Študentská vedecká aktivita

Použitie podmienených stromových štruktúr na predikciu úspešnosti bankovej marketingovej kampane

Filip Sochor

Sekcia: Kvantitatívne metódy a informatika

Konzultant: Mgr. Mária Stachová, PhD.

Banská Bystrica

2016

ABSTRAKT

SOCHOR, Filip Bc.: Použitie podmienených stromových štruktúr na predikciu úspešnosti bankovej marketingovej kampane. [Študentská vedecká aktivita] / Filip Sochor. - Univerzita Mateja Bela v Banskej Bystrici. Ekonomická fakulta; Katedra kvantitatívnych metód a informačných systémov. - Vedúci: Mgr. Mária Stachová, PhD.

Marketingová kampaň je v bankovom sektore jedným z kľúčových spôsobov ako si získať, prípadne udržať klienta. Úspešnosť kampane vždy závisí od explicitných aj implicitných faktorov. Táto práca sa zaoberá predikciou implicitných faktorov na úspešnosť bankovej marketingovej kampane, využívaním podmienených stromových štruktúr, ktoré sú súčasťou data miningových metód. Analyzovanými dátami sú verejne dostupné údaje portugalskej bankovej inštitúcie o klientoch portugalských bánk súvisiace s marketingovou kampaňou. Na všetky štatistické výpočty je využívané prostredie systému R.

Kľúčové slová: data mining. marketingová kampaň. štatistický program R.

OBSAH

ÚVOD

Bankové inštitúcie sa definujú ako subjekty, ktoré prijímajú vklady a poskytujú úvery. Táto definícia je všeobecným vyjadrením toho, čo je primárnou podnikateľskou náplňou banky. Členenie bankových produktov, je však v dnešnej dobe značne diverzifikované, takým spôsobom, aby mal každý klient možnosť vybrať si „produkt na mieru“ a zároveň je to aj spôsob akým sa banky navzájom odlišujú. Bankovníctvo sa stalo veľmi ziskovou činnosťou a preto aj pre subjekty pôsobiace v tomto sektore je získať si a udržať si klienta čoraz náročnejšie. Konkurencia stúpa či už z kvalitatívneho hľadiska alebo kvantitatívneho.

Banky preto využívajú rôzne typy analýz a výskumov, aby zistili ako ich marketingové kampane na získanie si, respektíve udržanie si klienta fungujú, aké sú faktory ich úspechu, prípadne aká je ich úspešnosť vôbec. Výstupy takýchto štúdií potom vedú k priamemu/selektívnemu marketingu, pri ktorom je kontaktovaná iba určitá cieľová skupina klientov, pre ktorú sa špeciálne nastavia parametre produktu, ktorý je ponúkaný.

To, ako jednotliví klienti reagujú na takúto marketingovú ponuku si banky uchovávajú a pri získaní dostatočného počtu údajov potom aj klientske správanie vyhodnocujú. Slúžia na to viaceré metódy, najčastejšie využívanými sú matematicko-štatistické metódy, ktoré sú schopné odhaliť implicitné vzťahy medzi rozhodnutím klienta a tým čo jeho rozhodnutie mohlo ovplyvniť. Je preto kľúčovým krokom, vedieť získané dáta rozhodnutí klientov správne analyzovať a vyhodnotiť. Práve tento faktor môže byť podstatným pri tvorbe produktov, ich nastavovaní, komunikácii aj distribúcii a práve dobrá analýza môže viesť k vyššej konkurencieschopnosti banky na konkurenčnom trhu.

V tejto práci budeme analyzovať dátovú množinu portugalskej bankovej inštitúcie obsahujúcej informácie o bankovej marketingovej kampani, pričom budeme využívať štatistické metódy podmienených stromových štruktúr zo štatistického smeru nazývaného data mining.

Našim cieľom bude zistiť, v akej miere sú metódy, ktoré budeme využívať aplikovateľné na tento typ praxe, ich schopnosťou štatistickej predikcie a tiež, ktoré faktory vplývajú na rozhodnutie sa klienta, reagovať na marketingovú kampaň pozitívne.

1 TEORETICKÁ ČASŤ

Práca bude analyzovať cielenú marketingovú kampaň v bankovom sektore a snažiť sa zhotoviť model predikcie jej úspešnosti. Na zhotovenie budú využité modely podmienených stromových štruktúr. V prvej časti práce opíšeme analyzované údaje marketingovej kampane a tiež teoreticky vymedzíme štatistické metódy, ktoré budeme aplikovať. V druhej časti práce budeme konkrétne analyzovať spomenutú dátovú množinu a popisovať výsledky predikčnej analýzy.

1.1 Opis dátovej množiny

Dátová množina, ktorú budeme využívať pre účely práce, obsahuje údaje portugalskej bankovej inštitúcie, ktoré poskytujú informácie o bankovej marketingovej kampani za roky 2008-2010. Vzorka s ktorou pracujeme zahŕňa 45 211 klientov, o ktorých je uverejnených 17 údajov (premenných). Celá databáza je voľne prístupná na webovom portáli: „UCI Machine Learning Repository“ pod názvom: „Bank Marketing Data Set“. V marketingovej kampani bol ako produkt, ponúkaný termínovaný vklad s vysokou úrokovou sadzbou a podiel úspešne oslovených klientov bol 13,2%. V nasledujúcej tabuľke sú uvedené využívané premenné s popisom charakteru.

Tabuľka 1: Premenné analyzovanej dátovej množiny

P.č.	Označenie premennej	Názov premennej	Charakter
1	O1	Vek	Numerická
2	O2	Povolanie	Kategorická
3	O3	Rodinný stav	Kategorická
4	O4	Vzdelanie	Kategorická
5	O5	Default úroku v histórii	Binárna
6	O6	Ročný priemerný stav prostriedkov na účte	Numerická
7	O7	Vlastní úver na bývanie	Binárna
8	O8	Vlastní úver	Binárna
9	K1	Typ kontaktovania klienta	Kategorická
10	K2	Deň mesiaca posledného kontaktu s klientom	Kategorická
11	K3	Mesiac posledného kontaktu s klientom	Kategorická
12	K4	Dĺžka posledného hovoru s klientom v sekundách	Numerická
13	I1	Počet kontaktov s klientom	Numerická
14	I2	Počet uplynutých dní od kedy bol klient kontaktovaný v predošlej kampani	Numerická
15	I3	Počet kontaktov s klientom v predošlých kampaniach	Numerická
16	I4	Výstup z predošlých marketingových kampaní	Kategorická
17	Y	Prijatie poskytovaného produktu	Binárna

Prameň: Vlastné spracovanie

1.2 Data mining

Čím ďalej, tým viac je veda a technika v rozmachu. S možnosťou vývinu veľkokapacitných zariadení na úschovu dát, sa možnosti štatistických metód vo veľkej miere výrazne zväčšili. Tvorba nových modelov, metód, priam aj odborov ktoré sa venujú práve pozorovaniu dát a hľadaniu vzťahov medzi nimi rastie rýchlym tempom. Takéto prístupy k hľadaniu skrytých informácii si našli svoje uplatnenie v mnohých odboroch od tých vedeckých až po výhradne komerčné. Data mining je jeden z prístupov, ktorý sa týmto smerom vyvíja a v tejto práci budeme využívať niektoré z jeho poznatkov práve na predikciu správania sa klientov voči ponúknutým produktom.

Frawley a kol. (1992) definujú data mining ako netriviálny proces extrakcie nepriamych, doteraz nepoznaných a potenciálne použiteľných informácií z dát. Larose (2005) tento pojem definuje ako proces odhaľovania nových zmysluplných korelácií, modelov a trendov preskúmavaním veľkých objemov dát uložených v dátových skladoch, používaním technológií rozoznávajúcich modely, tak ako aj štatistickými a matematickými metódami.

Data mining ako súhrn metód poskytuje veľmi veľký počet rôznych prístupov nahliadania na dáta. V tejto práci sa budeme zameriavať na hľadanie súvislostí pomocou podmienených stromových štruktúr. Podmienené stromové štruktúry patria do kategórie rozhodovacích stromov, kde súvislosť závislej premennej vysvetľuje jedna alebo viacero nezávislých premenných. Keďže rozhodovacie stromy sú základom pre metódu podmienených stromových štruktúr, ich popis začneme práve popisom rozhodovacích stromov.

1.3 Rozhodovacie stromy

Rozhodovací strom je triediaci nástroj (klasifikátor) s funkciou rekurzívneho delenia objektov dátových množín (Dahan, 2014). Rozhodovacie stromy sú vytvárané algoritmami, ktoré umožňujú rozdelenie dátových objektov do vetiev. Na základe kritéria v uzle prebehne rozdelenie do segmentov začínajúc prvým delením, ktoré nastáva v takzvanom počiatočnom uzle a tvorí vrchol stromu. Každá vetva je potom delená cez uzly (obsahujúce kritériá) na ďalšie vetvy až pokým nedôjde ku kritériu zastavenia algoritmu, ktoré je užívateľom zadané. Košatosť stromu teda určuje užívateľ (de Ville, 2006).

Rozhodovacie stromy sú často preferovanými metódami pre výskum vďaka jednoduchosti ich využitia či pochopenia a taktiež vďaka tomu že na rozdiel od iných

klasifikačných metód (ako napríklad logistická regresia), predpoklady ich využitia neobmedzujú dátovú množinu.

1.4 Podmienené stromy

Podmienené stromy sú metódou založenou na rozhodovacích stromoch, konkrétne algoritme CART (classification and regression trees), ktoré vylepšením techniky rozhodovacích stromov riešia práve problémy tohto algoritmu. Väčšina autorov hovorí o týchto dvoch nedostatkoch, ktoré prináša algoritmus CART:

1. strom vytvorený týmto algoritmom sa stane príliš košatým a vie správne predikovať len dáta na ktorých bol vytvorený (takzvaný overfitting)
2. algoritmus zvýhodňuje premenné, pri ktorých je možné aplikovať viac delení alebo tých ktoré obsahujú veľký počet chýbajúcich hodnôt

Prvý z týchto problémov je pomocou orezávania stromu možné vyriešiť po vytvorení stromu, je to však spojené s ručnou optimalizáciou. Druhý problém tohto algoritmu nie je možné ručne vyriešiť, lebo sa týka samotného tvorenia stromu.

Spomenuté nedostatky boli predmetom mnohých článkov a nových štatistických metód. V tejto práci sa budeme inšpirovať Hothornom, Hornikom a Zeileisom (2006), ktorý zaviedli metódu podmieneného binárneho delenia.

Postup tvorenia stromu

1) V prvom kroku sa testuje globálna nulová hypotéza nezávislosti medzi prediktorom a závislou premennou (nezávislých premenných môže byť aj viac). Algoritmus sa zastaví pokiaľ hypotéza nemôže byť zamietnutá. Inak je vybraný ako kritérium delenia prediktor s najsilnejšou závislosťou, ktorá je meraná p-hodnotou a nasleduje ďalší krok. 2) Na vybraný prediktor sa použije deliace kritérium, ktoré stanoví hodnoty rozdelenia v konkrétnom uzle. 3) Rekurzívne sa opakujú kroky 1), 2) (Hothorn, Hornik, Zeileis, 2006). Hladiny p-hodnoty sú väčšinou nastavené v závislosti od dátovej množiny. Odporúčané hodnoty sú 0,05 alebo 0,01.

1.5 Štatistický program R

Výpočty použité pre analýzu v našej práci sa realizujú v štatistickom systéme R. R je voľne šíriteľné softvérové prostredie slúžiace na štatistické výpočty a grafické znázornenia. Funguje pod rôznymi druhmi UNIX platforiem, Windowsom aj MacOS. Jazyk programu R je podobný jazyku S, s istými obmenami. Vytvorili a rozvíjali ho Bell Laboratories (John Chambers a kolektív). R je Open Source, čo dáva možnosť tvoriť vlastné funkcie, knižnice a tvorcovia jednotlivých knižníc iba požadujú aby ste ich pri využívaní ich knižníc v prácach citovali (R Core team, 2012).

Na účely našej práce budeme využívať prostredie RStudio, ktoré funguje pod programom R. To zahŕňa konzolu, možnosť priamej exekúcie príkazov, grafické znázorňovanie, históriu činností, manažment pracovného prostredia a rôzne iné možnosti.

1.6 Metodika postupu

V tejto práci budeme postupovať podľa nasledovnej štruktúry:

1. Implementovanie dát do systému R.
2. Skúmanie anomálii alebo nežiaduceho správania sa dát a ich úprava.
3. Vytvorenie trénovacej a testovacej množiny dát.
4. Implementovanie do algoritmov podmienených stromových štruktúr.
5. Vyhodnotenie dosiahnutých výsledkov a prípadné zopakovanie bodu 4 s inými parametrami.

2 ANALYTICKÁ ČASŤ

V tejto kapitole budeme bližšie skúmať množinu údajov výstupu z bankového marketingu. Zameriame sa na implementáciu dát do modelov a nastavenie modelov tak, aby výsledky vykazovali čo najväčšiu predikčnú presnosť a v neposlednom rade aj zhodnotenie faktorov, ktoré vplývali na úspech alebo neúspech ponúkaného produktu v kampani. Pred tým však, než môže nastať implementácia dát do modelov, je odporúčaným štatistickým postupom prezrieť si dátovú množinu , prípadne identifikovať a eliminovať anomálie v nej alebo iné nežiaduce správanie. Tieto kroky v práci nebudeme bližšie vysvetľovať ani ich popisovať nakoľko nie sú predmetom práce. Ich aplikácia, ktorá vychádza aj zo stanovenej metodiky postupu je samozrejmá.

Po odstránení anomálií ostalo v súbore 45 202 klientov a 17 premenných, ktoré pozorovaný stav klientskeho správania zachytávajú.

Vytvorenie trénovacej a testovacej množiny dát

Pred vytvorením modelov podmienených stromových štruktúr sme z pôvodného súboru dát vytvorili 2 súbory vo veľkostnom pomere 80:20. Súbor s 80% veľkosťou pôvodného súboru tvorí takzvanú trénovaciu množinu a súbor s 20% veľkosťou dát tvorí takzvanú testovaciu množinu, pričom v oboch súboroch je zachovaný rovnaký pomer klientov, ktorí pozitívne reagovali na marketingovú kampaň.

Tento krok je zavádzaný kvôli tomu, aby model vytvorený na dátach dosahoval čo najväčšiu predikčnú schopnosť na inej dátovej množine než na tej, na ktorej bol vytvorený. Preto je potom model po jeho vytvorení odskúšaný na takzvaných testovacích dátach a predikčnú schopnosť na týchto dátach sa snažíme maximalizovať.

2.1 Tvorba podmieneného stromu

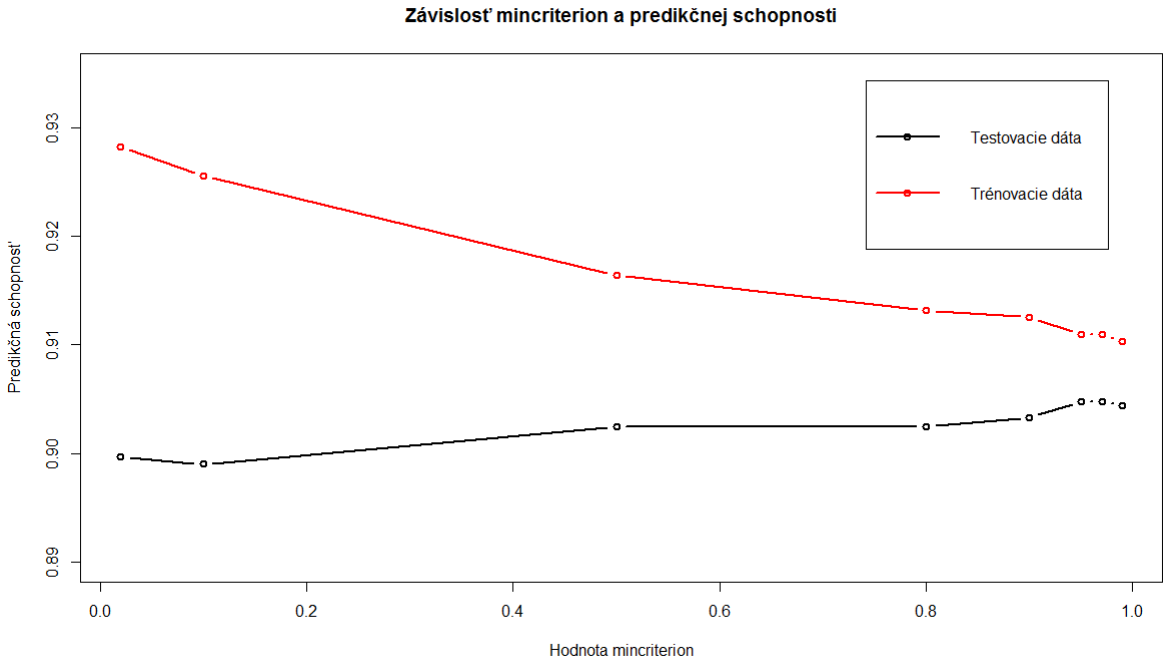
Na vytvorenie podmieneného stromu využívame funkciu *ctree()* z balíčka `party` (Hothorn, Hornik, Zeileis, 2006). Zápis v programe R na spustenie algoritmu bol nasledovný:

```
>objekt <- ctree(y ~ ., data=train, controls=ctree_control(testtype
= "Bonferroni", mincriterion = 0.95))
```

Pri nastavovaní tvorby modelu je v tejto funkcii možné využiť viacero kritérii, ktoré formujú jeho vznik. Jedným z nich je aj hodnota „mincriterion“, ktorá stanovuje hodnotu p pre zamietnutie nulovej hypotézy. Konkrétne je to hodnota (1-p) a teda, pokiaľ chceme aby

sa pri nedosiahnutí p-hodnoty menšej ako 0,05 algoritmus zastavil, zvolíme hodnotu mincriterion - 0,95 (ako je názorne ukázané aj v zadanom kóde). To, prečo je tento parameter podstatné optimálne nastaviť, ukážeme na obrázku nižšie.

Obrázok 1: Vplyv zmeny hodnoty parametra Mincriterion na predikčnú schopnosť



Prameň: Vlastné spracovanie

Uvedený obrázok poukazuje na závislosť predikčnej schopnosti modelu v závislosti od zvolenej hodnoty Mincriterion na trénovacej aj testovacej množine dát. Stanovené hodnoty Mincriterion sa pohybujú v intervale (0,02; 0,99). A týmto hodnotám odpovedajú výstupy predikčnej schopnosti z intervalu (0,899; 0,928).

Priamo pozorovateľný fakt je, že na testovacie dáta vplýva hodnota Mincriterion inak ako na trénovacie dáta. Je to pochopiteľné, pretože kým na trénovacej množine je nižšia hodnota závislosti vhodná na lepšiu predikciu, tak takto postavený model, s klesajúcou hodnotou Mincriterion sa stáva stále menej použiteľným na inú dátovú množinu než na tú na ktorej bol zostrojený a teda napríklad aj na množinu testovacích dát.

Je prirodzené, že predikčný rozdiel na testovacích a trénovacích dátach bude vždy, našou úlohou je tento rozdiel minimalizovať. V tomto modeli je rozdiel minimalizovaný pri najvyšších troch nameraných hodnotách Mincriterion a konkrétne sú to hodnoty: 0,95; 0,97;

0,99. Z tohto pozorovania vyplýva, že čím prísnejšie je kritérium na nulovú hypotézu, tým lepší výsledok predikcie modelu ako celku môžeme čakať.

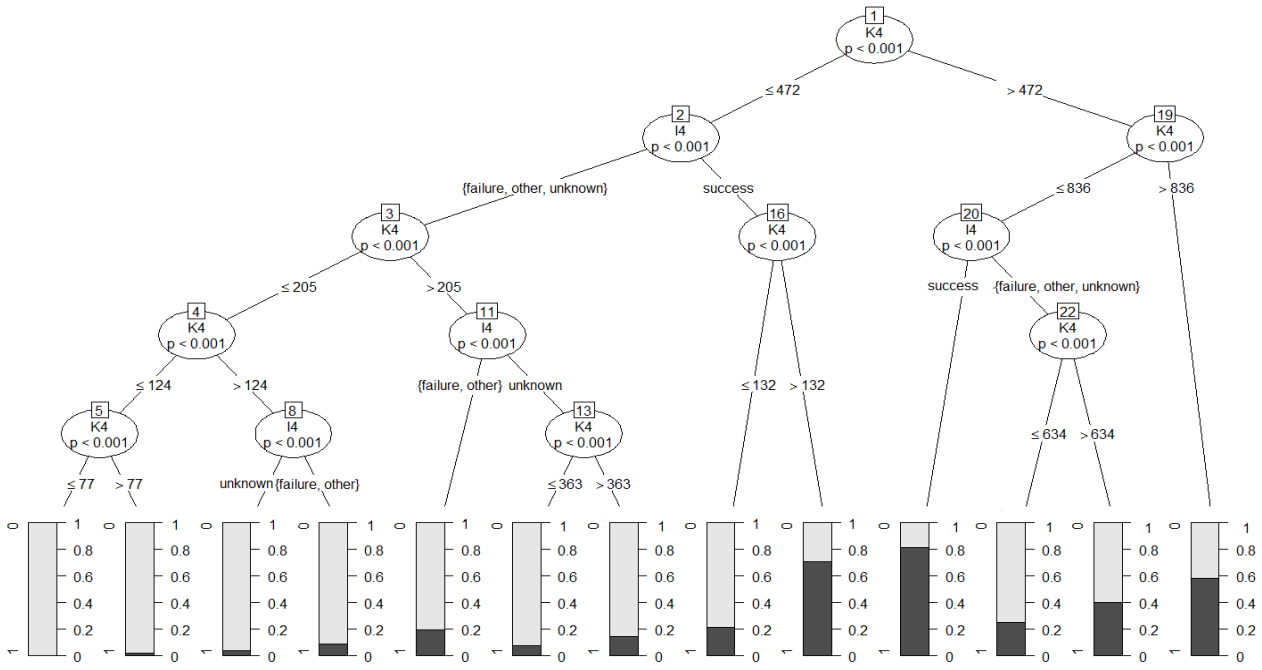
Ďalšou možnosťou nastavenia algoritmu, je nastavenie váh pre konkrétne hodnoty premenných. Pokiaľ určité pozorovanie obsahuje nameranú hodnotu premennej, ktorá je z intervalu pre nastavenie váh, priradí sa tomuto pozorovaniu vyššia váha oproti ostatným.

2.2 Popis modelom vytvoreného stromu

Model stromu, ktorý algoritmus vytvoril je graficky znázornený ako strom otočený vrchom nadol. Strom má svoj začiatok na vrchole odkiaľ sa postupne vetví do väčších rozmerov. Čím je uzol v hierarchii vetvenia vyššie, tým je jeho dôležitosť väčšia. Môžeme teda povedať, že premenné na základe ktorých sú klienti klasifikovaní majú väčšiu predikčnú schopnosť, čím vyššie sa v stromovej štruktúre nachádzajú.

Na obrázku nižšie je grafické zobrazenie podmieneného stromu. Pre účely ukážky, ako vyzerá výsledný model stromu sme zvolili zobrazenie len vrchných uzlov, pozostávajúce z 2 premenných, pretože zobrazenie celého stromu bolo príliš neprehľadné a nečitateľné z dôvodu jeho veľkej košatosti. Spomenuté premenné sú K4 a I4.

Obrázok 2: Ukážka podmieneného stromu



Prameň: Vlastné spracovanie

Na obrázku, tvary elipsy predstavujú uzly delenia stromu. Obsahujú premennú, ktorá v konkrétnom uzle zabezpečuje kritérium delenia a na vetvách po uzlami sú konkrétne hodnoty, ktoré boli na klasifikovaných klientoch namerané. Vytvorený model stromu v priebehu delenia dosiahol 13 listov (koncových bodov stromu) a každý z nich je graficky znázornený stĺpcovým grafom pod ukončením spodnej vetvy. Časť stĺpca s čiernou výplňou predstavuje podiel klientov, ktorí pozitívne reagovali na marketingovú kampaň a teda produkt od banky získali. Ako sme už spomenuli, Obrázok 2 je iba ukážkou toho ako strom vyzerá za použitie dvoch premenných. V texte nižšie popisujeme úplný skonštruovaný podmienený strom, ktorý v práci nie je graficky zobrazený, kvôli jeho neprehľadnej košatosti.

Úplný podmienený strom

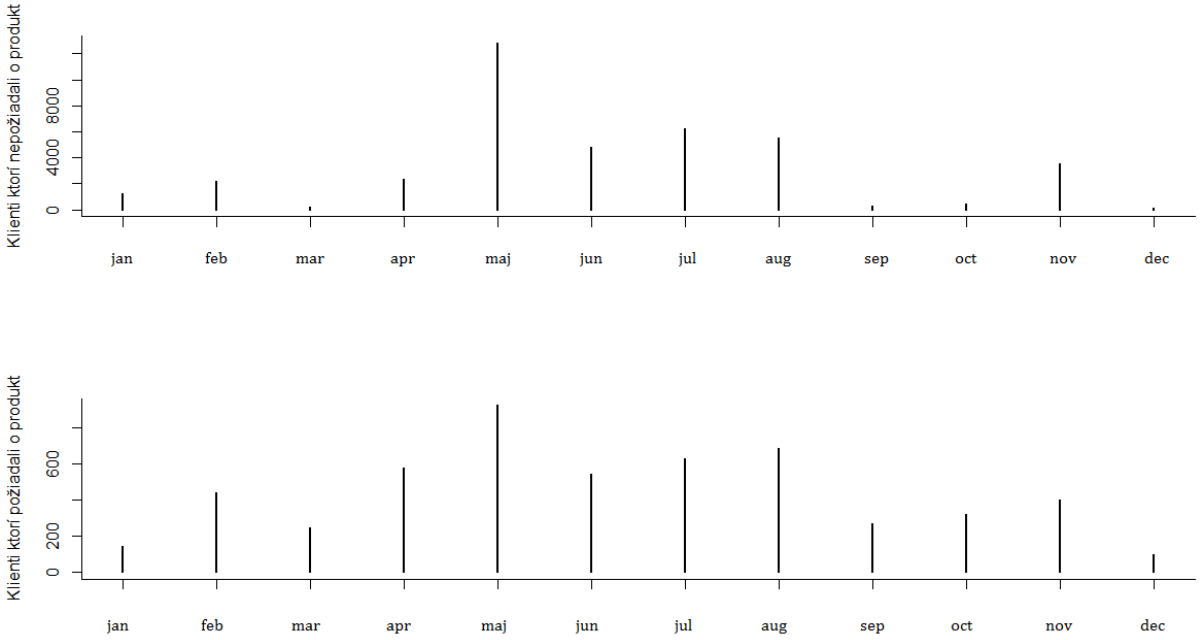
V hlavnom uzle (na vrchole stromu) sa strom delí na základe premennej K4 – dĺžka posledného kontaktu s klientom v sekundách. Pokiaľ dĺžka tohto styku s klientom bola menšia než 472 sekúnd (takmer 8 minút), klient ďalej pokračoval ľavou vetvou z hlavného uzlu, v opačnom prípade pravou vetvou. Takéto deliace kritérium má svoje logické opodstatnenie. Pokiaľ klient záujem o produkt, je pravdepodobné že nemal záujem s poskytovateľom volať viac než 8 minút. Na druhej strane, ak klient prejavoval o produkt záujem, je prirodzené že tento kontakt trval dlhšiu dobu v záujme získania informácií klientom.

Ďalším uzlom, z ľavej strany hlavného uzla bolo delenie na základe premennej I4 – výstup z minulých marketingových kampaní a teda to, ako klient po oslovení na kampaň reagoval. Táto premenná je kategorická a obsahuje 4 možnosti odpovede – „úspech, neúspech, neznáme, iné“. V konkrétnom uzle, pokiaľ výstup z predošlých kampaní bol buď: neúspech, neznáme alebo iné, títo klienti boli ďalej radení do ľavej vetvy, v prípade úspechu do pravej vetvy. Ďalším možným úsudkom pri pozorovaní tohto delenia môže byť, že klienti ktorí už v minulosti boli úspešne oslovení bankou, majú väčšiu tendenciu byť úspešne oslovení aj v ďalším kampaniach. Súvisí to zrejme s dobrou skúsenosťou klienta, menšou podozrievavosťou a inými faktormi.

Druhým kritérium po hlavnom uzle z pravej strany, bolo opäť delenie podľa premennej K4 a konkrétne šlo o klientov pri ktorých K4 > 129. Môžeme povedať že faktor dĺžky času v úspešnosti oslovenia klienta zohrával naozaj veľkú úlohu.

Medzi premenné na vrchu stromu patrilo K4, I4, K3, I1. Z vymenovaných premenných sú logickou úvahou pochopiteľné 3 z nich. Predikčná schopnosť premennej: „mesiac posledného kontaktu s klientom“ vzhľadom na neúplnú podrobnosť informácií s ktorými disponujeme nám nedáva logický súvis. Môže však ísť o určitú previazanosť medzi určitým mesiacom a klientskym správaním voči ponuke banky. V jednom z vrchných uzlov stromu sú deliacim kritériom premennej mesiace: marec, september a október. Hoci nemáme poznatky o tom, aké je prepojenie týchto mesiacov a klientskeho rozhodnutia, na grafe nižšie je jasne viditeľné, prečo sú práve tieto tri mesiace deliacim kritériom na vrcholných pozíciách podmieneného stromu.

Obrázok 3: Správanie klientov v závislosti od mesiacov v roku



Prameň: Vlastné spracovanie

Na obrázku vidíme 2 grafy. Graf na vrchu v jednotlivých mesiacoch ukazuje klientov, ktorí o produkt nepožiadali a graf nižšie zobrazuje klientov, ktorí o produkt požiadali. Práve mesiace: marec, september a október sú tými kde je rozdiel medzi tými ktorí nepožiadali o produkt v pomere s tými ktorí oň požiadali najviac rozdielny. Preto aj algoritmus môže najlepšie predikovať, práve na základe týchto mesiacov, kde je pomer klientskeho správania vzhľadom na ostatné mesiace neštandardný.

2.3 Klasifikačná tabuľka

Klasifikačná tabuľka je nástroj, ktorý slúži na to, aby sa bolo možné hlbšie skúmať predikčnú presnosť modelu. Táto tabuľka ukazuje, akým spôsobom bolo koľko pozorovaní klasifikovaných. Uvádzame tabuľku k prvému zostrojenému stromu, ktorý sme vyššie popísali.

Tabuľka 2: Klasifikačná tabuľka modelu

	Modelom klasifikovaní ako klienti, ktorí o produkt nepožiadali	**Modelom klasifikovaní ako klienti, ktorí o produkt požiadali**
Klienti, ktorí o produkt nepožiadali	7629	343
Klienti, ktorí o produkt požiadali	518	551

Prameň: Vlastné spracovanie

Tabuľka sa skladá z dvoch riadkov a dvoch stĺpcov. Výsledok súčtu prvého riadku je počet klientov ktorí nepožiadali o produkt. Analogicky, pri druhom riadku je výsledkom súčtu, počet klientov ktorí o produkt požiadali. Z hľadiska stĺpcov, súčet hodnôt prvého stĺpca hovorí o počte klientov, ktorých model predikoval ako tých, ktorí o produkt nepožiadali a v druhom stĺpci ako tých, ktorí oň požiadali. Na základe toho môžeme vidieť, že model nesprávne klasifikoval určitý počet klientov ktorí nepožiadali, aj tých ktorí o produkt požiadali.

Výpočet chybovosti modelu tých, ktorí nepožiadali o produkt a boli klasifikovaní ako tí čo požiadali vypočítame: =343/(343+7629). Výsledkom je 0,043 a teda 4,3%-ná chybovosť klasifikácie modelu tejto skupiny.

Podobne aj chybovosť klasifikácie tých, ktorí požiadali a boli označení ako keby nepožiadali vypočítame: =518/(518+551). Chybovosť pri týchto klientoch je 48,5%.

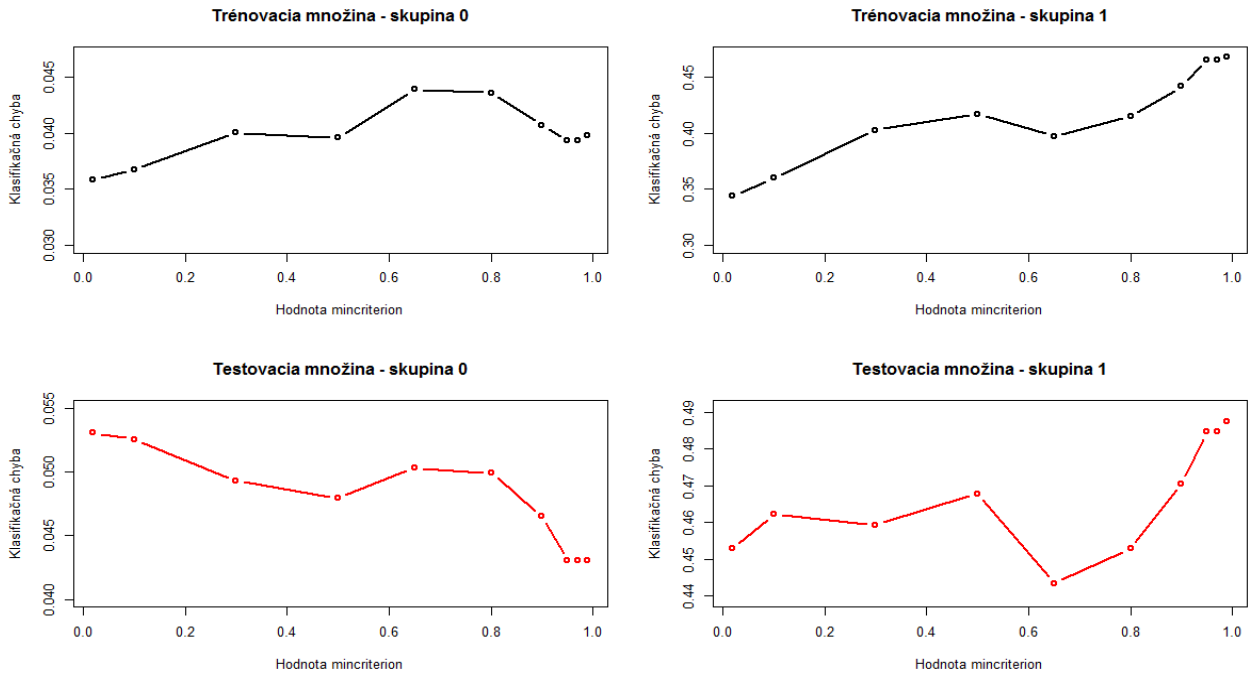
Ak chceme vypočítať celkovú chybovosť klasifikácie, tak postupujeme vydelením sumy hodnôt na diagonále celkovou sumou hodnôt: =(343+518)/(343+518+7629+551). Pre model ako celok je klasifikačná chyba 9,5%.

Jav ktorý nastal pri tomto predikčnom modeli, a teda že jedna skupina premennej má nízku chybovosť klasifikáciu a druhá relatívne vysokú vyzerá neštandardne. Avšak je to jav,

ktorý možno očakávať, pri podiele klientov ktorí o úver požiadali na celkovej vzorke klientov. Je to len 13,2% zo všetkých klientov. Keďže táto skupina má podstatne menší rozsah, je o to náročnejšie správne ju klasifikovať.

Pri snahe o zlepšenie predikčnej chyby je teda nevyhnutné zamerať sa na správnu klasifikáciu buď jednej skupiny alebo druhej. Vždy pôjde o zlepšenie chybovosti klasifikácie jednej skupiny na úkor druhej. Na nasledovnom obrázku, ukážeme ako sa v našom súbore klientov vyvíjajú klasifikačné chybovosti skupín na základe zmeny Mincriterion.

Obrázok 4: Porovnanie vývoja klasifikačnej chybovosti v skupinách



Prameň: Vlastné spracovanie

Na obrázku vidíme na ľavej strane vývoj klasifikačnej chyby v závislosti od nastavenia hodnoty Mincriterion na skupine ktorá nepožiadala o produkt. Na pravej strane ide naopak o skupinu, ktorá o produkt požiadala. Je dôležité upozorniť, že intervaly klasifikačnej chyby sú pre každý graf iné. Celá škála nie je zobrazená kvôli tomu aby bol jasne pozorovateľný vývoj jednotlivých skupín.

Kým na trénovacej množine dát Obrázka 2 je možné pozorovať podobný vývoj ako na Obrázku 1 (aj keď so značnými odchýlkami), na testovacej množine dát je tento vývoj veľmi nejednoznačný hlavne pri skupine, ktorá požiadala poskytovaný produkt. Ten samotný fakt

iba hovorí o tom, čo z obrázku nie je vidieť, a to je podiel veľkosti skupín na celkovej množine dát. Keďže je oveľa náročnejšie správne klasifikovať a teda aj predikovať pomocou modelu skupinu s takýmto malým podielom na dátach, vzniká v rámci modelovania zobrazené neštandardné správanie.

Z vyššie uvedeného obrázku je však tiež zrejmé, že pokiaľ by sme sa chceli zamerať na čo najlepšie odhadnutie skupiny klientov ktorí požiadali o produkt, hodnotu Mincriterion by sme nastavovali niekde okolo hodnoty 0,65 a ak by sme sa zameriavali na skupinu, klientov ktorí nepožiadali, hodnotu Mincriterion by sme nastavili okolo 0,95.

ZÁVER

Keďže konkurencieschopnosť banky, je v dobe s veľkým počtom bánk naprieč všetkými vyspelými krajinami kľúčovým faktorom úspechu, snaha dosiahnuť takýto stav je z hľadiska banky viac než žiadaná. Správna predikcia úspešnosti oslovenia nového klienta alebo úspešné predanie produktu dlhoročnému klientovi je základom podnikateľského plánu banky. Preto sme v tejto práci zostavovali model, ktorým je možné odhadnúť klientske správanie sa na základe získaných údajov o nich.

V rozsahu tejto práce sme vymedzili skúmanú problematiku: priamy bankový marketing a jeho aplikácie. Predstavili sme konkrétnu dátovú množinu, ktorá je voľne dostupná a obsahuje údaje z oblasti bankového marketingu. Na vyhodnocovanie týchto údajov a hľadania implicitných informácií v nich sme priblížili data miningové klasifikačné metódy, ktoré majú potenciál byť vhodným nástrojom riešenia tejto problematiky.

V analytickej časti sme na konkrétne, spomenuté dáta aplikovali algoritmus podmienených stromov v štatistickom programe R. Okrem jeho zobrazenia a popisu, sme venovali časť práce aj analýze toho, ako jednotlivé parametre, ktoré vieme v algoritme nastaviť, ovplyvňujú predikčnú schopnosť modelu najskôr na trénovacích ale hlavne na testovacích dátach.

Závery z tejto práce hovoria o tom, ktoré faktory (premenné) najlepšie vysvetľujú/predikujú to, či sa klient banky rozhodol pre produkt alebo nie spolu s logickými odôvodneniami toho, prečo sú to práve tie premenné. Tiež hovoria o tom ako a prečo je dôležité citlivo nastavovať parametre algoritmu podmienených stromov.

Premenné s najvyššou predikčnou schopnosťou sú:

- dĺžka posledného hovoru s klientom v sekundách (K4),
- výstup z predošlých marketingových kampaní (I4),
- mesiac posledného kontaktu s klientom (K3),
- počet kontaktov s klientom (I1).

Za predpokladu poznania týchto 4 údajov o klientovi, o ktorom nevieme ako reagoval na ponuku produktu v marketingovej kampani, vieme z poskytnutých premenných v súbore s najväčšou pravdepodobnosťou správne určiť medzi ktorých klientov by patril.

Nastavovanie algoritmu podmienených stromov má viacero možností. Jednou z nich je hodnota parametra Mincriterion. V práci sme ukázali, že je veľmi dôležité túto hodnotu správne nastaviť, pretože predikčné odchýlky ktoré môže nastavenie tohto parametra priniesť sú relatívne veľkého charakteru. Tiež je dôležité brať do úvahy fakt, že hodnota Mincriterion má iný vplyv na trénovacie a iný vplyv na testovacie dáta a tiež je veľmi dôležité stanoviť si, ktorú skupinu zo závislej premennej chceme ako analytici uprednostniť, resp. predikcia ktorej zo skupín klientov má pre nás vyššiu výpovednú hodnotu a na tú sa zamerať aj na úkor tej druhej skupiny.

Na záver môžeme skonštatovať, že vytvorený model podmieneného stromu má svoju vypovedaciu hodnotu a s nepresnosťou 10% vie správne predikovať zaradenie klienta do tej skupiny klientov, do ktorej patrí. Z tohto hľadiska a na základe tejto práce, môžeme data miningovú metódu podmienených stromov zaradiť medzi aplikovateľné metódy na analýzu oblasti bankového marketingu, ktorej značnou výhodou je relatívne jednoduchá interpretovateľnosť.

POUŽITÁ LITERATÚRA

1. DAHAN, H. a kol. 2014. *Proactive Data Mining with trees*. Springer. 2014. 94s. ISBN 978-1-4939-0538-6.
2. DE VILLE, B. 2006. *Decision trees for bussiness intelligence and Data Mining: Using SAS Enterprise Miner*. Cary: SAS Institute Inc., 2006. 241s. ISBN 978-1-59047-567-6.
3. FRAWLEY, W. J a kol., 1992. *Knowledge discovery in databases: An overview*. *AI magazine*, 1992, 13.3: 57.
4. HOTHORN, T. - HORNIK, K. - ZEILEIS, A. 2006. *Unbiased recursive partitioning*: *A conditional inference framework*. Journal of Computational and Graphical statistics, 2006, *15*(3), 651-674.
5. LAROSE, D. T. 2005. *Discovering knowledge in data*. New Jersey: John Wiley& Sons, 2005. 222s. ISBN 0-471-66657-2.