IBM SPSS Data Preparation

Präzisere Ergebnisse mit einer besseren Datenaufbereitung

Highlights

- Aufspüren verdächtiger oder ungültiger Fälle, Variablen und Datenwerte
- Anzeigen von Mustern fehlender Daten
- Zusammenfassen von Variablenverteilungen
- Präzisere, schnellere Datenaufbereitung für die Analyse

Alle Forscher müssen ihre Daten vor der Analyse aufbereiten. IBM SPSS Statistics umfasst bereits Tools zur Datenaufbereitung. Manchmal benötigen Sie dazu jedoch stärker spezialisierte Verfahren. Mit IBM SPSS Data Preparation können Sie ohne großen Aufwand verdächtige oder ungültige Fälle, Variablen und Datenwerte ermitteln. Außerdem können Sie Muster fehlender Daten anzeigen, Variablenverteilungen zusammenfassen und präziser mit Algorithmen arbeiten, die für nominale Attribute vorgesehen sind. Dadurch wird der Datenaufbereitungsprozess so optimiert, dass Sie schneller zur Analyse bereit sind und genauere Schlussfolgerungen ziehen können. Sie können eine der Prozeduren für automatisierte Datenaufbereitung auswählen, um Ergebnisse am schnellsten zu erzielen, oder eines der weiteren Verfahren nutzen, mit denen Sie komplexere Datasets verarbeiten können.

SPSS Data Preparation ist als reine Client-Software verfügbar. Für eine höhere Leistung und Skalierbarkeit ist darüber hinaus eine serverbasierte Version verfügbar.

Auswahl aus mehreren Datenaufbereitungsoptionen

Datenvalidierungsprozedur

Die Datenvalidierung war bisher in der Regel ein manueller Prozess. Sie können für die Daten eine Häufigkeitsprüfung durchführen, die Häufigkeiten ausgeben, zu korrigierende Daten eingrenzen und eine Prüfung auf Fallkennungen durchführen. Dies ist zeitintensiv. Wenn zudem jeder Analyst im Unternehmen ein geringfügig anderes Verfahren anwendet, kann die projektübergreifende Konsistenz problematisch sein.

Wenn Sie keine manuellen Überprüfungen durchführen möchten, können Sie die Datenvalidierungsprozedur (VALIDATEDATA) verwenden. Mit dieser Prozedur können Sie Regeln anwenden, um Datenprüfungen auf der Basis der Messniveaus der einzelnen Variablen (kategorial oder kontinuierlich) durchzuführen. Wenn Sie zum Beispiel Umfragedaten mit Variablen auf einer Fünfpunkteskala nach Likert analysieren, können Sie mithilfe der Datenvalidierungsprozedur eine Regel für Fünfpunkteskalen anwenden und alle Fälle mit Werten markieren, die außerhalb des Bereichs von 1 bis 5 liegen. Sie können Berichte zu ungültigen Fällen sowie Zusammenfassungen von Regelverstößen und zur Anzahl der betroffenen Fälle erhalten. Außerdem können Sie Validierungsregeln für einzelne Variablen angeben (zum Beispiel Bereichsprüfungen) und variablenübergreifend prüfen (zum Beispiel auf „schwanger und männlich“).

Auf der Grundlage dieses Wissens können Sie die Gültigkeit der Daten bestimmen und vor der Analyse verdächtige Fälle nach Ihrem eigenen Ermessen entfernen oder korrigieren.

Automatische Datenaufbereitung in einem einzigen Schritt

Die manuelle Datenaufbereitung ist ein komplexer Prozess, für den ein Analyst, bezogen auf ein Projekt, 40 bis 90 Prozent seiner Zeit aufwenden muss. Wenn schnelle Ergebnisse gefragt sind, erleichtert die Prozedur zur automatisierten Datenaufbereitung (ADP – Automated Data Preparation) das Erkennen und Korrigieren von Qualitätsfehlern und das Imputieren fehlender Werte in einem einzigen effizienten Schritt. Die Funktion ADP liefert einen leicht verständlichen Bericht mit vollständigen Empfehlungen und Visualisierungen zum einfacheren Ermitteln der in der Analyse zu verwendenden Daten.

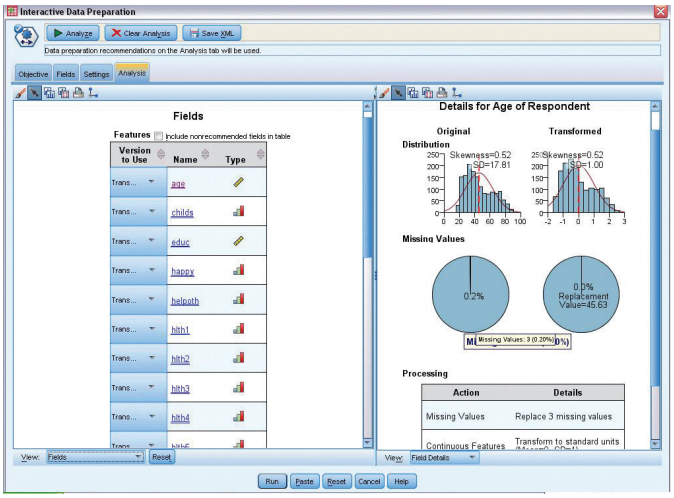

Abbildung 1: Die Funktion für die automatisierte Datenaufbereitung (ADP – Automated Data Preparation) stellt Empfehlungen bereit und ermöglicht es Anwendern, die Empfehlungen detailliert zu analysieren und zu untersuchen.

Prozedur zur Erkennung von Anomalien

Mithilfe der Prozedur zur Erkennung von Anomalien können Sie verhindern, dass Ausreißer zu einer ungleichen Verteilung bei Analysen führen. Diese Funktion sucht auf der Basis der Abweichungen von ähnlichen Fällen nach ungewöhnlichen Fällen und liefert Ursachen für derartige Abweichungen. Sie können Ausreißer markieren, indem Sie eine neue Variable erstellen. Nachdem Sie ungewöhnliche Fälle ermittelt haben, können Sie diese weiter untersuchen und festlegen, ob sie mitanalysiert werden sollen.

Optimales Binning

Damit für nominale Attribute konzipierte Algorithmen (zum Beispiel der Naïve-Bayes-Algorithmus und Logit-Modelle) verwendet werden können, müssen die Skalenvariablen vor der Modellerstellung einer Klasse zugeordnet werden. Falls die Skalenvariablen keiner Klasse zugeordnet sind, dauert die Verarbeitung von Algorithmen wie der multinomialen logistischen Regression (MLR) sehr lange oder sie konvergieren möglicherweise nicht (besonders bei einem umfangreichen Dataset). Darüber hinaus sind die Ergebnisse, die Sie erhalten, möglicherweise schwierig zu lesen oder zu interpretieren.

Mit optimalem Binning können Sie jedoch Trennwerte bestimmen, mit denen Sie das bestmögliche Ergebnis für Algorithmen erzielen können, die für nominale Attribute konzipiert sind.

Mit dieser Prozedur können Sie vor der Modellerstellung aus drei Binning-Typen für die Vorverarbeitung von Daten eine Auswahl treffen:

- Unüberwacht – Es werden Klassen mit der gleichen Anzahl erstellt.
- Überwacht – Bei der Bestimmung der Trennwerte wird die Zielvariable berücksichtigt. Dieses Verfahren ist genauer als das unüberwachte Verfahren. Allerdings ist es auch rechenintensiver.
- Hybrides Verfahren – Hierbei werden das überwachte und das unüberwachte Verfahren kombiniert. Dieses Verfahren ist besonders nützlich, wenn Sie über viele unterschiedliche Werte verfügen.

Damit Sie Assets effizient gemeinsam nutzen und wiederverwenden können, müssen diese so geschützt werden, dass interne und externe Konformitätsanforderungen erfüllt werden. Außerdem müssen die Ergebnisse so veröffentlicht werden, dass mehr Geschäftsbenutzer die Ergebnisse anzeigen und nutzen können. Zu diesem Zweck können Sie IBM SPSS Statistics durch IBM SPSS Collaboration and Deployment Services ergänzen. Weitere Informationen zu den enthaltenen wertvollen Funktionen erhalten Sie unter: **ibm.com**/spss/cds

Unsere Statistik-Software-Suite ist nun in drei Editionen erhältlich: IBM SPSS Statistics Standard, IBM SPSS Statistics Professional und IBM SPSS Statistics Premium. Diese Editionen fassen wichtige Funktionalität zusammen. So können Sie effizient sicherstellen, dass Ihrem gesamten Team oder Ihrer Abteilung alle Funktionen zu Verfügung stehen, die sie benötigen, um die Analysen zu erstellen, die den Erfolg Ihres Unternehmens garantieren.

Funktionen

Automatisierte Datenvorbereitung

Empfohlene Schritte zur Beschleunigung der Modellerstellung und zur Verbesserung der Vorhersagekraft:

- Festlegen des Ziels: Geschwindigkeit und Genauigkeit gegeneinander abwägen, Geschwindigkeit und Genauigkeit optimieren oder Analyse anpassen
- Vorbereiten der Termine und Uhrzeiten für die Modellierung:
  - Abgelaufene Zeit bis zu einem Referenzdatum berechnen
  - Abgelaufene Zeit bis zu einer Referenzuhrzeit berechnen
  - Zyklische Zeitelemente extrahieren
- Ausschließen von Eingabefeldern geringer Qualität:
  - Felder mit zu vielen fehlenden Werten ausschließen
  - Nominale Felder mit zu vielen eindeutigen Kategorien ausschließen
  - Kategoriale Felder mit zu vielen Werten in einer Einzelkategorie ausschließen
- Festlegen von Messniveaus:
  - Messniveaus numerischer Felder festlegen
- Vorbereiten von Feldern für die Verbesserung der Datenqualität:
  - Ausreißer behandeln
  - Fehlende Werte ersetzen
  - Nominale Felder umordnen
- Neuskalieren von Feldern:
  - Analyse gewichten
- Kontinuierliche Eingabefelder
- Kontinuierliche Zielfelder
- Transformieren von Feldern:
  - Kategoriale und/der kontinuierliche Eingabefelder verwenden
- Ausführen der Funktionsauswahl und der Erstellung
- Namensfelder:
  - Transformierte und erstellte Felder
  - Berechnete Zeiträume
  - Extrahierte zyklische Zeitelemente
- Anwenden von Transformationen auf Daten

Datenvalidierung

Sie können die Datenvalidierungsprozedur (VALIDATEDATA) verwenden, um Daten in der Arbeitsdatendatei zu prüfen. Basisprüfungen: Sie können Basisprüfungen angeben, die auf Variablen und auf Fälle in der Datei angewendet werden sollen.

- Sie können zum Beispiel Berichte erhalten, in denen Variablen mit einem hohen Prozentsatz fehlender Werte oder leerer Fälle angegeben sind:
  - Maximaler Prozentsatz für fehlende Werte
  - Maximaler Prozentsatz für Fälle in einer Einzelkategorie
  - Maximaler Prozentsatz für Fälle mit der Anzahl 1
  - Minimaler Variationskoeffizient
  - Minimale Standardabweichung
  - Unvollständige Kennungen markieren
  - Duplizierte Kennungen markieren
  - Leere Fälle markieren
- Standardregeln: Beschreiben der Daten, Anzeigen von Einzelvariablenregeln und deren Anwendung auf Analysevariablen:
  - Datenbeschreibung:
    - Verteilung: Zeigt für kategoriale Variablen ein Balkendiagramm oder für Skalenvariablen ein Histogramm in der Größe eines Piktogramms an.
    - Es werden minimale und maximale Datenwerte angezeigt.
  - Einzelvariablenregeln:
    - Regeln werden auf einzelne Variablen angewendet, um fehlende oder ungültige Werte zu ermitteln, zum Beispiel Werte außerhalb eines gültigen Bereichs.
    - Außerdem sind benutzerdefinierte Einzelvariablenregeln möglich.
- Angepasste Regeln: Es können variablenübergreifende Regelausdrücke definiert werden, in denen die Antworten der Befragten gegen die Logik verstoßen (zum Beispiel „schwanger und männlich“).
- Ausgabe: Berichte mit Beschreibungen ungültiger Daten:
  - Fallweiser Bericht, in dem die Verstöße gegen die Validierungsregeln nach Fällen aufgelistet sind:
    - Geben Sie die minimale Anzahl der Verstöße an, die erforderlich ist, damit ein Fall in den Bericht aufgenommen wird.
    - Geben Sie die maximale Anzahl der Fälle im Bericht an.
  - Standardberichte für Validierungsregeln:
    - Verstöße nach Analysevariablen zusammenfassen
    - Verstöße nach Regel zusammenfassen
    - Deskriptive Statistik anzeigen
- Speichern: Sie können Variablen speichern, die Regelverstöße aufzeichnen und diese verwenden, um Daten zu bereinigen und Fehlerfälle herauszufiltern:
  - Zusammenfassungsvariablen:
    - Indikator für leeren Fall
    - Indikator für duplizierte Kennung
    - Indikator für unvollständige Kennung
    - Verstoß gegen Validierungsregel (Gesamtzahl)
  - Indikatorvariable, in denen alle Verstöße gegen Validierungsregeln aufzeichnet werden

Ermitteln ungewöhnlicher Fälle

Mit der Prozedur zur Erkennung von Anomalien (DETECTANOMALY) wird nach ungewöhnlichen Fällen gesucht. Die Suche basiert auf Abweichungen von der zugehörigen Referenzgruppe und liefert Ursachen für derartige Abweichungen:

- Sie können mit dem Unterbefehl VARIABLES Variablen angeben, die von der Prozedur verwendet werden sollen. Sie können kategoriale und kontinuierliche Variablen sowie Kennungsvariablen (zur Ermittlung von Fällen) angeben und Variablen auflisten, die von der Analyse ausgeschlossen werden.
- Mit dem Unterbefehl HANDLEMISSING werden die Verfahren angegeben, mit denen in dieser Prozedur fehlende Werte verarbeitet werden:
  - Option für das Anwenden der Verarbeitung fehlender Werte (APPLY). Wenn diese Option ausgewählt ist, werden fehlende Werte von kontinuierlichen Variablen durch Gesamtmittel ersetzt und fehlende Kategorien kategorialer Variablen werden kombiniert und als eine gültige Kategorie behandelt. Die verarbeiteten Variablen werden anschließend in der Analyse verwendet. Wenn diese Option nicht ausgewählt ist, werden Fälle mit fehlenden Werten aus der Analyse ausgeschlossen.
  - Option zum Erstellen einer zusätzlichen Variablen für den fehlenden Anteil und zum Verwenden in der Analyse (CREATEMISPROPVAR). Falls dies ausgewählt ist, wird eine zusätzliche Variable erstellt, die als Variable für den fehlenden Anteil (Missing Proportion Variable) bezeichnet wird und den Anteil fehlender Variablen in den einzelnen Datensätzen darstellt, und in der Analyse verwendet. Falls dies nicht ausgewählt ist, wird die Variable für den fehlenden Anteil nicht erstellt.
- Mit dem Unterbefehl CRITERIA werden die folgenden Einstellungen angegeben:
  - Minimale und maximale Anzahl von Referenzgruppen
  - Korrekturgewichtung für das Messniveau
  - Anzahl der Ursachen in der Anomalieliste
  - Prozentsatz der Fälle, die als Anomalien betrachtet werden und in die Anomalieliste aufgenommen werden
  - Anzahl der Fälle, die als Anomalien betrachtet werden und in die Anomalieliste aufgenommen werden
  - Trennwert des Anomalie-Index zum Bestimmen, ob ein Fall als Anomalie betrachtet wird
- Zusätzliche Variablen können mit dem Unterbefehl SAVE in der Arbeitsdatendatei gespeichert werden:
  - Anomalie-Index
  - Referenzgruppenkennung
  - Referenzgruppengröße
  - Referenzgruppengröße als Prozentsatz
  - Variable, die einer Ursache zugeordnet ist
  - Variableneinflussmaß, das einer Ursache zugeordnet ist
  - Variablenwert, der einer Ursache zugeordnet ist
  - Normwert, der einer Ursache zugeordnet ist
- Mit dem Unterbefehl OUTFILE kann das Modell unter einem angegebenen Namen als XML-Datei gespeichert werden
- Mit dem Unterbefehl PRINT kann die Anzeige der Ausgabeergebnisse gesteuert werden.
- Sie können Folgendes ausgeben:
  - Zusammenfassung der Fallverarbeitung
  - Anomalie-Indexliste, Referenzgruppenkennung für Anomalien und Anomalie-Ursachenliste
  - Normentabelle für kontinuierliche Variable, falls bei der Analyse eine kontinuierliche Variable verwendet wird
  - Normentabelle für kategoriale Variable, falls bei der Analyse eine kategoriale Variable verwendet wird
  - Anomalie-Indexzusammenfassung
  - Zusammenfassungstabelle für die einzelnen Ursachen:
    - Die gesamte angezeigte Ausgabe mit Ausnahme der Notiztabelle und mit Ausnahme aller Warnungen unterdrücken

Optimales Binning

Daten können mithilfe des optimalen Binnings vorverarbeitet werden. Dabei werden kontinuierliche Variablen kategorisiert, indem die Werte der einzelnen Variablen auf Klassen verteilt werden. Dieses Verfahren ist nützlich, um die Anzahl der Werte in den angegebenen Binning-Eingabevariablen zu reduzieren. Dadurch kann sich die Leistung von Algorithmen beträchtlich erhöhen. Wenn Sie bestimmte Verfahren für optimales Binning anwenden, erleichtert eine Leitvariable das Ermitteln der Trennwerte. Dadurch wird die Beziehung zwischen der Leitvariablen und der in Klassen eingeteilten Variablen maximiert.

- Auswahl zwischen folgenden Verfahren:
  - Unüberwachtes Binning mithilfe des Algorithmus der gleichen Häufigkeiten. Bei diesem Verfahren wird der Algorithmus der gleichen Häufigkeiten verwendet, um die Binning-Eingabevariablen zu diskretisieren. Es ist keine Leitvariable erforderlich.
  - Überwachtes Binning mithilfe des MDLP-Algorithmus (MDLP – Minimal Description Length Principle). Bei diesem Verfahren werden die Binning-Eingabevariablen mithilfe des MDLP-Algorithmus und ohne Vorverarbeitung diskretisiert. Dies eignet sich für Datasets mit einer niedrigeren Fallanzahl. Es ist eine Leitvariable erforderlich.
  - Hybrides MDLP-Binning. Dieses umfasst die Vorverarbeitung über den Algorithmus der gleichen Häufigkeiten, gefolgt vom MDLP-Algorithmus. Dieses Verfahren eignet sich für Datasets mit einer größeren Anzahl von Fällen. Es ist eine Leitvariable erforderlich.
- Geben Sie die folgenden Kriterien an:
  - Wie der minimale Trennwert für die einzelnen Binning-Eingabevariablen definiert werden soll
  - Wie der maximale Trennwert für die einzelnen Binning-Eingabevariablen definiert werden soll
  - Wie die Untergrenze eines Intervalls definiert werden soll
  - Ob das Zusammenführen dünn besetzter Klassen erzwungen werden soll
  - Ob fehlende Werte mit der listenweisen oder mit der paarweisen Löschung verarbeitet werden sollen
- Speichern Sie Folgendes:
  - Neue Variable, die in Klassen eingeteilte Werte enthalten
  - Syntax für eine SPSS Statistics Base-Syntaxdatei
- Steuern Sie die Ergebnisanzeige mithilfe des Unterbefehls PRINT. Sie können Folgendes ausgeben:
  - Die Trennwertsätze der Binning-Eingabevariablen
  - Beschreibende Informationen für alle Binning-Eingabevariablen
  - Modellentropie für in Klassen eingeteilte Variablen

Systemvoraussetzungen

Anforderungen variieren je nach Plattform. Einzelheiten finden Sie unter: **ibm.com**/spss/requirements

Informationen zu IBM Business Analytics

IBM Business Analytics-Software stellt Entscheidern verlässliche Informationen zur Verfügung, die für fundierte Entscheidungen nötig sind. IBM bietet ein umfassendes, einheitliches Portfolio für Business Intelligence, vorausschauende und erweiterte Analyse, Financial Performance- und Strategiemanagement, Governance, Risikomanagement und Compliance sowie Analyseanwendungen.

Mit IBM Software können Unternehmen Trends, Muster und Unregelmäßigkeiten erkennen, „Was wäre, wenn"-Szenarien vergleichen, mögliche Bedrohungen und Chancen vorhersagen, kritische Geschäftsrisiken erkennen und minimieren sowie Ressourcen planen, budgetieren und prognostizieren. Durch diese umfassenden Analysefunktionen sind unsere Kunden rund um den Globus in der Lage, ihre Geschäftsergebnisse besser zu verstehen, vorauszusehen und zu beeinflussen.

Weitere Informationen

Weitere Informationen finden Sie unter:
ibm.com/de/spss

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
ibm.com/de

IBM Österreich
Obere Donaustrasse 95
1020 Wien
ibm.com/at

IBM Schweiz
Vulkanstrasse 106
8010 Zürich
ibm.com/ch

Die IBM Homepage finden Sie unter:
ibm.com

Bitte der Wiederverwertung zuführen

Business Analytics software

YTD03016-DEDE-04