IBM SPSS DATA PREPARATION

Zlepšete přípravu svých dat a získejte přesnější výsledky

Každý analytik by si měl před analýzou svá data řádně připravit. IBM SPSS Statistics Base obsahuje mnoho nástrojů přípravy dat, avšak pro úplnou a rychlou kontrolu jsou potřeba speciální techniky. Jedna z takových technik, která má zcela samostatnou pozici v procesu přípravy dat je obsažena v IBM® SPSS® Missing Values. Umožňuje porozumět příčinám vzniku chybějících dat a nabízí imputační postupy. Zásadní procedury pro přípravu dat však najdete v modulu IBM® SPSS® Data Preparation – v něm snadno identifikujete podezřelé či chybné případy, proměnné a hodnoty, zobrazíte strukturu vynechaných hodnot, prověříte rozložení proměnných a budete daleko přesněji pracovat s algoritmy navrženými pro nominální znaky. Urychlíte proces přípravy dat a bezpečně přejdete k jejich analýze. Zvolíte-li si zcela automatickou přípravu dat, dosáhnete výsledku ještě rychleji. V případě složitější situace, vyberte ručně z několika nabízených metod.
IBM SPSS Data Preparation lze provozovat v režimu lokální desktopové aplikace nebo v režimu klient – server, což přináší zrychlení výpočtů a větší škálovatelnost.

Nejdůležitější:

IBM SPSS Data Preparation umožňuje:

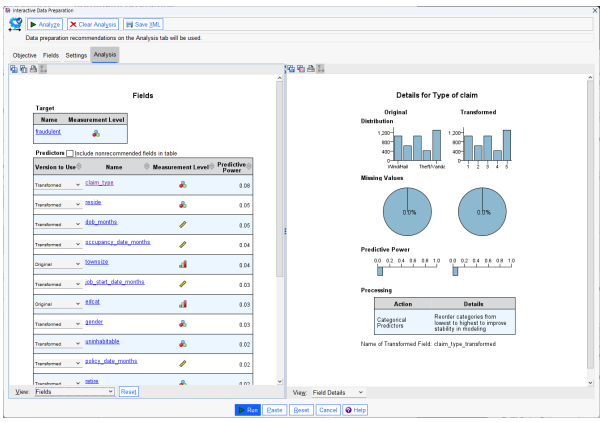
- **identifikovat podezřelé a neplatné případy, proměnné a hodnoty**
- **zobrazit struktury vynechaných hodnot**
- **sumarizovat rozložení proměnných**
- **přesněji a rychleji připravit data pro analýzu**

Procedura Validate Data

Kontrola dat byla dlouho typicky manuálním procesem. Běžně se zkonstruovaly a vytiskly tabulky četností, označily chyby a zjistila se identifikační čísla podezřelých případů. Takový postup je samozřejmě časově náročný. Navíc každý analytik mívá, i v rámci jedné organizace, svůj individuální postup kontroly a opravy chyb. Proto je vhodné tyto postupy sjednotit a zajistit standardizovaný proces přípravy dat ve všech projektech jako normu kontroly kvality.
Pomocí procedury Validate Data eliminujete manuální kontroly. Procedura vám nabízí kontrolní mechanismy zavedením pravidel pro každou proměnnou s ohledem na její typ, ať už se jedná o kategorizovanou či spojitou proměnnou. Například pro proměnnou měřenou na pětibodové škále procedura Validate Data zavede a aplikuje pravidlo, jež označí všechny případy, u kterých je hodnota mimo tuto stupnici.
Obdržíte zprávu o neplatných případech, přehledy o porušení pravidel a informaci o četnosti případů, kdy bylo pravidlo porušeno. Pravidla lze určit pro každou proměnnou zvlášť (jako v případě stupnice), ale i pro vztahy mezi více proměnnými (např. „gravidní muži“). Na základě této informace můžete před začátkem samotné analýzy určit celkovou validitu svých dat a podle svého uvážení odstranit či opravit chybné, resp. podezřelé případy.

Automatická příprava dat

Připravovat data ručně je náročný proces, který zabírá 40 až 90 procent času, který má analytik k dispozici na daný projekt. Potřebujete-li výsledek rychle, použijte proceduru Automated Data Preparation (ADP), která odhaluje a opravuje kvalitativní chyby dat a imputuje chybějící hodnoty v jediném kroku. Procedura poskytuje přehledné a srozumitelné reporty obsahující doporučení a grafy, pomocí kterých rozhodnete, zda je vhodné data zařadit do analýzy.



Procedura automatické přípravy dat poskytuje reporty obsahující doporučení a informaci pro jejich prozkoumání

Procedura Anomaly Detection

Použijte proceduru Anomaly Detection pro identifikaci odlehlých hodnot ve vícerozměrných datech a zabránili tak zkresleným závěrům. Procedura hledá neobvyklé případy na základě odchylek od skupin obdobných případů a informuje o tom, proč je případ považován za anomální. Neobvyklé případy můžete označit novou indikační proměnnou.

Poté, co jsou neobvyklé případy identifikovány, je později můžete přezkoumat a rozhodnout se, zda je do další analýzy zahrnete, či nikoli. Tato procedura se používá nejen pro kontrolu dat, ale také v analýze dat, např. pro identifikaci podvodů, pro hledání nových mutací apod.

Procedura Optimal Binning

Aby mohly být použity algoritmy navržené pro nominální proměnné (jako Naive Bayes a logitové modely), musí být spojité proměnné nejprve kategorizovány, než budou v modelech použity. V případě, že nejsou tyto úpravy provedeny, pak algoritmy jako multinomická logistická regrese trvají extrémně dlouho, nebo dokonce vůbec nekonvergují, a to zvláště v případě velkých datových souborů. Dosažené výsledky mohou být navíc špatně čitelné a nejasně interpretovatelné.

Optimální rozdělení stupnice spojité proměnné do intervalů určí hranice intervalů, které vám pomohou dosáhnout nejlepších možných výsledků s algoritmy navrženými pro nominální proměnné.

Při přípravě dat pro modelování si v této proceduře můžete vybrat ze tří typů dělení:

- ***nesupervizované dělení*** - vytváří kategorie se stejným počtem případů
- ***supervizované dělení*** - určuje intervaly maximalizací predikční síly pro zvolenou cílovou proměnnou. Metoda je přesnější než nekontrolované dělení v predikčních modelech, ale je početně náročnější a její výsledky jsou vztaženy k jedné konkrétní cílové proměnné
- ***smíšený přístup*** - kombinuje nesupervizované a supervizované dělení. Smíšená metoda je užitečná, zejména při velkém počtu odlišných hodnot

SPECIFIKACE

Procedura Automated Data Preparation

Doporučí potřebné kroky pro dosažení rychlejší tvorby modelu a zvýšení jeho predikční síly.

- *cíl přípravy dat: optimalizovat pro rychlost,optimalizovat pro přesnost, vyvážit rychlosta přesnost, vlastní nastavení*
- *příprava proměnných s informací o datu a čase*
  - *výpočet uplynulé doby od referenčního data*
  - *výpočet uplynulé doby od referenčního času*
  - *extrakce částí z cyklických údajů*
- *vyloučení vstupních proměnných s nízkou kvalitou*
  - *vyřazení proměnných s příliš velkým podílemvynechaných hodnot*
  - *vyloučení nominálních proměnných s přílišvelkým počtem odlišných hodnot*
  - *vyřazení kategorizovaných proměnných s přílišvelkým počtem případů v jediné kategorii*
- *úprava typu proměnných*
  - *změna typu ordinálních a číselných proměnnýchdle počtu jejich hodnot*
- *příprava proměnných pro zvýšení kvality dat*
  - *zpracování odlehlých pozorování*
  - *nahrazení chybějících hodnot*
  - *přeuspořádání kategorií nominální proměnné*
- *přeškálování proměnných*
  - *vážení*
- *transformace na z-skóre, přeškálování vstupníchčíselných proměnných*
- *Box-Coxova tnasformace cílové číselné proměnné*
- *transformace proměnných*
  - *loučení řídce zastoupených kategorií u kategorizovaných proměnných*
  - *rozdělení číselných proměnných do intervalů*
- *provedení výběru proměnných a konstrukce souhrnných proměnných*
- *pojmenování nových proměnných*
  - *transformované a souhrnné proměnné*
  - *vypočtené uplynulé doby*
  - *extrahované části cyklických údajů*
- *aplikace transformací na data*
- *připojení nových proměnných do původního datového souboru nebo vytvoření nového*

Procedura Validate data

Procedura **Validate Data** slouží pro kontrolu správnosti dat pracovního souboru.

- *základní kontroly: specifikace základních pravidel pro proměnné a případy datového souboru. Případ může být označen na základě různých kritérií:*
  - *maximální procento chybějících hodnot*
  - *maximální procento případů v jediné kategorii*
  - *maximální procento případů s četností jedna*
  - *minimální koeficient variability*
  - *minimální směrodatná odchylka*
  - *značení nekompletních identifikačních čísel*
  - *označení duplicitních identifikačních čísel*
  - *označení prázdných případů*
- *standardní pravidla: popisné statistiky a grafy, tvorba a aplikace pravidel pro jednotlivé proměnné.*
  - *popis dat:*
    - *distribuce: miniaturní sloupcové grafy pro kategorizované a histogramy pro spojité proměnné*
    - *zobrazení minim a maxim*
- *pravidla pro jednotlivé proměnné:*
  - *použití pravidel identifikace chybějících nebo neplatných hodnot pro individuální proměnné, například kontrola hodnot mimo škálu*
  - *lze definovat vlastní pravidla pro jednotlivé proměnné*
- *uživatelem definovaná pravidla: pravidla pro logické vztahy mezi proměnnými a hledání případů, které je porušují (například „gravidní muži“)*
- *výstup: reporty popisující nevalidní data*
  - *výpis jednotlivých případů, které porušují pravidla*
    - *specifikace minimálního počtu pravidel, která musí případ porušit, aby byl zahrnut do výpisu*
    - *specifikace maximálního počtu případů ve výpisu*
- *standardní správa validačních pravidel*
  - *sumarizace chyb dle proměnných*
  - *sumarizace chyb dle pravidel*
  - *zobrazení popisných statistik*
- *vytváření nových proměnných: nové proměnné identifikují porušení pravidel v datové matici (kvůli použití při procesu čištění dat a filtrování chybných případů)*
  - *sumární proměnné:*
    - *označení prázdného případu*
    - *označení duplicitních identifikačních čísel*
    - *označení neúplných identifikačních čísel*
    - *celkový počet porušení validačních pravidel*
  - *indikační proměnné, které zaznamenávají všechna* porušení validačních pravidel

Identifikace neobvyklých případů

Procedura **Anomaly Detection** vyhledává neobvyklé případy porovnáváním skupin podobných případů a vrací seznam důvodů, proč je případ odlišný od ostatních.

- *výběr proměnných, které bude procedura využívat pomocí specifikace **VARIABLES**. Specifikuje spojité, kategorizované a identifikační proměnné a seznam proměnných vyloučených z analýzy*
- *specifikace **HANDLEMISSING** určuje metodu zpracování chybějících hodnot*
  - *zpracování vynechaných hodnot: je-li tato volba vybrána, jsou pro účely analýzy vynechané hodnoty spojitých proměnných nahrazeny celkovým průměrem a všechny vynechané hodnoty kategorizovaných proměnných považovány za jedinou platnou kategorii; neníli tato volba vybrána, jsou případy s vynechanými hodnotami vyloučeny z analýzy*
  - *vytvoření přídavné proměnné s názvem Missing Proportion Variable a její použití v analýze: je-li vybrána tato volba, vytvoří se přídavná proměnná, která reprezentuje podíl vynechaných hodnot každého případu; není-li tato volba vybrána, přídavná proměnná se nevytvoří*
- *specifikace **CRITERIA** určuje následující kriteria:*
  - *minimální a maximální počet skupin podobných případů (homogenních klastrů)*
  - *váha pro korekci vlivu odlišných typů proměnných*
  - *počet důvodů v seznamu anomálních případů*
  - *procento případů považovaných za anomální a zahrnutých v seznamu anomálních případů*
  - *počet případů považovaných za anomálnía zahrnutých v seznamu anomálních případů*
  - *hranice indexu anomálie při rozhodování, zda případ je či není považován za anomální*
- *uložení přídatných proměnných do datové matice pomocí specifikace SAVE*
  - *index anomálie*
  - *identifikátor klastru*
  - *velikost klastru*
  - *velikost klastru v procentech*
  - *proměnná, díky které se případ považujeza anomální*
  - *míra anomálie pro výše uvedenou proměnnou*
  - *hodnota výše uvedené proměnné*
  - *standardní hodnota výše uvedené proměnné*
- *specifikace **OUTFILE** uloží model ve formátu XML pod zadaným jménem*
- *nastavení zobrazení výstupů příkazem **PRINT**:*
  - *souhrn všech případů*
  - *seznam indexů anomálií, identifikačních čísel podobných skupin a seznam důvodů, proč je případ považován za anomální*
  - *tabulka standardních hodnot spojitých proměnných (pokud jsou použity pro analýzu) a tabulka stand ardních hodnot kategorizovaných proměnných (pokud jsou použity pro analýzu)*
  - *souhrn indexů anomálií*
  - *přehled důvodů, proč jsou případy považovány za anomální*
    - *potlačení výstupu všech tabulek kromě poznámek a varování*

Optimální kategorizace

Procedura **Optimal Binning** kategorizuje jednu nebo více spojitých proměnných pomocí rozdělení hodnot jednotlivých proměnných do intervalů. Procedura je užitečná pro redukci počtu hodnot vstupní proměnné, která je potřeba kategorizovat, což výrazně zlepšuje kvalitu modelů. Použijete-li určitou metodu optimální kategorizace, řídící (závislá, určující, cílová) proměnná vám pomůže určit hranice intervalů, a tím maximalizovat vztah mezi touto a vstupní kategorizovanou proměnnou.

- *výběr metod:*
  - *nesupervizovaná kategorizace dělí proměnnou na stejně četné intervaly; řídící proměnná není zadána*
  - *supervizovaná kategorizace pomocí **algoritmu MDLP (Minimal Description Length Principle),** diskretizuje vstupní proměnné, je vhodná pro datové soubory s malým počtem případů; řídící proměnná je nutná*
  - *kombinovaná MDLP kategorizace vyžaduje předzpracování pomocí algoritmu založeném na stejných četnostech a následně používá MDLP algoritmus; metoda je vhodná pro datové soubory s velkým počtem případů; řídící proměnná je nutná*
- *nastavení:*
  - *procento případů dolní meze prvního intervalu pro každou kategorizovanou vstupní proměnnou*
  - *horní meze posledního intervalu pro každou kategorizovanou vstupní proměnnou*
  - *zahrnutí nebo vyloučení dolních mezí intervalů*
  - *sloučení řídce zastoupených kategorií*
  - *úplné vyloučení proměnné s vynechanými hodnotami (listwise) nebo vyloučení z aktuálního výpočtu (pairwise)*
- *uložení výsledků výsledky:*
  - *nové proměnné obsahující kategorizované hodnoty*
  - *syntax do souboru IBM SPSS Statistics s příponou *.spo*
- *zobrazení výsledků na výstupu:*
  - *hranice intervalů vstupních proměnných*
  - *popisné statistiky pro všechny vstupní proměnné, které byly kategorizovány*
- *entropie modelu s kategorizovanými proměnnými*