationen in der Tabelle sind in zwei oder in drei Spalten untergliedert:

- **Aktion.** Der Name der Aktion. Zum Beispiel "Stetige Prädiktoren". Weitere Informationen finden Sie im Thema „Aktionsdetails”.
- **Details.** Die Liste der durchgeführten Verarbeitung. Zum Beispiel "Zu Standardeinheiten transformieren".
- **Funktion.** Diese Spalte erscheint nur bei konstruierten Prädiktoren und zeigt die lineare Kombination von Eingabefeldern an, zum Beispiel 0,06*Alter + 1,21*Größe.

Aktionsdetails

Die verknüpfte Ansicht "Aktionsdetails" wird angezeigt, wenn Sie in der Hauptansicht "Aktionsübersicht" auf den unterstrichenen Link **Aktion** klicken, und enthält sowohl aktionsspezifische als auch allgemeine Informationen über jeden durchgeführten Verarbeitungsschritt. Die aktionsspezifischen Informationen erscheinen stets zuerst.

Für jede Aktion wird die Beschreibung als Titel im oberen Bereich der verknüpften Ansicht verwendet. Die aktionsspezifischen Informationen werden unter dem Titel angezeigt und enthalten gegebenenfalls Details zur Anzahl der abgeleiteten Prädiktoren, zu umgewandelten Feldern, zu Zieltransformationen, zu zusammengeführten oder neu sortierten Kategorien und zu konstruierten oder ausgeschlossenen Prädiktoren.

Bei der Verarbeitung jeder Aktion kann sich die für die Verarbeitung verwendete Anzahl an Prädiktoren ändern, wenn beispielsweise Prädiktoren ausgeschlossen oder zusammengeführt werden.

Hinweis: Wenn eine Aktion inaktiviert oder kein Ziel angegeben wurde, wird anstelle der Aktionsdetails eine Fehlernachricht angezeigt, wenn Sie in der Hauptansicht "Aktionsübersicht" auf die Aktion klicken.

Es gibt neun mögliche Aktionen, davon sind allerdings nicht alle notwendigerweise für jede Analyse aktiv.

Tabelle "Textfelder"

Die Tabelle zeigt folgende Anzahl:

- Von der Analyse ausgeschlossene Prädiktoren.

Tabelle "Prädiktoren für Datum und Uhrzeit"

Die Tabelle zeigt folgende Anzahl:

- Aus Variablen für Datum und Uhrzeit abgeleitete Dauer.
- Datums- und Uhrzeitelemente.
- Insgesamt abgeleitete Prädiktoren für Datum und Uhrzeit.

Das Referenzdatum oder die -uhrzeit wird als Fußnote angezeigt, wenn eine Datumsdauer berechnet wurde.

Tabelle "Prädiktorscreening"

Die Tabelle zeigt die Anzahl folgender von der Verarbeitung ausgeschlossener Prädiktoren:

- Konstanten.
- Prädiktoren mit zu vielen fehlenden Werten.
- Prädiktoren mit zu vielen Fällen in einer einzelnen Kategorie.
- Nominale Felder (Sets) mit zu vielen Kategorien.
- Insgesamt ausgeschlossene Prädiktoren.

Tabelle"Messniveau prüfen"

Die Tabelle zeigt die Anzahl umgewandelter Felder und teilt sich wie folgt auf:

- In stetige Feldern umgewandelte ordinale Felder (sortierte Sets).
- In ordinale Felder umgewandelte stetige Felder.
- Anzahl an Umwandlungen insgesamt.

Wenn keine Eingabefelder (Ziel oder Prädiktoren) stetig (kontinuierlich) oder ordinal waren, wird dies als Fußnote vermerkt.

Tabelle "Ausreißer"

Die Tabelle zeigt, ob und wie Ausreißer behandelt wurden.

- Entweder die Anzahl stetiger Felder, für die Ausreißer gefunden und entfernt wurden, oder die Anzahl stetiger Felder, für die Ausreißer gefunden und als fehlend eingestuft wurden, je nach Ihren Einstellungen im Feld "Eingaben & Ziel vorbereiten" auf der Registerkarte "Einstellungen".
- Die Anzahl stetiger Felder, die ausgeschlossen wurden, weil sie nach der Ausreißerbehandlung konstant waren.

Der Ausreißertrennwert wird in einer Fußnote vermerkt. Eine weitere Fußnote wird angezeigt, wenn keine Eingabefelder (Ziel oder Prädiktoren) stetig (kontinuierlich) waren.

Tabelle "Fehlende Werte"

Die Tabelle zeigt die Anzahl an Feldern, in denen fehlende Werte ersetzt wurden, und teilt sich wie folgt auf:

- Ziel. Diese Zeile wird nicht angezeigt, wenn kein Ziel angegeben wurde.
- Prädiktoren. Dies teilt sich weiter auf in Anzahl an "nominal (Set)", "ordinal (sortiertes Set)" und "stetig".
- Die gesamte Anzahl ersetzter fehlender Werte.

Tabelle "Ziel"

Die Tabelle zeigt wie folgt, ob das Ziel tranformiert wurde:

- Box-Cox-Transformation in Normalverteilung. Dies teilt sich weiter in Spalten auf, die die angegebenen Kriterien (Mittelwert und Standardabweichung) und Lambda zeigen.
- Zielkategorien zur Verbesserung der Stabilität neu sortiert.

Tabelle "Kategoriale Prädiktoren"

Die Tabelle zeigt folgende Anzahl kategorialer Prädiktoren:
- Wessen Kategorien wurden zur Verbesserung der Stabilität in aufsteigender Reihenfolge neu sortiert.
- Wessen Kategorien wurden zur Maximierung des Zielzusammenhangs zusammengeführt.
- Wessen Kategorien wurden zur Behandlung dünn besetzter Kategorien zusammengeführt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Zusammenführung konstant.

Wenn es keine kategorialen Prädiktoren gab, wird dies durch eine Fußnote vermerkt.

Tabelle "Stetige Prädiktoren"

Es gibt zwei Tabellen. Die erste zeigt eine der folgenden Transformationen:
- Zu Standardeinheiten transformierte Prädiktorwerte. Zusätzlich werden hier die Anzahl transformierter Prädiktoren, der angegebene Mittelwert und die Standardabweichung angezeigt.
- Einem gemeinsamen Bereich zugeordnete Prädiktorwerte. Zusätzlich werden hier die Anzahl der mithilfe einer **Min./Max. Transformation** transformierten Prädiktoren sowie die angegebenen Mindest- und Höchstwerte angezeigt.
- Klassierte Prädiktorwerte und die Anzahl klassierter Prädiktoren.

Die zweite Tabelle enthält Informationen über die Prädiktorerstellung, die als Anzahl folgender Prädiktoren angezeigt werden:
- Erstellt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Klassierung konstant.
- Ausgeschlossen, weil nach der Erstellung konstant.

Wenn keine stetigen (kontinuierlichen) Prädiktoren eingegeben wurden, wird dies durch eine Fußnote vermerkt.

Scores zurücktransformieren

Wenn ein Ziel durch ADP transformiert wurde, bewerten nachfolgend mithilfe des transformierten Zielscores erstellte Modelle die transformierten Einheiten. Um die Ergebnisse interpretieren und verwenden zu können, müssen Sie den vorhergesagten Wert wieder in das ursprüngliche metrische Maß zurückkonvertieren.

1. Wählen Sie in den Menüs Folgendes aus, um Scores zurückzutransformieren:

   Transformieren > **Daten für Modellierung vorbereiten** > **Scores zurücktransformieren...**
2. Wählen Sie ein Feld aus, das zurücktransformiert werden soll. Dieses Feld sollte vom Modell vorhergesagte Werte des transformierten Ziels enthalten.
3. Geben Sie ein Suffix für das neue Feld an. Dieses neue Feld enthält vom Modell vorhergesagte Werte im ursprünglichen metrischen Maß des nicht transformierten Ziels.
4. Geben Sie den Speicherort der XML-Datei mit den ADP-Transformationen an. Es sollte eine Datei sein, die aus den Dialogfeldern für interaktive oder automatische Datenaufbereitung heraus gespeichert wurde. Weitere Informationen finden Sie im Thema „Transformationen anwenden und speichern” auf Seite 17.

Kapitel 5. Ungewöhnliche Fälle identifizieren

Die Prozedur "Anomalieerkennung" sucht anhand von Abweichungen von den Normwerten der Gruppe nach ungewöhnlichen Fällen. Die Prozedur wurde für die Datenprüfung in der explorativen Datenanalyse konzipiert. Zweck der Prozedur ist das schnelle Erkennen von ungewöhnlichen Fällen, bevor mit anderen Analysen Schlüsse aus den Daten gezogen werden. Dieser Algorithmus dient der Erkennung von allgemeinen Anomalien. Dies bedeutet, dass sich die Definition eines anomalen Falls nicht auf eine bestimmte Anwendung beschränkt, bei der Anomalien sehr treffend definiert werden können, z. B. beim Erkennen von ungewöhnlichen Zahlungsmustern im Gesundheitswesen oder beim Aufdecken von Geldwäsche im Finanzwesen.

Beispiel. Ein Analytiker, der mit der Erstellung von Vorhersagemodellen für die Ergebnisse von Schlaganfallbehandlungen betraut wurde, ist über die Qualität der Daten besorgt, weil solche Modelle bei ungewöhnlichen Beobachtungen anfällig sein können. Einige dieser Randbeobachtungen stellen wirklich einzigartige Fälle dar und eignen sich deswegen nicht für eine Vorhersage. Andere Beobachtungen stellen Dateneingabefehler dar, wobei die Werte technisch gesehen "richtig" sind und deswegen nicht mit Datenvalidierungsprozeduren abgefangen werden können. Die Prozedur "Ungewöhnliche Fälle identifizieren" sucht Ausreißer und meldet diese, sodass der Analytiker entscheiden kann, wie mit diesen Fällen verfahren wird.

Statistiken. Die Prozedur erzeugt Peergruppen, Normwerte für Peergruppen bei stetigen und kategorialen Variablen, Anomalieindizes auf der Grundlage von Abweichungen von den Normwerten der Gruppen sowie Variableneinflusswerte für Variablen, die am meisten dazu beitragen, dass ein Falls als ungewöhnlich klassifiziert wird.

Erläuterung der Daten

Daten. Mit dieser Prozedur können sowohl stetige als auch kategoriale Variablen analysiert werden. Jede Zeile stellt eine eindeutige Beobachtung und jede Zeile eine eindeutige Variable als Grundlage für die Peergruppen dar. In der Datendatei kann eine Fall-ID-Variable zum Markieren der Ausgabe verfügbar sein. Diese Variable wird jedoch nicht in der Analyse verwendet. Fehlende Werte sind zulässig. Wenn die Gewichtungsvariable angegeben wurde, wird diese ignoriert.

Das Erkennungsmodell kann auf eine neue Testdatendatei angewendet werden. Die Elemente der Testdaten müssen dieselben wie die Elemente der Lerndaten sein. Abhängig von den Einstellungen des Algorithmus kann die Verarbeitung fehlender Werte, die beim Erstellen des Modells verwendet wird, vor dem Scoring auf die Testdaten angewendet werden.

Fallreihenfolge. Beachten Sie, dass die Lösung von der Fallreihenfolge abhängen kann. Um die Auswirkungen der Reihenfolge zu minimieren, mischen Sie die Fälle in zufälliger Reihenfolge. Prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie verschiedene Lösungen abrufen, bei denen die Fälle in einer unterschiedlichen, zufällig ausgewählten Reihenfolgen sortiert sind. In Situationen mit extrem umfangreichen Dateien können mehrere Durchgänge mit jeweils einer Stichprobe von Fällen durchgeführt werden, die in unterschiedlicher, zufällig ausgewählter Reihenfolge sortiert ist.

Annahmen. Der Algorithmus setzt voraus, dass alle Variablen nicht konstant und unabhängig sind. Es wird außerdem angenommen, dass kein Fall bei einer Eingabevariablen fehlende Werte aufweist. Für alle stetigen Variablen wird eine Normalverteilung (Gauß-Verteilung) und für alle kategorialen Variablen eine multinomiale Verteilung vorausgesetzt. Empirische interne Tests zeigen, dass die Prozedur wenig anfällig gegenüber Verletzungen hinsichtlich der Unabhängigkeitsannahme und der Verteilungsannahme ist. Dennoch sollten Sie darauf achten, wie genau diese Voraussetzungen erfüllt sind.

So identifizieren Sie ungewöhnliche Fälle:

1. Wählen Sie in den Menüs Folgendes aus:
   Daten > **Ungewöhnliche Fälle identifizieren...**
2. Wählen Sie mindestens eine Analysevariable aus.
3. Wahlweise können Sie eine Fall-ID-Variable zum Beschriften der Ausgabe auswählen.

Felder mit unbekanntem Messniveau

Der Messniveau-Alert wird angezeigt, wenn das Messniveau für mindestens eine Variable (ein Feld) im Dataset unbekannt ist. Da sich das Messniveau auf die Berechnung der Ergebnisse für diese Prozedur auswirkt, müssen alle Variablen ein definiertes Messniveau aufweisen.

Daten durchsuchen. Liest die Daten im aktiven Dataset und weist allen Feldern, deren Messniveau zurzeit nicht bekannt ist, das Standardmessniveau zu. Bei großen Datasets kann dieser Vorgang einige Zeit in Anspruch nehmen.

Manuell zuweisen. Öffnet ein Dialogfeld, in dem alle Felder mit unbekanntem Messniveau aufgeführt werden. Mit diesem Dialogfeld können Sie diesen Feldern ein Messniveau zuweisen. Außerdem können Sie in der Variablenansicht des Dateneditors ein Messniveau zuweisen.

Da das Messniveau für diese Prozedur bedeutsam ist, können Sie erst dann auf das Dialogfeld zur Ausführung dieser Prozedur zugreifen, wenn für alle Felder ein Messniveau definiert wurde.

Ungewöhnliche Fälle identifizieren: Ausgabe

Liste ungewöhnlicher Fälle und Ursachen für die Ungewöhnlichkeit. Bei dieser Option werden drei Tabellen erstellt:

- Die Liste der Indizes anomaler Fälle zeigt die als ungewöhnlich identifizierten Fälle und deren entsprechende Anomalieindexwerte an.
- Die Liste der Peergruppen-IDs anomaler Fälle zeigt ungewöhnliche Fälle und die Informationen über deren entsprechende Peergruppen an.
- Die Liste der Ursachen anomaler Fälle zeigt die Fallanzahl, die Ursachenvariable, den Variableneinflusswert, den Wert der Variablen und den Normwert der Variablen für jede Ursache an.

Alle Tabellen werden nach Anomalieindex in absteigender Reihenfolge sortiert. Darüber hinaus werden die IDs der Fälle angezeigt, wenn auf der Registerkarte "Variablen" eine Fall-ID-Variable angegeben wurde.

Auswertung. Mit den Steuerelementen in diesem Gruppenfeld werden Auswertungen der Verteilungen erstellt.

- **Normwerte der Peergruppen.** Bei dieser Option wird die Tabelle für die Normwerte der stetigen Variablen (wenn die Analyse stetige Variablen umfasst) und die Tabelle für die Normwerte der kategorialen Variablen (wenn die Analyse kategoriale Variable umfasst) angezeigt. Die Tabelle für die Normwerte der stetigen Variablen enthält den Mittelwert und die Standabweichung jeder stetigen Variablen für jede Peergruppe. Die Tabelle für die Normwerte der kategorialen Variablen enthält den Modalwert (die häufigste Kategorie), die Häufigkeit und die Häufigkeit in Prozent jeder kategorialen Variablen für jede Peergruppe. Der Mittelwert einer stetigen Variablen und der Modalwert einer kategorialen Variablen werden in der Analyse als Normwerte verwendet.
- **Anomalieindizes.** Die Auswertung des Anomalieindex enthält deskriptive Statistiken für die Anomalieindizes der Fälle, die als am ungewöhnlichsten identifiziert wurden.
- **Vorkommen der Ursache nach Analysevariablen.** Die Tabelle zeigt pro Ursache die Häufigkeit und die Häufigkeit in Prozent des Vorkommens jeder Variable als Ursache an. Die Tabelle führt auch deskriptive Statistiken über den Einfluss jeder Variablen auf. Wenn die maximale Anzahl von Ursachen auf der Registerkarte "Optionen" auf 0 festgelegt wurde, steht diese Option nicht zur Verfügung.

- **Verarbeitete Fälle.** Die Zusammenfassung der Fallverarbeitung enthält Häufigkeiten und Häufigkeiten in Prozent für alle Fälle im aktiven Dataset, die in die Analyse aufgenommenen und ausgeschlossenen Fälle und die Fälle in jeder Peergruppe.

Ungewöhnliche Fälle identifizieren: Speichern

Variablen speichern. Mithilfe der Steuerelemente in diesem Gruppenfeld können Sie Modellvariablen im aktiven Dataset speichern. Sie können auch festlegen, dass vorhandene Variablen ersetzt werden, deren Namen mit den zu speichernden Variablen kollidieren.

- **Anomalieindex.** Speichert für jeden Fall den Wert des Anomalieindex in einer Variablen mit dem angegebenen Namen.
- **Peergruppen.** Speichert die Peergruppen-ID, die Fallanzahl und die Größe als Prozentsatz für jeden Fall in Variablen mit dem angegebenen Stammnamen. Wenn für den Stammnamen zum Beispiel *Gruppe* angegeben wurde, werden die Variablen *GruppeID*, *GruppeGröße* und *GruppePrztGröße* erzeugt. *GruppeID* stellt die Peergruppen-ID des Falls dar, *GruppeGröße* die Gruppengröße und *GruppePrztGröße* die Gruppengröße als Prozentsatz.
- **Ursachen.** Speichert Sets von Ursachenvariablen mit dem angegebenen Stammnamen. Ein Set von Ursachenvariablen besteht aus dem Namen einer Variablen, die eine Ursache darstellt, dem Einflussmaß der Variablen, dem Variablenwert und dem Normwert. Die Anzahl der Sets hängt von der Anzahl der angeforderten Ursachen ab (angegeben auf der Registerkarte "Optionen"). Wenn als Stammname zum Beispiel *Ursache* angegeben wurde, werden die Variablen *UrsacheVar_k*, *UrsacheMaß_k*, *UrsacheWert_k* und *UrsacheNormwert_k* erzeugt, wobei *k* die *k*-te Ursache darstellt. Diese Option steht nicht zur Verfügung, wenn die Anzahl der Ursachen auf 0 festgelegt wurde.

Modelldatei exportieren. Hiermit können Sie das Modell im XML-Format speichern.

Ungewöhnliche Fälle identifizieren: Fehlende Werte

Auf der Registerkarte "Fehlende Werte" kann die Behandlung benutzerdefiniert und systemdefiniert fehlender Werte festgelegt werden.

- **Fehlende Werte aus der Analyse ausschließen.** Fälle mit fehlenden Werten werden aus der Analyse ausgeschlossen.
- **Fehlende Werte in die Analyse aufnehmen.** Fehlende Werte von stetigen Variablen werden durch deren entsprechenden Gesamtmittelwert ersetzt. Fehlende Kategorien von kategorialen Variablen werden gruppiert und als gültige Kategorie behandelt. Die verarbeiteten Variablen werden anschließend in der Analyse verwendet. Sie können die Erzeugung einer zusätzlichen Variable anfordern, die den Anteil der fehlenden Variablen in jedem Fall darstellt, und diese Variable in der Analyse verwenden.

Ungewöhnliche Fälle identifizieren: Optionen

Kriterien zum Identifizieren ungewöhnlicher Fälle. Diese Optionen bestimmen, wie viele Fälle in die Liste der Anomalien aufgenommen werden.

- **Prozentsatz der Fälle mit den höchsten Anomalieindexwerten.** Geben Sie eine positive Zahl kleiner oder gleich 100 ein.
- **Feste Anzahl von Fällen mit den höchsten Anomalieindexwerten.** Geben Sie eine positive Ganzzahl an, die kleiner oder gleich der Gesamtzahl der in der Analyse verwendeten Fälle im aktiven Dataset ist.
- **Nur Fälle identifizieren, deren Anomalieindex größer oder gleich einem Minimalwert ist.** Geben Sie eine nicht negative Zahl an. Ein Fall wird als Anomalie betrachtet, wenn sein Anomalieindex größer oder gleich dem angegebenen Trennwert ist. Diese Option wird zusammen mit den Optionen **Prozentsatz der Fälle** und **Feste Anzahl von Fällen** verwendet. Wenn Sie beispielsweise eine feste Anzahl von 50 Fällen und einen Trennwert von 2 angeben, besteht die Anomalieliste höchstens aus 50 Fällen, von denen jeder einen Anomalieindexwert größer oder gleich 2 aufweist.

Anzahl von Peergruppen. Die Prozedur sucht nach der besten Anzahl von Peergruppen zwischen dem angegebenen Minimal- und Maximalwert. Die Werte müssen positive Ganzzahlen sein, und das Minimum darf das Maximum nicht überschreiten. Wenn die angegebenen Werte gleich sind, setzt die Prozedur eine feste Anzahl von Peergruppen voraus.

Hinweis: Abhängig von der Variation in den Daten können Situationen auftreten, in denen die Daten weniger Peergruppen unterstützen können als als Minimum angegeben. In einer solchen Situation erzeugt die Prozedur eine kleinere Anzahl von Peergruppen.

Maximale Anzahl von Ursachen. Eine Ursache besteht aus dem Variableneinflussmaß, dem Variablennamen für diese Ursache, dem Wert der Variablen und dem Wert der entsprechenden Peergruppe. Geben Sie eine nicht negative Ganzzahl an. Wenn dieser Wert größer oder gleich der Anzahl der verarbeiteten Variablen ist, die in der Analyse verwendet werden, werden alle Variablen angezeigt.

Zusätzliche Funktionen beim Befehl DETECTANOMALY

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Sie können einige Variablen im aktiven Dataset aus der Analyse ausschließen, ohne dass ausdrücklich alle Analysevariablen angegebenen werden müssen (mit dem Unterbefehl `EXCEPT`).
- Sie können eine Korrektur angeben, um den Einfluss von stetigen und kategorialen Variablen auszutarieren (mit dem Schlüsselwort `MLWEIGHT` im Unterbefehl `CRITERIA`).

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

Kapitel 6. Optimale Klassierung

Die Prozedur "Optimale Klassierung" diskretisiert eine oder mehrere metrische Variablen (im Folgenden als **Klassierungseingabevariablen** bezeichnet), indem die Werte der einzelnen Variablen auf verschiedene Klassen verteilt werden. Die Klassenbildung ist in Bezug auf eine kategoriale Leitvariable optimal, die den Klassierungsvorgang "überwacht". Anstatt der ursprünglichen Datenwerte können dann die Klassen zur weiteren Analyse verwendet werden.

Beispiele. Für die Verringerung der unterschiedlichen Werte, die eine Variable annehmen kann, gibt es verschiedene Anwendungsmöglichkeiten. Hier einige Beispiele:

- Anforderungen anderer Prozeduren an die Daten. Diskretisierte Variablen können für die Verwendung in Prozeduren, bei denen kategoriale Variablen erforderlich sind, als kategorial behandelt werden. Beispielsweise müssen für die Prozedur "Kreuztabellen" alle Variablen kategorial sein.
- Datenschutz. Die Angabe von klassierten Werten anstelle der tatsächlichen Werte in Berichten kann zur Gewährleistung des Datenschutzes bei Ihren Datenquellen beitragen. Die Prozedur "Optimale Klassierung" kann eine Orientierung für die Auswahl der Klassen bieten.
- Schnellere Durchführung. Einige Prozeduren sind effizienter, wenn sie mit einer reduzierten Anzahl an unterschiedlichen Werten arbeiten. So lässt sich beispielsweise die Geschwindigkeit der multinomialen logistischen Regression durch die Verwendung diskretisierter Variablen erhöhen.
- Ermittlung vollständiger oder quasi vollständiger Datentrennung.

Optimale Klassierung im Vergleich zur visuellen Klassierung. In den Dialogfeldern von "Visuelle Klassierung" stehen Ihnen mehrere automatische Methoden zur Erstellung von Klassen ohne die Verwendung einer Leitvariablen zur Verfügung. Diese Regeln für unüberwachte Klassierung sind nützlich für die Erstellung deskriptiver Statistiken, wie beispielsweise Häufigkeitstabellen, "Optimale Klassierung" ist am besten, wenn das Endziel in der Erstellung eines Vorhersagemodells besteht.

Ausgabe. Mit dieser Prozedur werden Tabellen mit Trennwerten für die Klassen und deskriptive Statistiken für jede Klassierungseingabevariable erstellt. Zusätzlich können Sie neue Variablen im aktiven Dataset speichern, die die klassierten Werte der Klassierungseingabevariablen enthalten und die Klassierungsregeln als Befehlssyntax zur Verwendung bei der Diskretisierung neuer Daten speichern.

Erläuterungen der Daten für "Optimale Klassierung"

Daten. Bei dieser Prozedur wird davon ausgegangen, dass es sich bei den Klassierungseingabevariablen um metrische, numerische Variablen handelt. Die Leitvariable sollte kategorial sein. Es kann sich dabei um eine Zeichenfolgevariable oder eine numerische Variable handeln.

So erhalten Sie eine optimale Klassierung:

1. Wählen Sie in den Menüs Folgendes aus:
   Transformieren > **Optimale Klassierung...**
2. Wählen Sie mindestens eine Klassierungseingabevariable aus.
3. Wählen Sie eine Leitvariable aus.

Variablen, die die klassierten Datenwerte enthalten, werden nicht standardmäßig erstellt. Auf der Registerkarte Speichern können Sie diese Variablen speichern.

Optimale Klassierung: Ausgabe

Die Registerkarte "Ausgabe" steuert die Anzeige der Ergebnisse.

- **Endpunkte für Klassen.** Zeigt das Set an Endpunkten für die einzelnen Klassierungseingabevariablen an.
- **Beschreibende Statistiken für Klassierungsvariablen.** Diese Option zeigt für die einzelnen Klassierungseingabevariablen die Anzahl der Fälle mit gültigen Werten, die Anzahl der Fälle mit fehlenden Werten, die Anzahl der verschiedenen gültigen Werte sowie die Minimal- und Maximalwerte an. Für die Leitvariable zeigt diese Option die Klassenverteilung für alle zugehörigen Klassierungseingabevariablen an.
- **Modellentropie für Klassierungsvariablen.** Für jede Klassierungseingabevariable zeigt diese Option ein Maß für die Vorhersagegenauigkeit der Variablen hinsichtlich der Leitvariablen an.

Optimale Klassierung: Speichern

Variablen in aktivem Dataset speichern. In der weiteren Analyse können anstelle der ursprünglichen Variablen Variablen verwendet werden, die die klassierten Datenwerte enthalten.

Klassierungsregeln als Syntax speichern. Generiert Befehlssyntax, die für die Klassierung von anderen Datasets verwendet werden kann. Die Umcodierungsregeln beruhen auf den vom Klassierungsalgorithmus bestimmten Trennwerten.

Optimale Klassierung: Fehlende Werte

Auf der Registerkarte "Fehlende Werte" wird angegeben, ob der Umgang mit fehlenden Werten anhand eines listenweisen oder paarweisen Ausschlusses erfolgt. Benutzerdefiniert fehlende Werte werden immer als ungültig behandelt. Bei der Umcodierung der ursprünglichen Variablenwerte in eine neue Variable werden benutzerdefiniert fehlende Werte in systemdefiniert fehlende Werte umgewandelt.

- **Paarweise.** Diese Option operiert auf der Basis der einzelnen Paare aus Leitvariablen und Klassierungseingabevariablen. Die Prozedur verwendet alle Fälle mit nicht fehlenden Werten bei der Führungs- und Klassierungseingabevariablen.
- **Listenweise** Diese Option wird auf alle auf der Registerkarte "Variablen" angegebenen Variablen angewendet. Wenn bei einem Fall eine Variable fehlt, wird der gesamte Fall ausgeschlossen.

Optimale Klassierung: Optionen

Vorverarbeitung. Die "Vorklassierung" von Klassierungseingabevariablen mit vielen verschiedenen Werten kann die Verarbeitung ohne größere Qualitätseinbußen bei den endgültigen Klassen beschleunigen. Der Wert für die maximale Anzahl an Klassen stellt lediglich die Obergrenze für die Anzahl der erstellten Klassen dar. Wenn Sie also 1000 als Maximalwert angeben, eine Klassierungseingabevariable jedoch weniger als 1000 verschiedene Werte aufweist, werden so viele vorverarbeitete Klassen für die Klassierungseingabevariable erstellt wie verschiedene Klassen in der Klassierungseingabevariablen enthalten sind.

Dünn besetzte Klassen. Gelegentlich kann die Prozedur zu Klassen mit sehr wenigen Fällen führen. Mit der folgenden Strategie können diese Pseudotrennwerte gelöscht werden:

Angenommen, der Algorithmus hat für eine Variable $n_{\text{endgültig}}$ Trennwerte und daher $n_{\text{endgültig}}+1$ Klassen gefunden. Für die Klassen $i = 2,\ldots, n_{\text{endgültig}}$ (von der Klasse mit dem zweitniedrigsten Wert bis zur Klasse mit dem zweithöchsten Wert) wird Folgendes berechnet:

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

Dabei ist *sizeof(b)* die Anzahl der Fälle in der Klasse.

Wenn dieser Wert kleiner ist als der angegebene Zusammenführungsschwellenwert, wird b_i als dünn besetzt betrachtet und mit b_{i-1} oder b_{i+1} zusammengeführt, je nachdem, welche Klasse die niedrigere Klasseninformationsentropie aufweist.

Bei dieser Prozedur wird ein einzelner Durchlauf durch die Klassen vorgenommen.

Klassengrenzen Bei dieser Option wird angegeben, wie die Untergrenze eines Intervalls festgelegt wird. Da die Prozedur die Trennwerte automatisch ermittelt, ist dies weitgehend eine Frage der Präferenzen.

Erste (niedrigste) Klasse/Letzte (höchste) Klasse. Diese Optionen geben an, wie die minimalen und maximalen Trennwerte für die einzelnen Klassierungseingabevariablen festgelegt werden. Im Allgemeinen geht die Prozedur davon aus, dass die Klassierungseingabevariablen einen beliebigen Wert der reellen Zahlen annehmen können, aber wenn es theoretische oder praktische Gründe für die Begrenzung des Bereichs gibt, können Sie den gewünschten niedrigsten und/oder höchsten Wert angeben.

Zusätzliche Funktionen beim Befehl OPTIMAL BINNING

Die Befehlssyntax ermöglicht außerdem Folgendes:

- Sie können mithilfe der Methode der gleichen Häufigkeiten eine unüberwachte Klassierung durchführen (mit dem Unterbefehl `CRITERIA`).

Vollständige Informationen zur Syntax finden Sie in der Befehlssyntaxreferenz.

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

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

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

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

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

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

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

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

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.

COPYRIGHTLIZENZ:

Diese Veröffentlichung enthält Beispielanwendungsprogramme, die in Quellensprache geschrieben sind und Programmiertechniken in verschiedenen Betriebsumgebungen veranschaulichen. Sie dürfen diese Beispielprogramme kostenlos kopieren, ändern und verteilen, wenn dies zu dem Zweck geschieht, Anwendungsprogramme zu entwickeln, zu verwenden, zu vermarkten oder zu verteilen, die mit der Anwendungsprogrammierschnittstelle für die Betriebsumgebung konform sind, für die diese Beispielprogramme geschrieben werden. Diese Beispiele wurden nicht unter allen denkbaren Bedingungen getestet. Daher kann IBM die Zuverlässigkeit, Wartungsfreundlichkeit oder Funktion dieser Programme weder zusagen noch gewährleisten. Die Beispielprogramme werden ohne Wartung (auf "as-is"-Basis) und ohne jegliche Gewährleistung zur Verfügung gestellt. IBM übernimmt keine Haftung für Schäden, die durch die Verwendung der Beispielprogramme entstehen.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

© (Name Ihrer Firma) (Jahr). Teile des vorliegenden Codes wurden aus Beispielprogrammen der IBM Corporation abgeleitet.

© Copyright IBM Corp. _Jahr/Jahre angeben_. Alle Rechte vorbehalten.

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.

Index