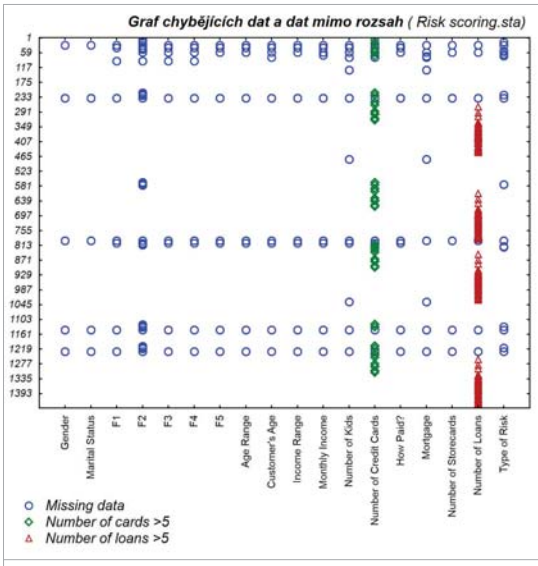
Kvalita dát alebo Čo predchádza analýzu

Kvalita dát je pojem, ktorý nemá jednoznačnú definíciu a ani ju mať nemôže, pretože na kvalitu možno nazerať z mnohých uhlov. Z hľadiska obsahu nás zaujíma predovšetkým úplnosť a správnosť uložených dát. Pri ich prezentovaní je zase prioritou dostupnosť a jednoznačnosť uložených záznamov.



Typické problémy v dátach

Najčastejšie problémy v dátach sú chýbajúce údaje v jednotlivých atribútoch. Ďalší výrazný problém je neaktuálnosť – napríklad e-mailových adries, na ktoré odosielame newsletter, a podobne. Pri číselných údajoch môžu byť prekážkou hodnoty mimo definovaného rozsahu, ako aj duplicitné premenné. Tieto problémy majú vplyv na analýzu dát (od výpočtu priemerných hodnôt až po zložitejšie odhady správania klientov – data mining), pretože nekorektné dáta významne skresľujú závery analýz (o rizikových klientoch, o využití produktov, o skutočných ziskoch atď.). Chybné výstupy z analýz potom môžu mať vplyv na rozhodovanie, napríklad o marketingovej investícii. Zlé rozhodnutia následne môžu viesť k finančným stratám alebo nízkej návratnosti, keď oslovené skupiny klientov nereagujú podľa očakávaní a peniaze na kampaň sú už vyčerpané. K tomu treba pripočítať stratené výnosy, ktoré by sme získali, keby sme pracovali s presnými dátami a oslovili iba najvhodnejšie skupiny klientov.

Konsolidácia dát

V praxi má mnoho spoločností rôzne softvérové nástroje na rôzne oblasti. Dáta z obvykle decentralizovaných systémov sú konsolidované v dátovom sklade spoločnosti. To je základ ich ďalšieho využitia. Konsolidácia nie je jednoduchý proces, ale ak chceme, aby konzultanti a obchodníci uložené informácie využívali, je to nevyhnutný proces. Dáta majú po konsolidácii štandardizovanú formu, čo zvyšuje šancu, že z nich vyťažíme informácie, ktoré nám reálne pomôžu. Uloženie dát v dátovom sklade však ešte neznamená ich jednoduché využitie. Prácu komplikujú napríklad redundantné a často ešte stále neúplné záznamy. Už tu môžu pomôcť štatistické metódy, vďaka ktorým sme schopní odhaľovať hodnoty mimo rozsahu, korelované premenné (ktoré sú duplicitami), neexistujúce kategórie a pod. Tento automatizovaný medzičlánok, ktorý pracuje už pri vstupe dát do dátového skladu, má významný vplyv na čistotu a kvalitu dát, čím výrazne rastie ich využiteľnosť pri ďalších rutinných i pokročilých analýzach.

Príprava dát na analýzu

Či už máme dáta z dátového skladu, alebo z transakčných systémov, prvý krok pred vlastnou analýzou je ich validácia. V nej sledujeme najmä počty chýbajúcich hodnôt v jednotlivých premenných a duplicitné záznamy.

Na tieto operácie slúži množstvo techník, ktoré nájdeme v analytických softvéroch. Spôsob spracovania určuje typ premennej (numerická, kategorická, textová). Podľa povahy dát potom aplikujeme vhodné metódy, ktoré v podstate spadajú do tzv. deskriptívnej štatistiky a základného testovania. Výstupy z analýz sú charakteristiky polohy, variability, výsledky testu odľahlých hodnôt a rozloženie dát, tabuľky početností a nevyhnutná vizualizácia, teda grafy rozpätia, chýbajúcich hodnôt, škatuľové grafy a pod. Každá premenná má svoje logické obmedzenia, ktoré musia jednotlivé záznamy spĺňať. Po prekódovaní chybných kategórií, nahradení alebo vynechaní chýbajúcich hodnôt už máme podstatnú časť práce za sebou. Údaje sú tak pripravené na získanie kľúčových informácií z ich obsahu. Máme teda k dispozícii tzv. analytické dáta.

Ďalšia práca s dátami sa bude líšiť podľa ich povahy a podľa cieľa analýzy, teda podľa toho, čo chceme z dát zistiť. V dátovom súbore ďalej môžeme vytvárať nové premenné, ktoré využívajú časť informácie z už existujúcich premenných. Typický príklad sú dátumy. Ak chceme napríklad porovnať frekvenciu uskutočnených transakcií, môžeme sa na ne pozrieť z pohľadu mesiacov, dní, ale aj v rámci hodín. Pri niektorých výstupoch môže byť žiaduce aj hľadisko minút. Zo základnej informácie vieme odvodiť všetky ostatné.

Z iných číselných premenných zase vytvárame kategórie (intervaly) alebo zisťujeme, či medzi premennými nie sú nejaké silné väzby, či sa nevysvetľujú navzájom, nie sú duplicitné a pod.

Ako teda zabezpečiť kvalitu dát?

Príčinou vzniku nekorektných záznamov v databáze je zvyčajne zle nastavený proces – či už v softvéri, ktorý firma používa, alebo pri jeho obsluhe. Všetky dáta by mali mať svojho vlastníka. Definovanie zodpovednosti má byť prvý krok. Ďalší krok je štandardizácia ukladaných záznamov. Pri importe dát do dátového skladu z transakčných systémov spoločnosti možno využiť štandardizované techniky a ušetriť tým čas s prípravou dát na analýzy. Platí, že závery by sme mali robiť až s „vyčistenými" analytickými dátami. Výstupy pokročilejších metód budú len také dobré, aké dobré sú dáta na vstupe.

FOTO: ARCHÍV STATSOFT

MILOŠ ULDRICH
Autor pôsobí ako odborný konzultant a analytik v spoločnosti StatSoft CR s.r.o. (Dell | Software Group)