IBM SPSS Data Preparation

Nauwkeuriger resultaten dankzij betere datapreparatie

In het kort

- Verdachte of ongeldige gevallen, variabelen en gegevenswaarden opsporen.
- Patronen ontdekken in ontbrekende gegevens.
- Een overzicht geven van de verdeling van variabelen.
- Gegevens sneller en nauwkeuriger gereedmaken voor analyse.

Voordat onderzoekers een analyse kunnen starten, moeten ze de beschikbare gegevens voorbewerken. IBM SPSS Statistics bevat weliswaar tools voor het voorbewerken van gegevens, maar soms zijn meer specialistische technieken nodig om uw data gereed te maken. IBM SPSS Data Preparation maakt het mogelijk om verdachte of ongeldige gevallen, variabelen en gegevenswaarden op te sporen en om patronen van ontbrekende gegevens te bekijken. Ook kunt u de verdeling van variabelen vaststellen en nauwkeuriger werken met algoritmen die bedoeld zijn voor nominale attributen. Hiermee stroomlijnt u de voorbewerking van uw gegevens – zodat u sneller aan de slag kunt met uw analyse en hieruit meer nauwkeurige conclusies kunt trekken. Voor de snelste resultaten kiest u ervoor om uw gegevens geheel automatisch te laten voorbewerken, maar u kunt ook een keuze maken uit verschillende andere methoden voor het aanpakken van meer uitdagende datasets.

SPSS Data Preparation is verkrijgbaar in de vorm van clientsoftware, maar ten behoeve van een betere performance en een grotere schaalbaarheid is er ook een serverversie beschikbaar.

Verschillende opties voor datapreparatie

De datavalidatie-procedure

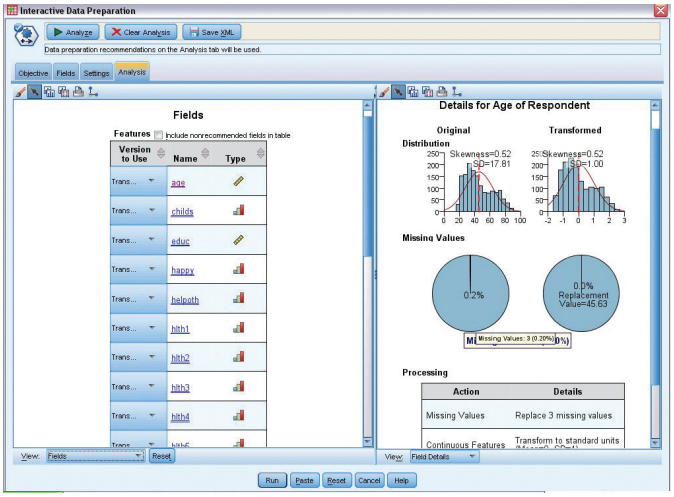
Validatie van gegevens is altijd een typisch handmatig proces geweest. U creëert een frequentietabel op basis van uw data en drukt deze gegevens af. Vervolgens omcirkelt u de gegevens die gecorrigeerd moeten worden en zoekt u case-ID's op. Dit is niet alleen zeer tijdrovend, maar omdat elke analist in uw organisatie waarschijnlijk een iets andere methode gebruikt, is het moeilijk om van project tot project de consistentie te bewaren.

Om handmatige controles te vermijden, kunt u gebruikmaken van de datavalidatie-procedure. Met deze procedure kunt u gegevenscontroles uitvoeren op basis van regels die betrekking hebben op het meetniveau van elke variabele (hetzij categorisch, hetzij doorlopend). Als u bijvoorbeeld een analyse uitvoert op onderzoeksgegevens met variabelen op een vijfpunts Likert-schaal, kunt u met de datavalidatie-procedure een regel voor dergelijke vijfpuntsschalen opstellen en alle gevallen laten markeren die een waarde hebben die buiten de schaal van 1-5 valt. U kunt rapporten van ongeldige gevallen laten genereren en overzichten van overtredingen van de regels en het aantal betrokken gevallen laten opstellen. En u kunt validatieregels definiëren voor afzonderlijke variabelen (bijvoorbeeld of alle waarden binnen het bereik vallen) en voor kruiscontroles tussen variabelen (bijvoorbeeld "mannen in verwachting").

Aan de hand van deze kennis kunt u bepalen in hoeverre de gegevens betrouwbaar zijn en kunt u, voordat de analyse begint, naar eigen inzicht verdachte gevallen verwijderen of corrigeren.

Gegevens in één stap automatisch bewerken

Het handmatig voorbewerken van gegevens is een complex proces waar gewoonlijk zo'n 40 tot 90 procent van de tijd in gaat zitten die een analist aan een bepaald project kan besteden. Als u de resultaten snel nodig hebt, is de procedure Automated Data Preparation (ADP) nuttig voor het detecteren en corrigeren van kwaliteitsfouten en het aanvullen van ontbrekende waarden, dit alles in één efficiënte stap. De ADP-functie genereert overzichtelijke rapporten met volledige aanbevelingen en visualisaties, aan de hand waarvan u kunt bepalen welke gegevens u in uw analyse wilt gebruiken.



Figuur 1: De functie Automated Data Preparation levert aanbevelingen en biedt gebruikers de mogelijkheid om in te zoomen op specifieke aanbevelingen.

De procedure Anomaly Detection

Met de procedure Anomaly Detection kunt u voorkomen dat de analyse wordt verdraaid door uitschieters. Deze procedure gaat, op basis van afwijkingen van vergelijkbare gevallen, op zoek naar ongebruikelijke gevallen en geeft verklaringen voor dergelijke afwijkingen. Uitschieters kunt u markeren door een nieuwe variabele te creëren. En als u de ongebruikelijke gevallen eenmaal hebt opgespoord, kunt u ze verder onderzoeken en bepalen of ze moeten worden meegenomen in uw analyse.

Optimal Binning

Om gebruik te kunnen maken van algoritmen die bedoeld zijn voor nominale attributen (zoals Naïve Bayes- en logitmodellen), moet u uw schaalvariabelen "binnen" voordat u een model opstelt. Worden de schaalvariabelen niet gebind, dan neemt de verwerking van bepaalde algoritmen, zoals multinomiale logistieke regressie, zeer veel tijd in beslag of worden deze algoritmen in het geheel niet geconvergeerd, vooral als er sprake is van grote datasets. Bovendien zijn de resultaten die u krijgt vaak moeilijk te lezen of te interpreteren.

Optimal Binning maakt het echter mogelijk om de grenswaarden te bepalen, zodat algoritmen die bedoeld zijn voor nominale attributen, de best mogelijke resultaten opleveren.

Met deze procedure kunt u, alvorens een model op te stellen, kiezen uit drie soorten binning voor de voorbewerking van gegevens:

- Zonder toezicht – Bins met gelijke aantallen maken
- Met toezicht – Bij het vaststellen van grenswaarden rekening houden met de doelvariabele. Deze methode is nauwkeuriger dan de methode zonder toezicht, maar er is wel meer rekenkracht voor nodig.
- Hybride aanpak – Een combinatie van de methoden zonder en met toezicht. Deze methode komt vooral goed van pas als u grote aantallen verschillende waarden hebt.

Om assets efficiënt te kunnen uitwisselen en hergebruiken, om ze te beschermen volgens interne en externe compliance-eisen en om resultaten zo te publiceren dat grotere aantallen zakelijke gebruikers ze kunnen bekijken en gebruiken, is het een overweging om uw SPSS Statistics-software uit te breiden met IBM SPSS Collaboration and Deployment Services. Meer informatie over de interessante mogelijkheden hiervan vindt u op: **ibm.com**/spss/cds

Ons pakket van statistische software is nu beschikbaar in drie uitgaven: IBM SPSS Statistics Standard, IBM SPSS Statistics Professional en IBM SPSS Statistics Premium. Doordat er in deze pakketten essentiële functies bij elkaar zijn gebracht, beschikt uw gehele team of afdeling op een efficiënte manier over de voorzieningen en functies die nodig zijn om de analyses uit te voeren die bijdragen aan het succes van uw organisatie.

Kenmerken

Automatische datapreparatie

Procedures aanbevelen om het bouwen van modellen te versnellen en de voorspellingskracht te vergroten:

- Doel vaststellen: Snelheid en nauwkeurigheid in evenwicht, afgestemd op snelheid, afgestemd op nauwkeurigheid, of aangepaste analyse
- Datums en tijden gereedmaken voor modellering:
  - Verstreken tijd tot referentiedatum berekenen
  - Verstreken tijd tot referentietijd berekenen
  - Cyclische tijdselementen extraheren
- Invoervelden van slechte kwaliteit uitsluiten:
  - Velden met te veel ontbrekende waarden uitsluiten
  - Nominale velden met te veel unieke categorieën uitsluiten
  - Categorische velden met te veel waarden in een enkele categorie uitsluiten
- Meetniveaus aanpassen:
  - Meetniveaus van numerieke velden aanpassen
- Velden gereedmaken voor een betere datakwaliteit:
  - Uitschieters afhandelen
  - Ontbrekende waarden invullen
  - Nominale velden opnieuw rangschikken
- Velden opnieuw inschalen:
  - Gewicht van analyse
- Doorlopende invoervelden
- Doorlopende doelvelden
- Velden transformeren:
  - Categorische en/of doorlopende invoervelden gebruiken
- Kenmerken selecteren en construeren
- Naamvelden:
  - Getransformeerde en geconstrueerde velden
  - Berekende waarden voor de duur
  - Geëxtraheerde cyclische tijdselementen
- Transformaties uitvoeren op gegevens

Gegevens valideren

De datavalidatie-procedure wordt gebruikt voor het valideren van gegevens in het werkdatabestand. Elementaire controles: specificeer welke elementaire controles er moeten worden uitgevoerd op de variabelen en gevallen in uw bestand.

- Bijvoorbeeld: Rapporten laten genereren waarin wordt aangegeven welke variabelen een hoog percentage ontbrekende gegevens of lege gevallen hebben.
  - Maximumpercentage ontbrekende waarden
  - Maximumpercentage van gevallen in een enkele categorie
  - Maximumpercentage van gevallen met 1 als aantal
  - Minimum variatiecoëfficiënt
  - Minimum standaarddeviatie
  - Onvolledige ID's markeren
  - Duplicaat-ID's markeren
  - Lege gevallen markeren
- Standaardregels: De gegevens beschrijven, afzonderlijke variabele regels bekijken en deze regels toepassen op analysevariabelen:
  - Beschrijving van de gegevens:
    - Distributie: Een miniatuur staafdiagram voor categorische variabelen of een histogram voor schaalvariabelen
    - Minimale en maximale gegevenswaarden worden weergegeven
  - Regels voor afzonderlijke variabelen:
    - Regels toepassen op afzonderlijke variabelen die ontbrekende of ongeldige waarden aangeven, zoals waarden die buiten het geldige bereik vallen
    - De gebruiker kan ook zelf regels voor afzonderlijke variabelen definiëren
- Regels op maat: Regelexpressies voor meerdere variabelen waarin de antwoorden van respondenten niet voldoen aan de regels van de logica (bijvoorbeeld: "mannen die in verwachting zijn")
- Uitvoer: Rapporten waarin ongeldige gegevens worden beschreven:
  - Rapport per geval met lijsten van overtredingen van de validatieregels per geval:
    - Opgeven hoeveel overtredingen er minimaal in een geval moeten optreden om te worden opgenomen in het rapport
    - Het maximum aantal gevallen in het rapport opgeven
  - Rapporten met standaard validatieregels:
    - Overzicht van overtredingen per analysevariabele
    - Overzicht van overtredingen per regel
    - Beschrijvende cijfers
- Opslaan: Maakt het mogelijk om variabelen waarmee overtredingen van regels worden vastgelegd, op te slaan en te gebruiken om de gegevens op te schonen en slechte gevallen uit te filteren:
  - Overzichtsvariabelen:
    - Indicator lege gevallen
    - Indicator duplicaat-ID's
    - Indicator onvolledige ID's
    - Overtredingen van validatieregels (totaal aantal)
  - Indicatorvariabelen waarmee alle overtredingen van regels worden vastgelegd

Ongebruikelijke gevallen opsporen

De procedure Anomaly Detection gaat, op basis van afwijkingen van vergelijkbare gevallen, op zoek naar ongebruikelijke gevallen en geeft verklaringen voor dergelijke afwijkingen:

- Variabelen die door de procedure moeten worden gebruikt, opgeven met de subopdracht VARIABLES. Categorische, doorlopende en ID-variabelen opgeven (om gevallen aan te geven) en lijsten oproepen van variabelen die zijn uitgesloten van de analyse.
- Met de subopdracht HANDLEMISSING wordt aangegeven op welke manieren ontbrekende waarden in deze procedure worden afgehandeld:
  - Afhandeling van ontbrekende waarden toepassen. Met deze optie worden ontbrekende waarden van doorlopende variabelen vervangen door het groot gemiddelde (grand mean) en worden ontbrekende categorieën van categorische variabelen gecombineerd en behandeld als een geldige categorie. De verwerkte variabelen worden vervolgens gebruikt in de analyse. Is deze optie niet geselecteerd, dan worden gevallen met ontbrekende waarden uitgesloten van de analyse.
  - Een extra Missing Proportion Variable maken en deze gebruiken in de analyse. Met deze optie wordt er een ontbrekende variabele Missing Proportion Variable gemaakt. Deze variabele, die aangeeft hoe groot het aandeel ontbrekende variabelen in elk record is, wordt gebruikt in de analyse. Is deze optie niet geselecteerd, dan wordt de Missing Proportion Variable niet gecreëerd.
- Met de subopdracht CRITERIA worden de volgende instellingen opgegeven:
  - Minimum en maximum aantal peer groups
  - Aanpassingsgewicht op meetniveau
  - Aantal redenen in de anomalielijst
  - Percentage van gevallen dat als anomalie wordt beschouwd en dat is opgenomen in de anomalielijst
  - Aantal gevallen dat als anomalie wordt beschouwd en dat is opgenomen in de anomalielijst
  - Grenswaarde van de anomalie-index om te bepalen of een geval wordt beschouwd als een anomalie
- Extra variabelen opslaan in het werkdatabestand, met de subopdracht SAVE:
  - Anomalie-index
  - Peer group-ID
  - Grootte van de peer group
  - Grootte van de peer group als percentage
  - De variabele, gekoppeld aan een reden
  - De mate van impact van de variabele, gekoppeld aan een reden
  - De waarde van de variabele, gekoppeld aan een reden
  - De normwaarde, gekoppeld aan een reden
- Het model wegschrijven naar een XML-bestand met een opgegeven naam, met de subopdracht OUTFILE
- De weergave van de resultaten besturen, met de subopdracht PRINT
- U kunt het volgende afdrukken:
  - Samenvatting van case-verwerking
  - Voor anomalieën: de indexlijst, de lijst van peer-ID's en de lijst van redenen
  - De tabel Continuous Variable Norms (als er in de analyse gebruik wordt gemaakt van doorlopende variabelen)
  - De tabel Categorical Variable Norms (als er in de analyse gebruik wordt gemaakt van categorische variabelen)
  - Overzicht van de anomalie-index
  - Overzichtstabel van redenen, en voor elke reden:
    - Alle afgebeelde uitvoer onderdrukken, behalve de notitietabel en alle waarschuwingen

Optimal Binning

Met het voorbewerken van gegevens met behulp van Optimal Binning, worden een of meer doorlopende variabelen gecategoriseerd door de waarden van elke variabele te verdelen over vergaarbakken ("bins"). Met deze procedure kan het aantal waarden van de invoervariabelen worden verkleind, waardoor de performance van algoritmen sterk toeneemt. Bij gebruik van bepaalde methoden voor Optimal Binning wordt er een gidsvariabele gebruikt om de grenswaarden te bepalen. Daardoor wordt de relatie tussen de gidsvariabele en de variabele in de "bin" maximaal versterkt.

- U kunt kiezen uit de volgende methoden:
  - Binning zonder toezicht, via het algoritme voor gelijke frequenties. Bij deze methoden wordt er voor binning van de invoervariabelen gebruikgemaakt van een algoritme voor gelijke frequenties. Hiervoor is er geen gidsvariabele vereist.
  - Binning onder toezicht, via het MDLP-algoritme (Minimal Description Length Principle). Deze methode maakt de invoervariabelen voor binning concreet, zonder enige voorbewerking, met behulp van het MDLP-algoritme. Deze methode is geschikt voor datasets met een klein aantal gevallen. Er is een gidsvariabele vereist.
  - Hybride MDLP-binning. Hierbij vindt er voorbewerking plaats met het algoritme voor gelijke frequenties, gevolgd door het MDLP-algoritme. Deze methode is geschikt voor datasets met grote aantallen gevallen. Er is een gidsvariabele vereist.
- U kunt de volgende criteria opgeven:
  - Hoe de minimum grenswaarde voor elke invoervariabele voor binning wordt gedefinieerd
  - Hoe de maximum grenswaarde voor elke invoervariabele voor binning wordt gedefinieerd
  - Hoe de ondergrens van een interval wordt gedefinieerd
  - Of matig gevulde bins gedwongen moeten worden samengevoegd
  - Of ontbrekende waarden worden afgehandeld door ze lijstgewijs resp. paarsgewijs te wissen
- U kunt het volgende opslaan:
  - Nieuwe variabelen met waarden uit de bins
  - Syntaxis, in een SPSS Statistics Base-syntaxisbestand
- De weergave van de uitvoer besturen met de subopdracht PRINT. U kunt het volgende afdrukken:
  - De grenswaarden van de invoervariabelen voor binning
  - Beschrijvingen van alle invoervariabelen voor binning
  - Entropie van model voor variabelen in bins

Systeemvereisten

Vereisten verschillen per platform. Zie voor nadere informatie **ibm.com**/spss/requirements

Informatie over IBM Business Analytics

IBM Business Analytics-software biedt de inzichten die beslissers nodig hebben om actie te ondernemen en de bedrijfsprestaties te verbeteren. IBM levert een uitgebreide portfolio aan applicaties voor business intelligence, predictive en advanced analytics, financieel en strategisch prestatiemanagement, governance, risico en compliance en analyse.

Met IBM-software kunnen bedrijven trends spotten, patronen en afwijkingen ontdekken, "what if"-scenario's vergelijken, kansen en bedreigingen opsporen, bedrijfsrisico's aanpakken en de inzet van resources voorspellen en begroten. Dankzij deze diepgaande analysemogelijkheden kunnen organisaties overal ter wereld hun resultaten doorgronden, voorspellen en daarop anticiperen.

Meer informatie

Voor meer informatie gaat u naar
ibm.com/business-analytics

Gesprek aanvragen

Om een gesprek aan te vragen of een vraag te stellen, gaat u naar **ibm.com**/business-analytics/contactus
Een IBM-vertegenwoordiger zal binnen twee werkdagen reageren op uw informatieverzoek.

IBM Software Group
Johan Huizingalaan 765
1066 VH Amsterdam
Netherlands

De homepage van IBM is te vinden op:
ibm.com

Please Recycle

Business Analytics software

YTD03016-NLNL-04