ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS
SBORNÍK MENDELOVY ZEMĚDĚLSKÉ A LESNICKÉ UNIVERZITY V BRNĚ

Ročník LVII 4 Číslo 3, 2009

PREDIKCE CHOVÁNÍ ZÁKAZNÍKA PROSTŘEDKY DATAMININGU

N. Chalupová

Došlo: 10. února 2009

Abstract

CHALUPOVÁ, N.,: *Prediction of customer behaviour through datamining assets.* Acta univ. agric. et silvic. Mendel. Brun., 2009, LVII, No. 3, pp. 43–54

Business managers accounting for commercial success or non-success of the organization have to gain knowledge needful for correct decision acceptance. These knowledge represent sophisticated information hidden in enterprise data. One possibility, how to extract mentioned knowledge from data, is to use so-called datamining assets.
The paper deals with an application of chosen basic methods of knowledge discovering in databases for area of customer-provider relation and it presents, how to avail acquired knowledge as basis of managerial decisions leading to improving of customer relationship management. It solves prediction, whose aim is, on the basis of some attributes of exploring objects, to predict future behaviour of objects with these attributes. This way acquired knowledge, as the output of prediction, then can markedly help competent enterprise manager with planning of marketing strategies, for example so-called cross-selling and up-selling. The contribution describes a whole operation of available data processing: from its purifying, over its preparation for mining task, to self processing by the help of SAS Enterprise Miner tool. Regression analysis, neural network and decision tree, whose principles are briefly explained in this paper too, were used for knowledge mining. The estimation of customer behaviour was tested by two mining task varying in attribute using and in categories number of one of predicive attributes. The results of these two tasks are confronted by the help of prediction fruitfulness charts.

knowledge discovery in databases, datamining, prediction, customer, decision process, control

Chce-li firma obstát v současném konkurenčním prostředí trhu, je nezbytné, aby sledovala chování svých zákazníků. Za obchodní úspěch či neúspěch organizace odpovídají podnikoví manažeři, a ti proto musí získávat znalosti potřebné pro přijetí správného rozhodnutí. Tyto znalosti představují sofistikované informace ukryté v datech, která má podnik k dispozici. Novotný, Pour a Slánský (2005) uvádějí, že objem dat se v podniku zdvojnásobí v průměru každých pět let, což znamená, že v současné době již není problém data získat a uchovat, ale efektivně je zpracovat a využít jejich potenciál.

Možností, jak zmiňované znalosti z dat získat, je využít prostředků tzv. datamineringu. Tento obor se zabývá otázkami, jak nalézt v datech souvislosti, které nejsou přímo zřejmé, a které napomáhají lépe porozumět firemním procesům. Jednou z významných úloh datamineringu je predikce, jejímž cílem je na základě určitých vlastností zkoumaných objektů předpovědět budoucí chování objektů s těmito vlastnostmi. Výsledný výstup (předpověď) pak může výrazně napomoci při tzv. *křížovém prodeji* – snahy, jejichž účelem je navýšit objednávku zákazníka doporučením jiných produktů nabízených společností (Clemente, 2004) a *následném prodeji* – aktivity, jejichž cílem je nabídnout zákazníkovi vyšší/pokročilejší, a tedy i dražší model/verzi produktu) (Parr Rud, 2001). Nalezené znalosti lze využít například při plánování marketingových strategií přesnějším zacílením kampaně, při péči o zákazníky apod.

Cílem článku je aplikovat vybrané základní metody získávání znalostí z databází na oblast vztahu zákazníka a obchodu a získanou znalost využít s ohledem na vztah k řešenému problému – jako podklady pro manažerská rozhodnutí vedoucí ke zlepšení řízení vztahu se zákazníky.

MATERIÁL A METODY

Jako nástroj realizace predikce v následující úloze byl použit software Enterprise Miner společnosti SAS Institute Inc. Pro projekty dolování dat implementuje metodiku tzv. SEMMA, podle níž kroky analýzy zahrnují (SAS Institute Inc., 2008; Berka, 2003):

- výběr vhodného vzorku zkoumaných objektů (Sample)
- prozkoumání struktury dat (Explore)
- datové transformace (Modify)
- analýza dat – vytvoření modelu dat (Model)
- zhodnocení využitelnosti modelu, interpretace výsledků (Asses).

Dostupná jazyková verze tohoto nástroje byla pouze v anglické mutaci, proto i textové popisky ve schématech či grafech (generovaných použitým nástrojem) prezentovaných v této práci jsou anglické. Vše je však patřičně vysvětleno v textu článku.

Zdrojová data

K realizaci dolovací úlohy byla použita data poskytnutá Ing. Ladislavem Stejskalem, Ph.D., partnerem a koordinátorem šetření Omnibus 2007 za Ústav marketingu a obchodu Provozně ekonomické fakulty Mendelovy zemědělské a lesnické univerzity v Brně.

Popis a obsah dat

Zpracovávaná data reprezentují odpovědi dotazovaných respondentů na jednotlivé otázky z Dotazníku pro občany v rámci šetření OMNIBUS 2007. Jedná se o dotazníkové šetření pořádané Vysokou školou evropských a regionálních studií, o.p.s. v Českých Budějovicích ve spolupráci s Českou zemědělskou univerzitou v Praze, Západočeskou univerzitou v Plzni, Vysokou školou polytechnickou v Jihlavě, Mendelovou zemědělskou a lesnickou univerzitou v Brně, Stredoeurópskou vysokou školou ve Skalici a Slovenskou poľnohospodárskou univerzitou v Nitre.

Cílem uvedeného šetření je zjištění názorů občanů na otázky týkající se zejména problematiky investičního rozhodování, regionálního rozvoje a veřejné správy, spotřebitelského chování, trhu cestovního ruchu a trhu potravin.

Data, která jsou zpracovávána v rámci této dolovací úlohy, mají podobu jedné tabulky o necelých stopadesáti sloupcích a více než dvou tisících řádcích. Každý záznam (řádek) představuje jeden vyplněný dotazník. Jednotlivé atributy (sloupce) představují konkrétní odpověď respondenta na určitou otázku v dotazníku a jsou systematicky pojmenovány podle toho, které části výzkumu se týkají:

- první znak prefixu (písmeno) udává, ke které části šetření se váže atribut, jehož prefix začíná tímto písmenem (sled otázek v dotazníku lze jednoduše rozdělit na skupiny otázek týkajících se jednotlivých výše uvedených problematik šetření)
- druhý znak prefixu (číslo) představuje pořadí otázky v příslušné části šetření
- zbylá část názvu by měla symbolicky zachytit obsah otázky.

Například názvy atributů D4_penzijni a D4_stavebni, D4_vkladni, atd. znamenají, že se jedná o odpovědi na otázky týkající se investičního rozhodování. Tyto atributy mohou také být pouze částí odpovědi, a to v případě, že v odpovědi bylo možné vybrat více variant nebo určit důležitost varianty – každá varianta představovala jeden atribut, který mohl nabývat více hodnot.

Data byla získána pomocí několika technik sběru dat, např. papírové dotazníky, různé varianty webových formulářových dotazníků (každá instituce podílející se na výzkumu shromažďovala data do svých databází). Z této skutečnosti pak pramenila potřeba sjednotit podobu dílčích datových zdrojů.

Předzpracování dat

Z důvodu zmíněné různorodosti zdrojů a i dalších nedostatků v datech bylo nutné všechna data soustředit do jediného zdroje a nadále je upravit. Snahou těchto transformací bylo upravit data do jednotného formátu (struktury) vhodného pro dolování.

Nejvíce frekventovaným jevem, který se v datech vyskytoval, byl tzv. konflikt hodnot – případ, kdy hodnoty, které znamenají stejnou věc, jsou různé. Pravděpodobně nejčastěji byl způsobený skutečností, že jako odpověď na otázku bylo možné uvést něco jiného, než bylo obsaženo v nabízených možnostech. Tento problém byl manuálně řešen jedním ze způsobů:

- podobné odpovědi byly sdruženy do jedné „nové možnosti“ (např. v otázce zaměřené na nejčastěji využívané způsoby dopravy byla často zvolena možnost „jiné – uveďte“ a zde byly uvedeny „vlak“, „vlak + bus“, „ČD“, „ČSAD“ apod., které byly sloučeny do nové kategorie „další hromadná doprava“)
- odpovědi, které byly velmi blízké některým z nabízených možností, byly zahrnuty pod nabízenou možnost.

Nežádoucím jevem v datech byly různé logické chyby, například v části dotazníku zjišťující od respondenta základní identifikační údaje, docházelo k tomu, že v jedné otázce bylo zadáno státní občanství a v jedné z dalších otázek, nezávisle na výše uvedené odpovědi, vybrán region bydliště, přičemž bylo možné jako státní občanství zadat např. Českou republiku a zároveň z regionů vybrat např. Bratislavský kraj. Tento nesoulad bylo naštěstí možné ve většině případů odstranit dohledáním regionu bydliště respondenta podle uvedené obce a upravením příslušných atributů (špatně uvedeného státu nebo kraje) – jiná část šetření se totiž zabývala spokojeností s různými oblastmi života v místě bydliště respondenta a toto bydliště zde bylo také uvedeno. Tímto způsobem často byly i doplněny některé chybějící hodnoty atributů, které bylo možné odvodit z atributů jiných.

Z určitých skupin dat byly odstraněny další nesrovnalosti způsobené integrací několika zmiňova-

ných datových zdrojů. V některých skupinách dat bylo u příslušného atributu uvedeno „ano“ nebo „ne“ (označený příslušný checkbox ve webovém formuláři), v jiných podmnožinách dat byly tyto atributy prázdné a jiný atribut obsahoval souhrnnou odpověď – řetězec obsahující označení jednotlivých položek vybraných respondentem (např. mezerami či jinak oddělená písmena a, b, c atd.). Z těchto řetězců byla tato jednotlivá označení (písmena) extrahována a do příslušného sloupce přenesena správná hodnota – např. v MS Excelu v buňkách příslušného sloupce funkcí `=KDYŽ(JE.CHYBHODN(NAJÍT(„a“; <buňka_s_řetězcem>; 1)); „“; „ano“)`.

Použitím nejen uvedených způsobů vedoucích k vyčištění a zhodnocení dat se však všechna negativa odstranit nepodařilo. Pro predikci chování zákazníka v úloze popisované dále ale byly použity pouze atributy, jejichž negativa bylo možné odstranit. Zmiňované nedostatky se objevují v atributech, jejichž hodnoty z převážné většiny nebylo možné zařadit do několika (max. deseti) kategorií. Takovými jsou např. uvedení různých názorů nebo zdůvodnění spokojenosti či nespokojenosti zákazníka s produktem.

Analytické metody

Jak bude uvedeno dále, pro získání znalostí užitečných v oblasti řízení vztahu se zákazníky bylo využito regresní analýzy, neuronových sítí a rozhodovacího stromu, jejichž principy jsou stručně vysvětleny v následujícím textu.

Jednotlivé metody člení různí autoři podle různých hledisek. Pro účely této práce je dostačující rozdělení na (Parr Rud, 2001):

- statistické (lineární regrese, logistická regrese atd.)
- nestatistické (neuronové sítě, genetické algoritmy) a
- smíšené (rozhodovací stromy, bayesovská klasifikace a další).

Statistika nabízí celou řadu teoreticky dobře prozkoumaných a zdůvodněných a léty praxe ověřených metod pro analýzu dat. Statistickou analýzou experimentálních dat se podrobně zabývají například Meloun a Militký (2006). Dostatečné vysvětlení principů použitých statistických metod lze nalézt v publikacích Benjamini a Leshno (2005), Han a Kamber (2006) či Parr Rud (2001). Poslední dvě zmiňované publikace seznamují čtenáře také se základy použitých nestatistických a smíšených metod. Podrobněji se neuronovým sítím ve vztahu k dataminingu věnuje Zhang (2005), jejich aplikací se zabývají Dostál, Rais a Sojka (2005). Z kategorie smíšených metod byl použit rozhodovací strom. Principy této metody a její využití v dataminingu rozebírají, kromě posledních dvou již uvedených, také Rokach a Maimon (2005).

Regresní analýza

Regresní analýza je označení statistických metod, jejichž pomocí se odhaduje hodnota jisté náhodné veličiny (takzvané závisle proměnné, cílové proměnné, regresandu anebo vysvětlované proměnné) na základě znalosti jiných veličin (nezávisle proměnných, kovariát, regresorů anebo vysvětlujících proměnných). Snahou regresní analýzy je nalézt takovou „idealizující“ matematickou funkci, aby co nejlépe vyjadřovala charakter závislosti a co nejvěrněji zobrazovala průběh změn podmíněných průměrů závisle proměnné. Jde tedy o řešení úlohy aproximace pozorovaných hodnot daným typem funkce, ovšem s neznámými parametry.

Prostá lineární regresní analýza je statistickou metodou, která kvantifikuje závislost (hledány jsou parametry této závislosti) mezi dvěma spojitými proměnnými: závislou proměnnou čili proměnnou, která je predikována, a nezávislou, tedy prediktivní proměnnou (jejíž hodnoty slouží k predikci). V průběhu lineární regrese je hledána taková přímka procházející mezi jednotlivými body, pro niž platí, že součet druhých mocnin odchylek od každého bodu je co nejmenší.

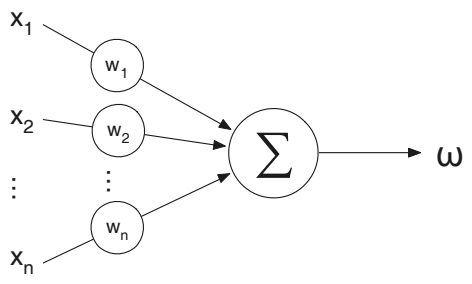
Nelineární regrese je velmi podobná lineární. V tomto případě je ale předpokládána složitější funkční závislost mezi nezávislou a závislou proměnnou. Proto je zpravidla nutné transformovat prediktivní (nezávislou) proměnnou tak, aby umožňovala najít lepší proložení (křivkou kvadratickou, obecně polynomickou, exponenciální atd.).

Logistická regrese je zajímavým případem nelineární regrese. Liší se hlavně v tom, že závislá proměnná není spojitá, je diskrétní neboli kategoriální. Tím se logistická regrese stává velmi užitečnou například v marketingu, protože v této oblasti je často vyvíjeno úsilí předvídat diskrétní akci, například odezvu na nabídku nebo nesplácení půjčky (Parr Rud, 2001). Proto se zde nemodeluje přímo závislá (diskrétní) veličina, ale spojitá hodnota reprezentující funkci pravděpodobnosti výskytu události neboli pravděpodobnost, že tato veličina má konkrétní hodnotu v závislosti na kombinaci hodnot nezávislých veličin.

Neuronová síť

Umělé neuronové sítě jsou velmi zjednodušeným modelem myšlení lidského mozku. Jeho činnost je umožněna neuronovými propojeními, která vytvářejí život člověka včetně jeho procesu učení.

Z biologické interpretace funkce neuronu byla sestavena jednoduchá varianta matematické interpretace neuronu (perceptron) (Obr. 1), které jsou pak navzájem propojovány v síť.



1: *Matematický model neuronu*

U neuronových sítí nelze znát detailně vnitřní strukturu systému, je na něj kladeno pouze několik předpokladů, jež umožní popsat chování systému funkcemi, které provádějí transformaci vstup – výstup.

Principem neuronových sítí je nastavení parametrů jednotlivých neuronů v procesu učení se z tréninkových dat tak, aby výsledná konfigurace co nejlépe vyhovovala následné klasifikaci a predikci. Základní logickou jednotkou je neuron – uzel, do kterého vstupují údaje (x_1, x_2, …, x_n na Obr. 1) buď z vnějšího vstupu sítě nebo z výstupu jiného neuronu. Z tohoto uzlu vystupují zpracované údaje (ω na Obr. 1) do dalšího neuronu nebo na výstup sítě.

Umělá neuronová síť pracuje ve dvou fázích. V první vystupuje síť (model složitého systému) v roli „zvídavého žáka“, tj. učí se nastavit své parametry tak, aby co nejlépe vyhovovaly požadované topologii sítě. Ve druhé fázi se stává síť „odborníkem“, neboť produkuje výstupy na základě znalostí získaných v první fázi.

Při konstrukci každé neuronové sítě je nutné definovat její jednotlivé vrstvy (vstupní, skryté, výstupní), jednotlivé vstupní a výstupní neurony, způsoby propojení neuronů navzájem mezi sebou (formulace přenosové funkce neuronů mezi skrytými vrstvami – Σ na Obr. 1), způsob její výuky (bez učitele, s učitelem, v epochách) a proces získávání poznatků. Dříve než začne vlastní proces zpracování, rozdělují se data na tři skupiny: trénovací, testovací a data pro finální validaci. Poté se každému z uzlů v první vrstvě (vstupní vrstva) přiřadí váhy (w_1, w_2, …, w_n na Obr. 1). Během každé iterace jsou vstupy (tréninková data) zpracovány systémem a porovnány se skutečnou hodnotou (testovací data). Změří se chyba a předá se ke zpracování systému, aby upravil původní váhy. Proces končí v okamžiku, kdy je dosaženo předem určené minimální chyby.

Neuronové sítě je vhodné použít v případě, ve kterém značnou roli v modelovaném procesu hra je náhoda a deterministické závislosti jsou natolik složité a provázané, že je nelze separovat a analyticky identifikovat, což představuje jednu z výhod neuronové sítě (Dostál, Rais, Sojka; 2005). Další výhodou je skutečnost, že neuronová síť díky své koncepci (informace v procesoru i paměti je rozprostřena po celé síti) může fungovat při neúplných nebo zašuměných datech (Berka, 2003). Naopak nevýhodou může být tendence sítě přizpůsobovat si data příliš, čímž model při aplikaci na nová data zastarává (Parr Rud, 2001), nebo na první pohled hůře srozumitelné nalezené znalosti (Berka, 2003).

Rozhodovací strom

Cílem rozhodovacích stromů je nalezení pravidel a vztahů v datovém souboru pomocí systematického rozdělování a větvení na nižší úrovně. Jsou snadno interpretovatelné, což umožňuje uživatelům rychle a jednoduše vyhodnocovat získané výsledky, identifikovat klíčové položky a vyhledávat zajímavé segmenty případů.

Při tvorbě rozhodovacího stromu se postupuje metodou „rozděl a panuj“ (divide and concquer). Trénovací data se postupně rozdělují na menší a menší podmnožiny tak, aby v těchto podmnožinách převládaly příklady jedné třídy. Od kořene stromu se na základě odpovědí na otázky (umístěné v nelistových uzlech) postupuje příslušnou větví stále hlouběji, až do listového uzlu. V hieratické struktuře jsou tedy nejvýše umístěné uzly, které mají na podřízené uzly největší vliv, respektive nejvíce odlišují příhlady různých tříd.

Klasifikační stromy rozdělují objekty popsané různými atributy do tříd, které jsou ve stromu reprezentovány listovými uzly. Užitečné jsou v oblastech, ve kterých lze hodnoty proměnných rozčlenit do relativně malého počtu skupin. Na druhou stranu nejsou vhodné pro případy, kdy je úkolem předpovězení kvantitativních hodnot.

Regresní stromy umožňují odhadnout hodnoty nějakého numerického atributu. V listových uzlech mají takové stromy například konstantu, která odpovídá průměrné hodnotě cílového atributu pro příklady v tomto uzlu.

VÝSLEDKY A DISKUSE

V dolovacích úlohách je nutné vědět, jaký druh modelu dat se snažíme z dat vytěžit. Typy dolovacích úloh se dělí do dvou základních skupin (Han, Kamber, 2006):

- deskriptivní – charakterizují obecné vlastnosti analyzovaných dat
- prediktivní – na základě analýzy současných dat provádějí dedukci pro předpověď budoucího chování.

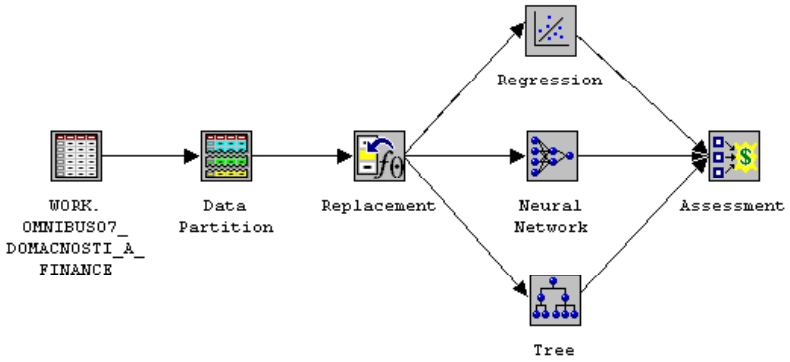
Úloha řešená v rámci této práce, již podle svého názvu, patří do kategorie prediktivních dolovacích úloh, neboť jejím cílem je zjistit nejpravděpodobnější budoucí chování objektů na základě vlastností těchto objektů.

Rozbor a popis úlohy

Úloha je zpracována z pohledu společností poskytujících své služby v odvětví investičních a spořicích nástrojů. Predikce chování jejich zákazníků pro ně může znamenat významné úspory nákladů na různé reklamní kampaně, které mohou přesněji zacílit právě na ty klienty, pro které budou s vysokou pravděpodobností nabízené produkty zajímavé, a proto na kampaň pravděpodobně pozitivně zareagují. Tím by bylo možné eliminovat zbytečné zasílání nabídek těm klientům, kteří by si pravděpodobně tyto nabídky ani nepřečetli a ihned je vyhodili.

Následující obrázek (Obr. 2) znázorňuje blokové schéma uvedené dolovací úlohy.

Komponenta `WORK.OMNIBUS07_DOMACNOSTI_A_FINANCE` symbolizuje zdroj dat pro dolování a zajišťuje nahrání těchto dat do úlohy a nastavení rolí jednotlivých proměnných v modelu (zejména jde o to, které proměnné do modelu vstupují a které jsou cílové). Pro dolování byl jako cílová proměnná zvolen údaj o tom, zda zákazník má či nemá stavební spoření, proměnnými, ze kterých se odvozoval cí-



2: *Blokové schéma úlohy predikce*

lový atribut (vstupní proměnné), byly sociodemografické atributy věk, vzdělání a příjmová skupina klienta a dále také vlastnictví (či nevlastnictví) důchodového spoření.

V sekci `Data Partition` jsou zdrojová data náhodně rozdělena do tří skupin. Jejich velikosti jsou pro tuto úlohu nastaveny takto:

- Jedna skupina obsahuje polovinu všech dat a je určena k učení prediktorů (učí se z výsledků),
- druhá skupina obsahující čtvrtinu dat je použita k validaci (prediktory kontrolují své výpočty podle výsledků) a
- třetí skupina obsahující zbývající čtvrtinu dat slouží k testování (prediktory již pouze predikují).

Prvek `Replacement` nahrazuje chybějící hodnoty podle různých kritérií (nahrazení nejčastěji se vyskytující hodnotou ve zbytku dat, nahrazení průměrnou hodnotou zbytku dat apod.). Její použití je nutné v případě, že data s chybějícími hodnotami dále zpracovává neuronová síť (komponenta `Neural Network`) nebo regresní analýza (komponenta `Regression`), což je případ této úlohy. Rozhodovací strom (komponenta `Tree`) umí nekompletní záznamy ignorovat – zde není potřeba předzpracovat data složkou Replacement.

Již zmiňované komponenty Neural Network, Regression a Tree reprezentují analytické metody pro zpracování definovaných dat.

Poslední člen, `Assessment`, vyhodnocuje výsledky dodané analytickými metodami.

Pro odhad chování zákazníka – zda si sjedná nebo nesjedná stavební spoření, byly vytvořeny dvě dolovací úlohy lišící se v použití atributů a v počtu kategorií atributu týkajícího se vzdělání respondenta (viz dále).

Příprava dat

Před spuštěním vlastního dolování byly vybrány relevantní sloupce z tabulky zdrojových dat, tzn. proměnné H2_vek, H3_vzdelani, H5_prijem, D4_penzijni a D4_stavebni. Hodnoty posledních dvou jmenovaných atributů byly navíc transformovány tak, že hodnota „ano“ reprezentující skutečnost, že daný respondent má sjednané stavební spoření, byla nahrazena hodnotou „1“ a hodnota „ne“ (respondent nemá stavební spoření) nahrazena hodnotou „0“. Tato transformace byla provedena proto, aby byla predikována žádaná událost, což je situace, kdy klient vlastní stavební spoření – prediktory jsou totiž nastaveny tak, že predikují maximální hodnotu (a použitý software neumožnil toto nastavení změnit). V původních datech tudíž byla stále jako cílová hodnota predikována hodnota „ne“, což je z hlediska využití výsledků predikce nezajímavá událost.

Proměnná H5_vzdelani byla nejprve použita v nezměněné podobě, ve druhé predikční úloze nebyl zohledněn typ středoškolského vzdělání, tzn. že hodnoty „střední odborné“ a „střední všeobecné“ byly nahrazeny hodnotou „středoškolské“.

Zdroj dat použitý v této úloze pak obsahoval šest sloupců a více než dva tisíce (přesně 2017) řádků. Větší přehled o analyzovaných datech lze získat z obrázků č. 3–7, znázorňujících rozložení četností hodnot jednotlivých proměnných.

Rozhodnutí, na jaké atributy zaměřit predikci (stavební spoření), bylo ovlivněno právě výše uvedeným histogramem proměnné D4_stavebni, který vypovídá o tom, že vzorek relevantních dat (těch, která obsahují cílovou hodnotu cílového atributu – hodnotu „ano“ či po transformaci „1“ proměnné D4_stavebni) je dostatečně velký, a tudíž by i po rozdělení dat (pro učení, validaci a testování) mělo být možné dostatečně úspěšně naučit prediktory předpovídat cílovou hodnotu.

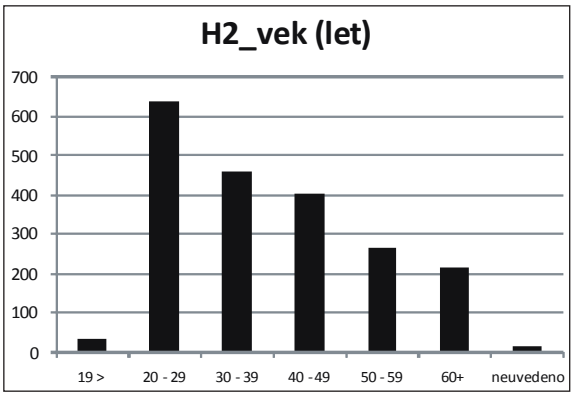
Výsledky dolování a jejich interpretace

Jak je uvedeno výše, pro odhad chování zákazníka byly vytvořeny dvě dolovací úlohy:

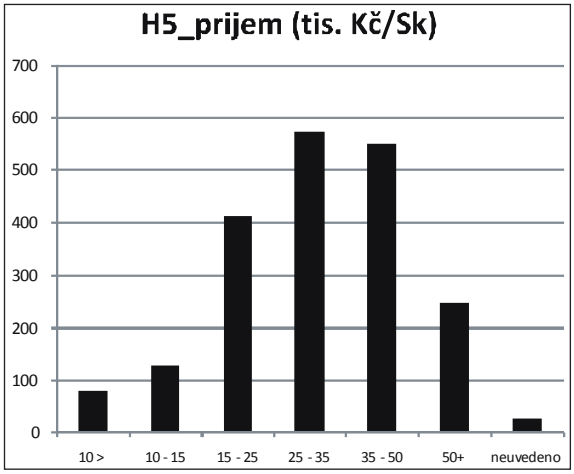
„původní úloha“ – predikce bez vlivu penzijního spoření a s rozlišením typů středoškolského vzdělání (vstupními atributy jsou věk, příjmová skupina a vzdělání v původní podobě)

„upravená úloha“ – predikce s vlivem penzijního spoření a bez rozlišení typů středoškolského vzdělání (vstupními atributy jsou věk, příjmová skupina, vlastnictví penzijního spoření a vzdělání v „redukované“ podobě s rozlišením).

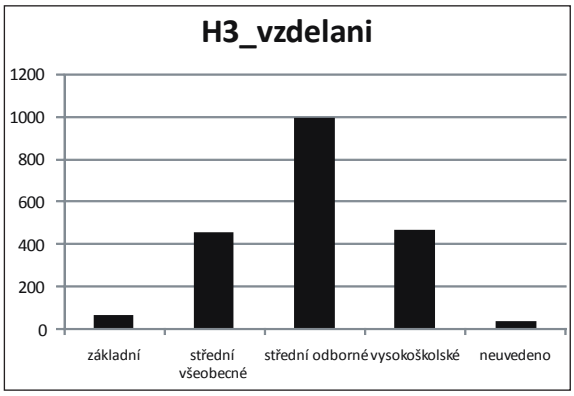
Následuje konfrontace výsledků jednotlivých použitých metod (rozhodovací strom, neuronová síť a regresní analýza) v těchto úlohách.



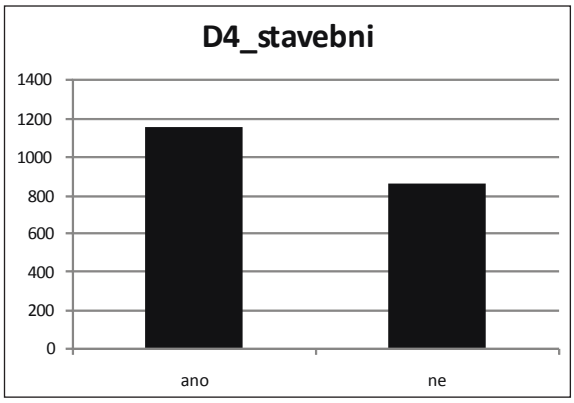
3: *Rozložení věkových skupin respondentů*



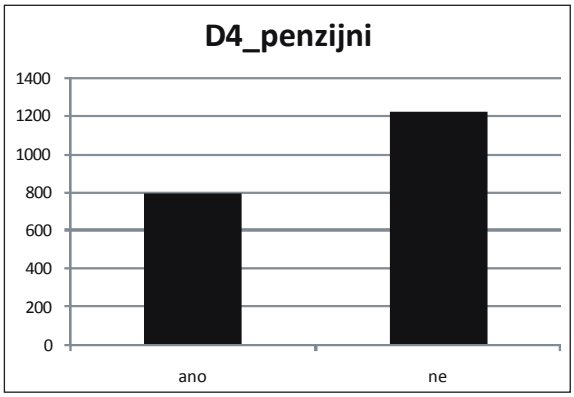
4: *Rozložení příjmových skupin respondentů*



5: *Rozložení respondentů podle úrovně dosaženého vzdělání*



6: *Rozložení respondentů podle vlastnictví stavebního spoření*



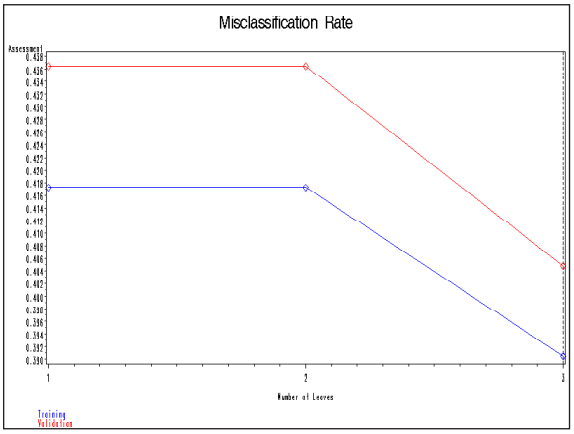
7: *Rozložení respondentů podle vlastnictví důchodového spoření*

Rozhodovací strom

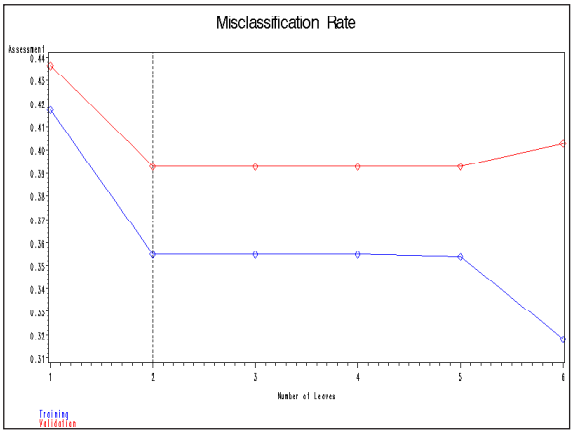
Prvním použitým prostředkem byl rozhodovací strom, jehož grafy učení a validace jsou znázorněny na obrázcích č. 8 a 9.

Z obou grafů (Obr. 8 a 9) chybovosti lze vyčíst, že upravená úloha vykazuje nepatrně lepší výsledky míry chybovosti než původní úloha. V původní úloze byl dosažen rozhodovací práh 0.40 po třech iteracích. V upravené úloze byla chyba minimalizována o jednu iteraci dříve, kde její hraniční bod činil 0.39 a nadále již tato chyba téměř nevzrostla (červená křivka „Validation" v grafech).

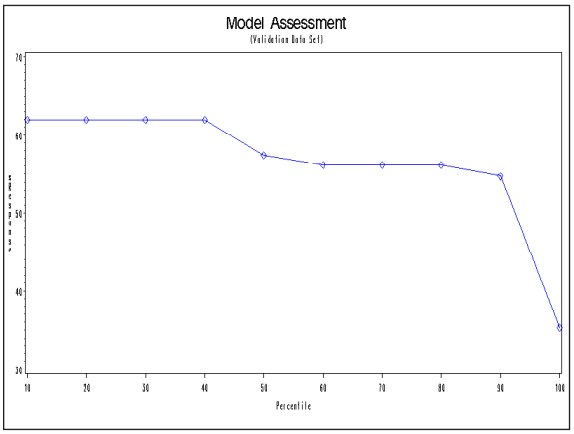
Z grafů na Obr. 10 a 11 reprezentujících přesnosti odhadů v jednotlivých dolovacích úlohách je patrné, že upravená úloha vykazuje opět o trochu lepší výsledky než původní úloha.



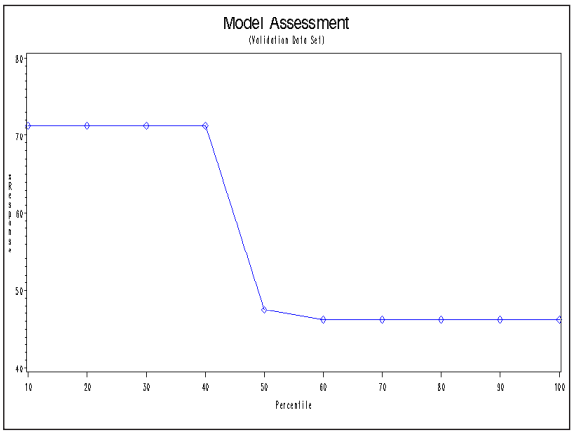
8: *Dělení větví stromu v původní úloze*



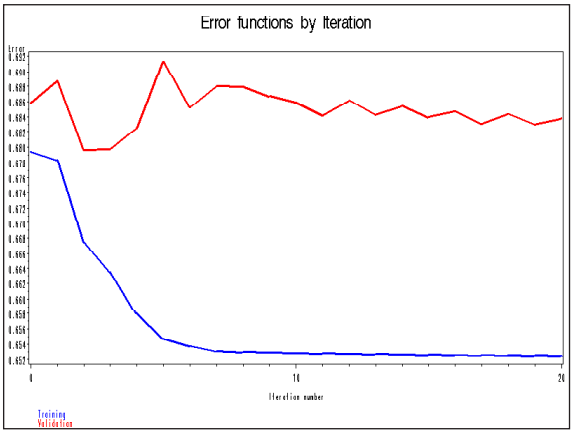
9: *Dělení větví stromu v upravené úloze*



10: *Přesnost predikce v původní úloze*



11: *Přesnost predikce v upravené úloze*



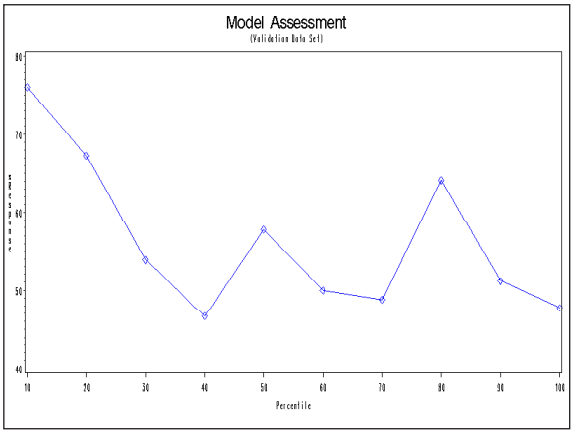
12: *Chybová funkce v původní úloze*

13: *Chybová funkce v upravené úloze*

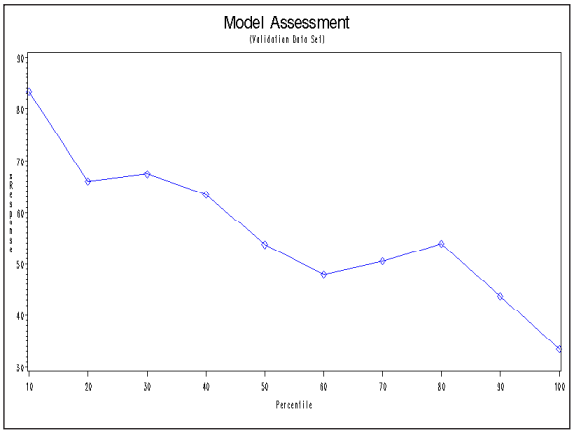
Odezva v původní úloze pro 40% dat činila 62%, poté mírně klesla na cca 55% a na této hodnotě se dlouhou dobu držela. V upravené úloze byla odezva pro stejnou část dat přibližně o 10% lepší, avšak poté náhle klesla hlouběji na necelých 50%, na kterých již zůstala. Výsledné grafy lze interpretovat tak, že zhruba o 40% respondentech lze s přibližně 60% (potažmo 70% v upravené úloze) pravděpodobností říci, že si uzavřou stavební spoření.

Neuronová síť

Druhou použitou analytickou metodou byla neuronová síť, jejíž průběhy učení a validace jsou zachyceny obrázky č. 12 a 13, z nichž je patrné (červená



14: *Přesnost předpovědi v původní úloze*



15: *Přesnost předpovědi v upravené úloze*

křivka „Validation“), že ani jedna z úloh nevykazuje příliš dobré výsledky: nejnižší hodnota chyby v první úloze se pohybuje kolem 0.68 a v druhé úloze slabě pod 0.67, což je nepatrný rozdíl. V upravené úloze je alespoň zpočátku průběh míry chybovosti mírně stabilnější než v původní úloze.

Na grafech (Obr. 14 a 15) jsou vyobrazeny průběhy přesností odhadu neuronové sítě v obou dolovacích úlohách.

Odezva predikce v původní úloze začíná na cca 75%, v upravené úloze začíná na přibližně o 10% lepší hodnotě, postupně však v obou případech klesá na 50% v původní úloze a na 40% v upravené úloze. Zde nelze jednoznačně říci, který průběh je lepší, neboť sice v upravené úloze je zpočátku odezva lepší než v upravené úloze, ale tato lepší odezva postupně klesá na horší hodnoty než v původní úloze.

Regresní analýza

Posledním použitým prediktorem je regresní analýza, konkrétně se jedná o logistickou regresi. Její grafy vypočtených vah jsou znázorněny pomocí grafů (Obr. 16 a 17), z nichž lze vyčíst, že původní úlohu nejvíce ovlivnily dva vstupní parametry: věk a příjmová skupina, v úloze upravené o použití atributu vlastnictví penzijního připojištění tento atribut také sehrál výraznou roli. Vzdělání, ať již redukované na menší počet kategorií či ponechané v původní podobě, predikci ovlivnilo nejméně ze všech použitých atributů.

V dalších vypočtených statistikách modelu byla míra chybovosti původní úlohy na úrovni 0.4345238095, u upravené úlohy činila 0.4027777778, což je nepříliš významné zlepšení.

Grafy na Obr. 18 a 19 zobrazují přesnosti odhadu v jednotlivých dolovacích úlohách.

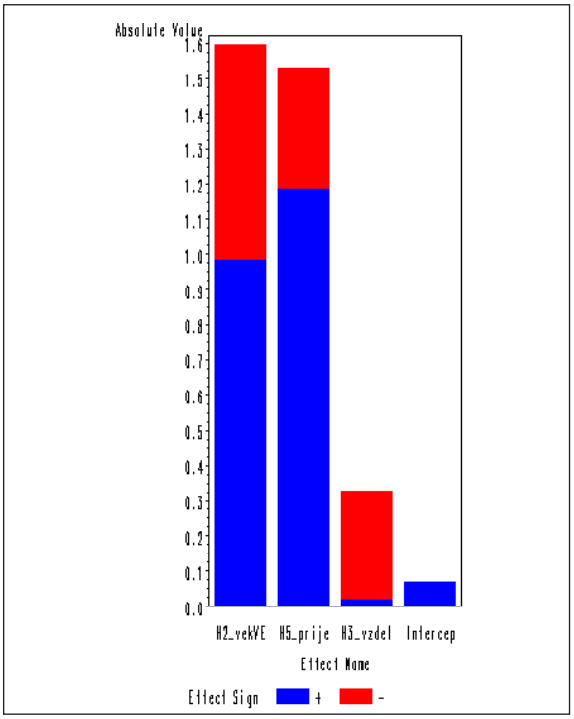
V upravené úloze je průběh sice „hladší“ – má méně výkyvů, ale má větší rozsah hodnot: úspěšnost předpovědi začíná na 85%, což není špatné, ale záhy klesá až na hodnotu téměř 30%, což je velmi špatný výsledek. Odezva v původní úloze sice začíná na o 10% horší hodnotě, ale klesá pozvolněji na o 10% lepší hodnotu než v upravené úloze, což ale také není příliš optimistický výsledek.

ZÁVĚR

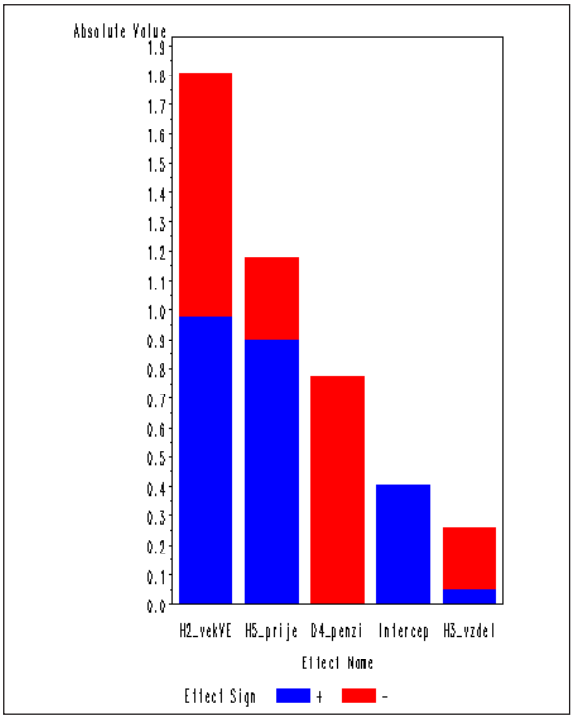
Nejstabilnější výsledky v obou úlohách vykazoval rozhodovací strom. Neuronová síť i logistická regrese především v upravené úloze měly téměř stejný průběh – jejich přesnost předpovědi začínala lepší hodnotou než rozhodovací strom, ale bohužel postupně odezva klesala k velmi nízkým hodnotám.

Celkově lze tedy konstatovat, že původní úloha není příliš vhodná pro predikci (jako nejvhodnější metoda predikce zde byla neuronová síť), jejím upravením ale bylo dosaženo mírně lepších výsledků, což by mohlo znamenat, že další podobné úpravy (zejména přidání vlivu dalších souvisejících atributů) by mohly přesnost predikce ještě vylepšit. Jako nejlepší metoda predikce se zde jeví rozhodovací strom. Uvedené skutečnosti demonstrují obrázky grafů č. 20 a 21.

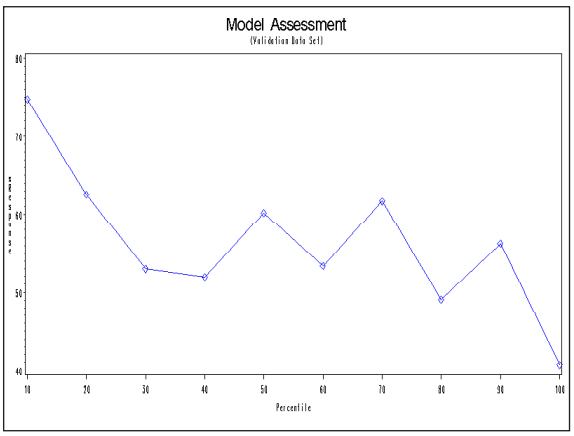
Je však nutné podotknout, že tento článek pouze demonstruje možnosti využití prostředků dataminingu v oblasti řízení vztahu se zákazníky (CRM – Customer Relationship Management). Důležité je také zdůraznit skutečnost, že ač se jeví použité nástroje a metody jako nepříliš účinné, nemusí to být a pravděpodobně to ani není způsobeno „neschopností“ (jakožto negativní vlastností) těchto technologií. V tomto případě je uvedená nízká účinnost patrně způsobena podstatou zkoumaných dat – jedná se totiž o výsledky dotazníkového šetření, kde není možné ošetřit pravdivost odpovědí nebo náhodnost toho, že někdo v dotazníku „něco rychle nakliká (či zaškrtne)“ na rozdíl například od reálných podnikových dat, která obsahují skutečné výsledky chování zákazníků (jejich skutečné nákupy, placení apod.)



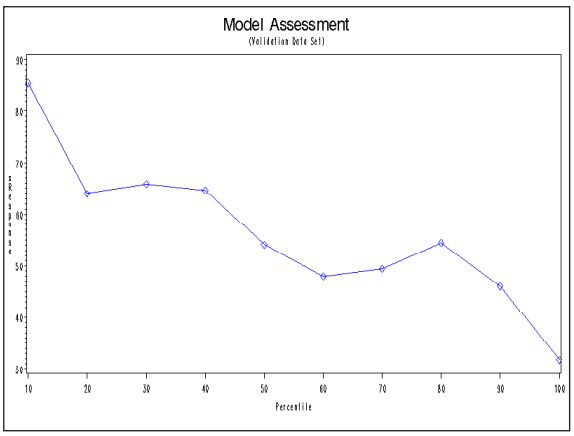
16: *Váhy parametrů v původní úloze*



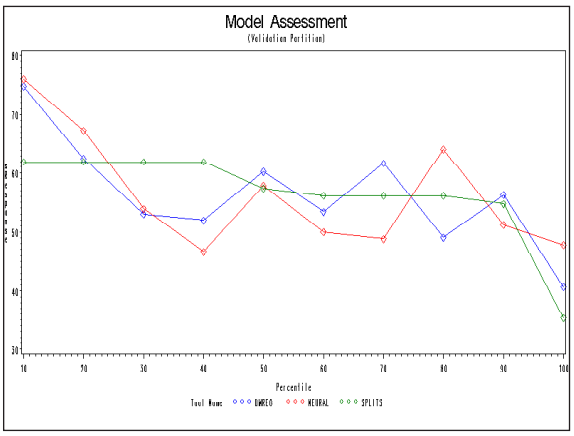
17: *Váhy parametrů v upravené úloze*



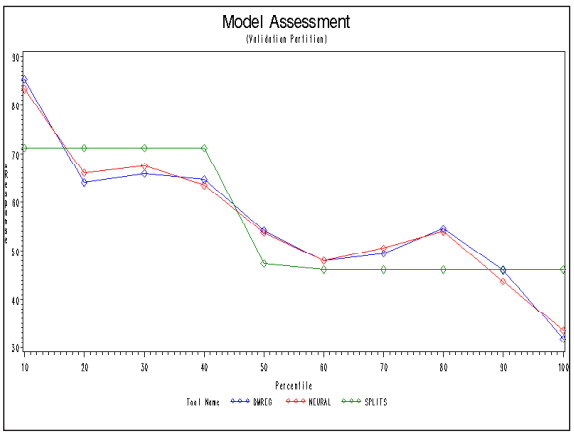
18: *Přesnost prediktoru v původní úloze*



19: *Přesnost prediktoru v upravené úloze*



20: *Výsledky predikce původní úlohy*



21: *Výsledky predikce upravené úlohy*

SOUHRN

Pokud chce podnik obstát v dnešním konkurenčním prostředí trhu, je nutné, aby sledoval chování svých zákazníků. Za obchodní úspěch či neúspěch organizace odpovídají podnikoví manažeři, kteří proto musí získávat znalosti potřebné pro přijetí správného rozhodnutí. Tyto znalosti představují sofistikované informace ukryté v datech, která má podnik k dispozici. Uvádí se, že objem dat se v podniku zdvojnásobí v průměru každých pět let, což znamená, že v současné době již není problém data získat a uchovat, ale efektivně je zpracovat a využít jejich potenciál. Možností, jak zmiňované znalosti z dat získat, je využít prostředků tzv. dataminingu.

Článek se zabývá aplikací vybraných základních metod získávání znalostí z databází na oblast vztahu zákazníka a obchodu a prezentuje, jak získanou znalost využít s ohledem na vztah k řešenému problému – jako podklady pro manažerská rozhodnutí vedoucí ke zlepšení řízení vztahu se zákazníky. Konkrétně řeší predikci, jejímž cílem je na základě určitých vlastností zkoumaných objektů předpovědět budoucí chování objektů s těmito vlastnostmi. Takto získaná znalost, jakožto výstup predikce, pak může příslušné odpovědné osobě (podnikovému manažerovi) výrazně napomoci při plánování marketingových strategií, například tzv. křížového prodeje (snahy, jejichž účelem je navýšit objednávku zákazníka doporučením jiných produktů nabízených společností) a následného prodeje (aktivity, jejichž cílem je nabídnout zákazníkovi vyšší/pokročilejší a tedy i dražší model/verzi produktu).

Příspěvek popisuje celý proces zpracování dostupných dat: od jejich čištění pomocí filtrů a různých funkcí nástroje MS Excel, přes jejich přípravu pro dolovací úlohu, po vlastní zpracování pomocí nástroje SAS Enterprise Miner. Pro dolování znalostí bylo použito regresní analýzy, neuronové sítě a rozhodovacího stromu, jejichž principy jsou v článku též stručně vysvětleny. Odhad chování zákazníka byl testován na dvou dolovacích úlohách lišících se v použití atributů a v počtu kategorií jednoho z prediktivních atributů. Výsledky těchto dvou úloh jsou konfrontovány pomocí grafů úspěšnosti predikce.

získávání znalostí z databází, datamining, predikce, zákazník, rozhodování, řízení

SUMMARY

If a company wants to compete in today's competitive market environment, it is necessary to monitor the behaviour of its customers. Business managers accounting for commercial success or non-success of organisation, therefore these managers have to gain knowledge needful for correct decision acceptance. These knowledge represent sophisticated information hidden in data that are at disposal for enterprise. It is mentioned that a data volume in enterprise will double every five years on average, this means that the problem of the present time is not to obtain data, but to cultivate it and to avail its potential. One possibility, how to extract mentioned knowledge from data, is to use so-called datamining assets.

The paper deals with an application of chosen basic methods of knowledge discovering in databases for area of customer-provider relation and it presents, how to avail acquired knowledge with respect to reference to solving problem – as basis of managerial decisions leading to improving of customer relationship management. In the concrete it solves prediction, whose aim is, on the basis of some attributes of exploring objects, to predict future behaviour of objects with these attributes. This way acquired knowledge, as the output of prediction, then can markedly help competent responsible person (enterprise manager) with planning of marketing strategies, for example so-called cross-selling (tendencies, whose aim is to increase the customer order by recommendation of other products offered by the company) and up-selling (activities, whose aim is to offer the customer superior/more advanced and thus more expensive product model/version).

The contributon describes a whole operation of available data processing: from its purifying by the help of filters and various functions of MS Excel tool, over its preparation for mining task, to self processing by the help of SAS Enterprise Miner tool. Regression analysis, neural network and decision tree, whose principles are briefly explained in this paper too, were used for knowledge mining. The estimation of customer behaviour was tested by two mining task varying in attribute using and in cathegories number of one of predicive attributes. The results of these two tasks are confronted by the help of prediction fruitfulness charts.

Článek vznikl za podpory výzkumného záměru Provozně ekonomické fakulty Mendelovy zemědělské a lesnické univerzity v Brně MSM 6215648904/03/03/02 a projektu IG 180601/2102/116 s názvem Analýza a návrh využitelnosti prostředků dataminingu při monitorování interakcí subjektů účastnících se procesu obchodování.

LITERATURA

BENJAMINI, Y., LESHNO, M., 2005: Statistical Methods For Data Mning. In: Maimon, O., Rokach, L. ed. *The Data Mining and Knowledge Discovery Handbook*. 1. vyd. New York: Springer, 565–587. ISBN 0-387-24435-2.

BERKA, P., 2003: *Dobývání znalostí z databází*. 1. vyd. Praha: Academia, 368. s. ISBN 80-200-1062-9.

CLEMENTE, M. N., 2004: *Slovník marketingu*. 1. vyd. Brno: Computer Press, 378 s. ISBN 80-251-0228-9.

DOSTÁL, P., RAIS, K., SOJKA, Z., 2005: *Pokročilé metody manažerského rozhodování*. 1. vyd. Praha: Grada Publishing, 168 s. ISBN 80-247-1338-1.

HAN, J., KAMBER, M., 2006: Data Mining Concepts and Techniques. 2. vyd. San Francisco: Morgan Kaufmann, 800 s. ISBN 1-55860-901-6.

MELOUN, M., MILITKÝ, J., 2006: *Kompendium statistického zpracování dat*. 2. vyd. Praha: Academia, 984 s. ISBN 80-200-1396-2.

NOVOTNÝ, O., POUR, J., SLÁNSKÝ, D., 2005: *Business Intelligence: Jak využít bohatství ve vašich datech*. 1. vyd. Praha: Grada Publishing, 256 s. ISBN 80-247-1094-3.

PARR RUD, O., 2001: *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*. 1. vyd. New York: John Wiley & Sons, 367 s. ISBN 0-471-38564-6.

ROKACH, L., MAIMON, O., 2005: Decision Trees. In: Maimon, O., Rokach, L. ed. *The Data Mining and Knowledge Discovery Handbook*. 1. vyd. New York: Springer, 165–192. ISBN 0-387-24435-2.

SAS INSTITUTE INC., 2008: *Data mining with SAS® Enterprise Miner™* [online]. poslední aktualizace: 2008 [cit 24. 11. 2008]. URL http://www.sas.com/technologies/analytics/datamining/miner/index.html.

ZHANG, P. G., 2005: Neural Networks. In: Maimon, O., Rokach, L. ed. *The Data Mining and Knowledge Discovery Handbook*. 1. vyd. New York: Springer, 487–516. ISBN 0-387-24435-2.

Adresa

Ing. Naděžda Chalupová, Ústav informatiky, Mendelova zemědělská a lesnická univerzita v Brně, Zemědělská 1, 613 00 Brno, Česká republika, e-mail: nadule@pef.mendelu.cz