Poznámky k hodnocení a údržbě marketingových predikčních a segmentačních modelů

Nikola Kaspříková, http://data.tulipany.cz

klíčová slova: data mining, marketing, hodnocení kvality modelu, predikce, segmentace

Abstrakt: V textu jsou diskutovány otázky a přístupy spojené s problematikou hodnocení a údržby marketingových data mining modelů a to jak modelů predikčních, využívaných například pro cílení kampaní nebo předpověď odchodu zákazníka ke konkurenci, tak modelů segmentačních. Vývoj a aplikace data mining modelů uvedených typů již patří k postupům mnohými společnostmi využívaným pro účely zvýšení efektivity marketingového řízení. Hodnocení modelů a jejich údržba patří k podstatným součástem analytického procesu, kterým je třeba věnovat dostatečnou pozornost.

Úvod

Data miningové modely patří k prostředkům, které mohou zajistit podstatný přínos pro účely efektivního řízení vztahů se zákazníky. Toto se projevuje zejména ve společnostech disponujících rozsáhlejší datovou základnou o klientech a historii jejich chování ve vztahu k dané společnosti, jejím produktům a službám. V dnešní době se již v mnohých firmách pro cílení marketingových kampaní běžně využívají predikční modely založené na odhadu sklonu zákazníka ke koupi určitého produktu tzv. propensity-to-buy modely, podle kterých je vybrána optimální cílová skupina klientů, kterým bude nabídnut určitý produkt. Takové modely mají povahu predikčních modelů, jde o zjednodušený obraz reality, který má formu například regresní rovnice nebo rozhodovacího stromu a který má předpovědět, kteří klienti pozitivně zareagují na nabídku daného produktu. Další důležitou aplikační oblastí pro predikční modely je churn management, kde se pracuje s modely předpovídajícími, kteří zákazníci odejdou ke konkurenci nebo omezí objem využívaných produktů a služeb dané společnosti.

Svou úlohu při podpoře marketingového řízení mají vedle predikčních modelů také modely segmentační [http://data.tulipany.cz/segmentaceCRM.pdf], které jsou vytvářeny za účelem nalezení nejvhodnějších pravidel, podle kterých je klientská základna rozčleněna do vzájemně odlišných a vnitřně homogenních segmentů.

Analytický proces vytěžování dat (data mining) obvykle představuje vykonání určité posloupnosti činností, přičemž obsah jednotlivých fází je vždy dán povahou řešeného problému. Postup při vytěžování dat lze zjednodušeně shrnout do několika fází:

(1) zadání úlohy, porozumění problému

(2) příprava dat

(3) modelování – vlastní vytváření (učení) modelu z dat

(4) validace, hodnocení modelu

(5) využití výsledků v praxi

(6) údržba modelu

Celý proces začíná vymezením problému, který je potřeba řešit. Podle toho se určí zdroje dat potřebných pro analýzu a zvolí se typ modelu, který bude použit. V průběhu dalšího postupu se může ukázat, že vybraný postup nevede k uspokojivému závěru a je nutné buď znovu provést některé kroky jinak nebo projekt zastavit. Také se může stát, že při řešení úlohy se neočekávaně projeví nějaké překvapivé skutečnosti, které dají podnět k dalším analýzám. Přehled o data miningu a možnostech využití data mining postupů pro marketing poskytuje například dokument [http://data.tulipany.cz/dataminingCRM.php]. Dále se v textu budeme věnovat problematice hodnocení modelu. Hodnocení vhodnosti modelu je třeba provést po dokončení vývoje modelu při rozhodování o volbě modelu, který má být použit. Po nasazení modelu do provozu se potom průběžné hodnocení modelu provádí v rámci údržby modelu a je východiskem pro rozhodování, zda vyvinutý model ještě stále dobře vyhovuje požadavkům, pro které byl vytvořen, a může být používán nadále, nebo by již měl být nahrazen modelem novým.

Je třeba připomenout, že volba model hodnoticích kritérií by měla vždy vycházet z požadavků, které jsou určeny povahou konkrétního řešeného věcného problému. Existují však i obecné principy hodnocení modelů, které je vhodné uvažovat. Statistické softwarové nástroje obsahují dostatek prostředků k diagnostice modelů, je užitečné rozumět principům, na kterých jsou tyto nástroje založeny, znát souvislosti jejich aplikace a předpoklady jejich korektního použití.

1 Obecné poznámky k problematice hodnocení modelů

Pro predikční i segmentační modely, které mají být dobře použitelné v praxi, platí, že pro účely ověření správnosti modelu (a použitelnosti na další data) je potřeba prozkoumat, zda model umožňuje správně vystihnout obecněji platné vztahy, než jaké se projevily zrovna v datech použitých přímo k vytvoření modelu (*trénovací data*). Proto se obvykle přímo k vytváření modelu (tj. odhadu struktury a parametrů) nepoužijí všechna dostupná data a část dat (*testovací data*) je ponechána pro testování přesnosti modelu. Když je potom pozorováno podstatné snížení výkonnosti modelu při aplikaci na testovací data oproti výsledkům na trénovacích datech, tedy se projeví jev nazývaný *přeučení modelu* (overfitting, model je příliš přizpůsoben trénovacím datům), je rozumným postupem ustoupit od modelu s vyšší zjištěnou přesností na trénovacích datech a přijmout raději jednodušší a na trénovacích datech třeba méně přesný model, který je ale nakonec pro praktické aplikace vyhovující spíše.

Při řešení otázky vhodného způsobu definování hodnoticí funkce je třeba si uvědomit, že přesnější model (tj. při volbě přístupu maximalizace věrohodnostní funkce takový model M, u kterého je poměrně vysoká pravděpodobnost, že by mohl generovat analyzovaná data D, jde o maximalizaci výrazu $P(D|M)$ výběrem vhodného M) bývá také složitější z hlediska struktury či počtu parametrů a tedy pro praktické využití z některých hledisek méně užitečný. Proto běžně používané vztahy pro hodnocení modelů[1] obsahují nejen složky kladně ohodnocující dobrý soulad modelu s analyzovanými daty, ale i členy penalizující složitost modelu. To je vyjádřeno *principem MDL, Minimum Description Length)* [Lam-Bacchus], který je motivován úvahou, že absolutně přesný (a složitý) popis dat[2] není vždy žádoucí, zejména v případě, kdy data jsou zatížena chybami a výsledky modelování (zjištěné vztahy) by měly být dostatečně obecné na to, aby byly platné i pro jiná než k vytvoření modelu použitá data. Proto je snaha minimalizovat (při určitém zvoleném způsobu kódování) celkovou délku popisu dat s využitím modelu, totiž součet délky popisu modelu a uchovávaných od-

[1] jako například BIC (bayesovské informační kritérium) nebo AIC; pro velice přístupnou diskuzi problematiky hodnocení modelů a definici informačních kritérií viz například [Pekár]

[2] který je dostupný vždy (vlastně jsou jím i data samotná)

chylek hodnot skutečných pozorování od hodnot modelem předpovídaných. Přitom zřejmě je mezi délkami obou složek vztah negativní substituce (tradeoff), vyjadřující nutnost volby mezi přesností a jednoduchostí (resp. možnostmi interpretace modelu a jeho praktickou využitelností). Pokud se pracuje s přesným (a složitějším) modelem, často mohou být chyby menší, naopak při použití jednoduššího (a méně přesného) modelu je potřeba počítat s většími odchylkami předpovídaných a skutečných hodnot.

2 Segmentační modely

2.1 Testování přesnosti segmentace

Hodnocení přesnosti modelu pro segmentaci spočívá v posouzení vnitroshlukové variability a v splňování požadavku na dobrou rozlišitelnost mezi jednotlivými segmenty. Přitom je požadováno, aby si zákazníci ze stejného segmentu byli hodně podobní (vnitroshluková variabilita nízká), což přispívá použitelnosti segmentace ať už pro přímé využití k řízení vztahů se zákazníky (kdy je užitečné, jestliže charakteristiky všech zákazníků do daného segmentu zařazených mohou být poměrně přesně vystiženy popisem vlastností typického, resp. průměrného představitele daného segmentu) nebo pro účel zpřesnění predikčních modelů. U modelů pracujících se vzdálenostmi (případ klasické shlukové analýzy) se sledují vzdálenosti vlastností jednotlivých klientů v segmentu například od vlastností reprezentanta či středu segmentu[1], u modelů pracujících s pravděpodobnostmi se sledují pravděpodobnosti příslušnosti klienta do jednotlivých segmentů, obdobně například při fuzzy shlukové analýze [http://data.tulipany.cz/ClustR.pdf] se pracuje s koeficienty příslušnosti. Pokud nastane případ, že u významné části klientů jsou u jednotlivých klientů rozdíly vzdáleností od středů několika nejbližších segmentů poměrně malé, případně několik nejvyšších pravděpodobností náležení do segmentů se příliš neliší (a jsou poměrně nízké), není přiřazení klientů do segmentů dostatečně jednoznačné a naznačuje podobnost segmentů a případně jejich částečný překryv.

Testování přesnosti segmentačního modelu je představováno získáním pravidel zařazování klientů do jednotlivých segmentů a jejich aplikací na nové (přesněji řečeno do segmentů dosud nezařazené) zákazníky (testovací data), přičemž se sledují parametry uvedené výše (variabilita uvnitř jednotlivých segmentů, jednoznačnost přiřazení zákazníků do segmentů) a srovnávají se s hodnotami dosaženými na trénovacích datech.

2.2 Údržba segmentačního modelu a časové aspekty segmentace

Obvyklým požadavkem na marketingový segmentační model je jeho dobrá srozumitelnost, totiž jasný význam segmentačních charakteristik a srozumitelné vyjádření vlastností jednotlivých segmentů. Vedle toho jsou často podstatné také technické vlastnosti modelu spočívající v dostatečně jednoznačné, přiřazení zákazníků do segmentů a dobrou stabilitou segmentů. Analogicky ověřování přesnosti modelu na testovacích datech se u modelů určených pro dlouhodobější používání sleduje, zda časem nedojde ke zhoršení sledovaných parametrů při používání modelu na nových datech a řeší se otázka načasování vytvoření nového modelu s použitím nově dostupných dat. Připomeňme, že u segmentace pro účely marketingového řízení je nežádoucí zbytečně častá resegmentace (ve smyslu znovuobjevování struktury dat a určování počtu a charakteristik segmentů a obecných pravidel pro náležení k nim; nikoliv opakovaného přiřazení klientů, dříve již do některého segmentu zařazených, do (možná jiného) segmentu s využitím aktuálních dat, což je krok méně zásadní) a je ceněná určitá stálost vymezení segmentů, která usnadňuje jejich řízení. Nicméně i přesto je potřebné resegmentaci provést v situaci, kdy noví klienti přestanou snadno zapadávat do některého ze stávajících segmentů (klesá míra jednoznačnosti přiřazení klientů do segmentů), je po-

[1] přesněji řečeno se pracuje s vektory příznaků

zorováno zvýšení variability v rámci jednotlivých segmentů (popisy segmentů přestávají mít dobrou vypovídací schopnost) nebo (v tomto bodě podobně jako u predikčních modelů) dojde ke změně struktury dat – jsou měřeny jiné charakteristiky, než podle kterých byly segmenty vytvářeny: například při změnách podmínek poskytování jednotlivých produktů, zavádění nových služeb a podobně.

Při dlouhodobějším charakteru využívání segmentace je užitečné sledovat přechody zákaz-níků mezi segmenty v čase. Zvyklosti, potřeby a chování některých zákazníků nemusejí zůstávat neměnné a mohou procházet určitým vývojem. Charakteristiky takových klientů pak již nemusejí odpovídat segmentu, do kterého byli přiřazeni původně, mohou však začít vyhovovat vlastnostem jiného segmentu; pak je postačující pracovat se stávajícími segmenty (a není nutné provádět kompletní resegmentaci jako v případech naznačených výše) a provést znovu pouze přiřazení klientů segmentům podle obecných pravidel nalezených dříve. Otázka frekvence znovupřiřazování klientů k segmentům se řeší v závislosti na povaze charakteristik pro segmentaci používaných, dostupnosti dat, potřebnosti znalosti aktuálního přiřazení klientů do segmentů a očekávané proměnlivosti charakteristik klientů v čase. Přitom může být vhodné pracovat také se speciálním technickým segmentem nových klientů, kam jsou apriori (tj. nezávisle na výsledcích modelování) zařazováni nově získaní klienti, u kterých lze po určitou dobu zpočátku klientského vztahu očekávat specifické chování. Představu o některých časových aspektech segmentace je možné získat z matice přechodů mezi segmenty, jejíž prvek v i-tém řádku a j-tém sloupci udává počet klientů, kteří byli v určitém období zařazeni do segmentu i a v období následujícím byli přiřazeni segmentu j[1]. Z matice přechodů mezi segmenty lze snadno odvodit matici pravděpodobností přechodů mezi segmenty, která je pro některé účely více informativní a přehlednější. Příklad takové matice je uveden níže:

Segment	*Segment 1*	*Segment 2*	*Segment 3*	*Segment 4*
Segment 1	***0,93***	0,05	0,01	0,01
Segment 2	0,03	***0,91***	0,02	0,04
Segment 3	0,01	0,02	***0,95***	0,02
Segment 4	0,03	0,03	0,02	***0,92***
Noví klienti	0,05	0,10	0,64	0,21

Z uvedené matice pravděpodobností přechodů mezi segmenty je patrné, že v našem zjednodušeném příkladě se čtyřmi segmenty lze pozorovat poměrně velkou stabilitu kteréhokoliv ze segmentů 1 až 4: více než 90% klientů zůstává ve stejném segmentu i v příštím období. Je na zvážení, zda neprodloužit interval mezi opakovaným zjišťováním příslušnosti do segmentů. Noví klienti přecházejí nejčastěji do *segmentu 3*, často také do *segmentu 4*, nejméně do *segmentu 1*. Poměrně málo časté jsou přímé fluktuace mezi *segmentem 1* a *segmentem 3* a to v kterémkoli směru.

Někdy může být zajímavé jít v analýze segmentačního modelu ještě o krok dále a pracovat s několika historickými maticemi pravděpodobností přechodů mezi segmenty za dvojice po sobě jdoucích období a zkoumat, jak proměnlivé jsou pravděpodobnosti přechodů v čase nebo zda pravděpodobnost, že klient v následujícím období přejde do určitého segmentu

[1] v této souvislosti je vhodné mít na paměti, že přiřazování do segmentů podle transakčních charakteristik se děje na základě historických dat (data za aktuální, resp. následující období ještě nejsou dostupná a tak vlastně nikdy nemůžeme okamžitě disponovat přesnou znalostí segmentu klienta pro právě nadcházející období), tj. klienti jsou zařazeni do segmentu zpětně podle projevů v průběhu uplynulého období (např. měsíce nebo kvartálu)

(resp. setrvá ve stávajícím), bude závislá pouze na tom, do kterého segmentu právě patří, a nebude podstatné, přes které segmenty se do něj dostal.

3 Validace predikčních modelů

Hodnocení přesnosti modelu v případě predikčního modelování znamená zjištění a ohodnocení míry souladu hodnot modelem předpovídaných a skutečně pozorovaných hodnot zkoumané veličiny. Přitom pro ověření správnosti se sleduje především přesnost modelu na datech, která nebyla použita k jeho vytváření, totiž na *testovacích datech*.

Připomeňme, že pokud je předpovídaná proměnná spojitá, často se chyba předpovědi L definuje jako součet čtverců odchylek jednotlivých předpovědí $\hat{y}_i$ od skutečných hodnot y_i:

$$L = \sum_i (\hat{y}_i - y_i)^2,$$

ale samozřejmě je možné podle potřeby definovat chybu i jinak. Pokud je zkoumaná veličina kategoriální a v nejjednodušším a obvyklém případě dvouhodnotová (zobecnění níže uvedených přístupů na případ vícehodnotové proměnné je poměrně přímočaré; pro zjednodušení interpretace předpokládejme, že u zkoumané proměnné značení p znamená kategorii nazvanou pozitivní a hodnota n negativní), bývá rozumné pracovat s hodnotící funkcí ve tvaru

$$L = w_1 c_1 P(\tilde{n}|p) + w_2 c_2 P(\tilde{p}|n),$$

kde $P(\tilde{n}|p)$ je pravděpodobnost, že pozorování p bude modelem chybně klasifikováno jako n (chyba 1. druhu), podobně $P(\tilde{p}|n)$ vyjadřuje pravděpodobnost, že pozorování, které je ve skutečnosti n bude chybně klasifikováno jako p (chyba 2. druhu); c_i vyjadřuje cenu chybného rozhodnutí i-tého druhu, w_1 je očekávaná relativní četnost výskytu pozorování p v datovém souboru, na který je zamýšleno vytvářený model použít, podobně w_2 pro n. Členy $P(\tilde{n}|p)$ a $P(\tilde{p}|n)$ vyjadřují "technickou" správnost predikce, členy w_i zajišťují správnou kalibraci modelu a c_i zohledňuje obecně různé náklady chybných rozhodnutí, jejichž přesná kvantifikace je v některých případech obtížná, ale stojí za to pokusit se je nějak odhadnout. Chyba při přijetí žádosti o úvěr u klienta, u kterého dojde k defaultu, bude spíše závažnější než neposkytnutí úvěru klientovi, který by splácel bez problémů. Podobně je vhodné například při predikci churnu (odchodu zákazníka od firmy) pozorněji sledovat přesnost předpovědí u bonitních klientů, vzhledem ke tvaru výše uvedené hodnoticí funkce se jeví jako efektivní v rámci přípravy dat v datovém souboru použitém pro analýzu zvýšit váhu (relativní četnost výskytu) klientů s vysokými hodnotami proměnných vázaných na hodnotu klienta.

Základní představu o technické přesnosti modelu lze získat z *matice záměn*, jejímiž prvky jsou četnosti jednotlivých kombinací skutečné hodnoty zkoumané proměnné a hodnoty předpovídané modelem. Struktura matice záměn je následující:

$$\begin{array}{c} \\ \tilde{p} \\ \tilde{n} \end{array} \begin{array}{c} \begin{array}{cc} p & n \end{array} \\ \begin{pmatrix} Tp & Fp \\ Fn & Tn \end{pmatrix} \end{array}$$

přitom prvky na diagonále představují počty případů klasifikovaných správně jako pozitivní (Tp), resp. negativní (Tn); prvky mimo diagonálu představují počet případů chybně klasifikovaných jako pozitivní (Fp) resp. jako negativní (Fn). Absolutně přesně klasifikující model má diagonální matici záměn, jako je například:

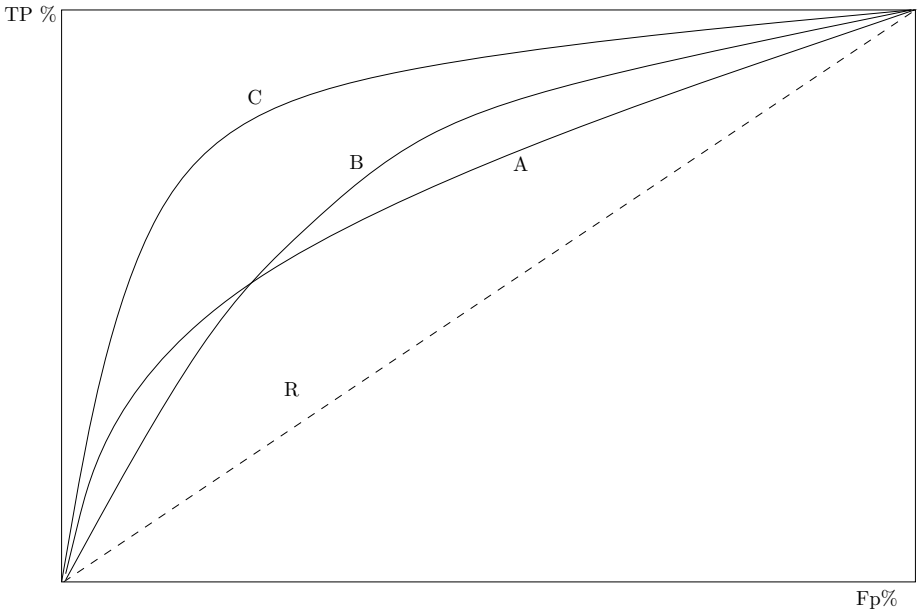
$$\begin{array}{c} \\ \tilde{p} \\ \tilde{n} \end{array} \begin{array}{c} \begin{array}{cc} p & n \end{array} \\ \begin{pmatrix} 40 & 0 \\ 0 & 960 \end{pmatrix} \end{array}$$

všimněme si, že takový absolutně přesný model přináší pouze nepatrné zlepšení přesnosti proti zcela jednoduchému modelu, který nezávisle na zjištěných hodnotách prediktorů zařazuje všechny případy do majoritní třídy negativní a jehož matice záměn na stejných datech vypadá takto:

$$\begin{array}{c} \\ \tilde{p} \\ \tilde{n} \end{array} \begin{array}{c} \begin{array}{cc} p & n \end{array} \\ \begin{pmatrix} 0 & 0 \\ 40 & 960 \end{pmatrix} \end{array}$$

a dosahuje tak 96% správnosti klasifikace. V závislosti na cenách chybných rozhodnutí se ale (například v již zmíněném případě predikce defaultu) může při použití byť o málo technicky přesnějšího modelu jednat o významné zlepšení.

Určitým problémem je z pohledu možností používat matici záměn hodnocení přesnosti a rozhodování o způsobu využívání u modelů některých typů, jejichž výstupem při aplikaci na data, u kterých je potřeba rozhodnout o hodnotě zkoumané proměnné, je výpočet pravděpodobností jednotlivých hodnot proměnné, případně výpočet skóre. Potom je pro praktické aplikace potřeba určit prahovou hodnotu. Výše uvedené matice záměn umožňují (po případné úpravě výsledků podle hodnoticí funkce L) snadno srovnat přesnost dvou modelů při určité zvolené prahové hodnotě; pro přehledné srovnání výkonnosti modelů při různě volených prahových hodnotách je ale vhodné použít operační charakteristiku, jakou je ROC křivka [Stein, M.: Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation. Moody's KMV Technical Report], která vyjadřuje vztah mezi Fp v % a Tp v %. Postup konstrukce křivky je takový, že se nejprve sestupně seřadí případy podle modelem určené pravděpodobnosti (resp. skóre) výskytu hodnoty pozitivní a potom postupně pro jednotlivé prahy se nanáší na osu x procento případů nesprávně klasifikovaných jako pozitivní a na osu y procento případů správně předpovězených jako pozitivní. Ukázka ROC křivek několika alternativních modelů je na obrázku, kde křivka R odpovídá neinformovanému modelu (náhodné rozhodování), křivka C modelu realizujícímu rozhodovací funkci, která dominuje rozhodovací funkci realizované kterýmkoli z modelů vystižených křivkami A, B, R. O vztahu přesnosti modelů vystižených křivkami A, B se nelze vyjádřit jednoznačně (jsou však přesnější než náhodně rozhodující model), je například možné oba modely srovnat alespoň podle plochy pod křivkou. Pro některé účely může být užitečnější model popsaný křivkou A spíše než B - například pro cílení kampaní s malým požadovaným rozsahem cílové skupiny, jindy může být vhodnější druhý model.



Dalším často používaným prostředkem pro hodnocení přesnosti, resp. srovnání výkonnosti modelů v marketingu je *křivka navýšení* (lift chart), často používaná u modelů předpovídajících responzi klientů na marketingovou kampaň, kde se sleduje, jak se ve srovnání s případem náhodného výběru cílové skupiny vyvíjí *response rate* (relativní četnost klientů, kteří na kampaň zareagovali, v souboru všech oslovených klientů) při postupném oslovování stále většího počtu klientů, kdy přednostně jsou oslovováni klienti s (podle modelu) větší pravděpodobností responze.

U predikčních modelů je nutné podobně jako u modelů segmentačních provádět údržbu modelu, pokud je záměr model využívat opakovaně, resp. dlouhodobě. Bývá užitečné po každém použití modelu pokusit se zkonstruovat a vyhodnotit například empirickou ROC křivku a pokud se podstatně zhoršila, je vhodné pokusit se vytvořit nový model. Při vyhodnocování efektivity predikčního modelu například s pomocí sestrojení empirické ROC křivky je důležité zahrnout do analýzy všechny podstatné faktory, aby měření výkonnosti modelu bylo korektní. Například pokud se hodnotí kvalita modelu pro predikci odchodu zákazníka ke konkurenci a tento model byl použit na výběr klientů, kteří by měli být se strany společnosti během stanoveného období kontaktováni s nabídkou výhodnějších podmínek, která je měla pomoci u společnosti udržet, je nutné analyzovat vývoj poměru zákazníků, kteří od společnosti po skončení kampaně odešli, v závislosti na modelem předpovídané pravděpodobnosti odchodu zvlášť ve skupině zákazníků, kteří byli během stanoveného období skutečně kontaktováni, a zvlášť ve skupině zákazníků, u nichž kontakt neproběhl.

4 Závěr

Hodnocení a údržba predikčních a segmentačních modelů patří k podstatným součástem analytických úloh v marketingové praxi. Při vytváření a výběru modelů pro marketing je obvykle smysluplné dávat přednost robustním modelům, které sice nemusejí vykazovat nejlepší výsledky nad daty použitými pro odhady parametrů modelu, ale jsou díky své jednodušší struktuře snadno interpretovatelné a nedochází k podstatnému zhoršení jejich kvality při použití na nová data. Pro hodnocení predikčních i segmentačních modelů existují poměrně dobře dostupné a propracované techniky a prostředky, které je vhodné použít i pro účel sledování vývoje vlastností používaného modelu v čase.

Reference

LAM, W. - BACCHUS, F. Learning Bayesian Belief Networks, An approach based on the MDL Principle. Computational Intelligence 10:4. Blackwell Publishing, 1994. ISSN 0824-7935.

PEKÁR, S. - BRABEC, M. Moderní analýza biologických dat. Zobecněné lineární modely v prostředí R. Scientia 2009. ISBN 978-80-86960-44-9.