



Ekonomická teória a prax z pohľadu študentov

Problémy portácie prediktívnych analytických modelov
z komerčných na open - source aplikácie.

Ivana Veljačiková

Sekcia: kvantitatívne metódy a informatika

Konzultant: Doc. RNDr. Juraj Pančík, CSc

Banská Bystrica
2011

Obsah

Zoznam grafov.....	3
Zoznam ilustrácií.....	3
Úvod.....	4
1 Teoretické východiská dataminingu a rizika stornovania zmluvy.....	5
1.1 Riziko stornovanosti zmluvy.....	5
1.2 Datamining.....	6
1.3 Definícia dataminingu.....	6
1.4 Úlohy data miningu.....	7
1.5 Techniky využívané v dataminingu.....	7
1.6 Dataminingové metódy.....	10
1.7 Datamining v poisťovníctve:.....	12
1.8 Softvérová podpora – WEKA.....	12
1.9 Charakteristika Poisťovne a produktové portfólio.....	14
2 Tvorba modelov rizika s využitím softvéru WEKA.....	16
2.1 Štruktúra údajov.....	16
2.2 Tvorba dataminingových modelov.....	23
Záver.....	27
Zoznam bibliografických odkazov.....	28
Prílohy.....	29

Zoznam grafov

Garf 1: Veková štruktúra klientov, ktorí zrušili zmluvu pri produkte 1A

Garf 2: Spôsoby zrušenia zmlúv

Garf 3: Veková štruktúra poistencov pri produkte 3A

Garf 4: Analýza aktívnych a zrušených zmlúv podľa krajov

Zoznam ilustrácií

Obrázok 1: CRISP-DM proces

Obrázok 2: Weka Explorer

Obrázok 3: ROC krivka pre produkt 2A

Úvod

Cieľom práce je portácia prediktívnych modelov medzi dvoma programovými prostrediami - z komerčného prostredia programovej sady SAS do cenovo výhodnejšieho prostredia voľne šíriteľného programového vybavenia WEKA a analytických nástrojov, ktoré poskytuje.

Pri tvorbe modelov sme postupovali podľa analýzy a modelov, ktoré zostavili Ing. Hromada a Ing. Hujíková. Pôvodné modely boli vytvorené komerčným nástrojom SAS Enterprise Miner (SAS EM). Vstupné údaje pootrebné na analýzu a tvorbu modelov nám poskytla poisťovňa fungujúca na slovenskom poistnom trhu, ktorú však nebudeme menovať.

V prvej kapitole sme charakterizovali riziko stornovania zmluvy, pretože modely sú postavené na skúmaní správania sa klienta práve pri stornovaní poistnej zmluvy. Ďalej je stručne opísaná problematika dataminingu a metódy, ktoré využíva. V tejto kapitole je predstavený aj softvér WEKA. Taktiež je stručne charakterizovaná Poisťovňa a jej produktové portfólio.

V druhej kapitole sa zaoberáme samotnou analýzou a tvorbou modelov. Načrtli sme štruktúru a štatistické parametre vstupných údajov a opísali sme postup a výsledky samotných modelov.

1 Teoretické východiská dataminingu a rizika stornovania zmluvy

1.1 Riziko stornovanosti zmluvy

Špecifickým rizikom, ktorým sa budeme zaoberať v tejto kapitole je riziko predčasného ukončenia zmlúv - tzv. riziko stornovanosti. K ukončeniu zmluvy dochádza buď zo strany klienta, alebo nezaplatenia predpísaného poistného.

Získanie klienta a uzatvorenie zmluvy je pre spoločnosť finančne náročná záležitosť, preto je z nákladového hľadiska výhodnejšie si klienta udržať, hoci to nie je vždy jednoduché. Na to, aby si poisťovne udržali svojich klientov, musia vynaložiť značné úsilie v podobe starostlivosti o klienta, poskytnúť mu dostatočné informácie o danom produkte a navrhnúť mu čo najvhodnejšie riešenie.

K rušeniu zmlúv zo strany klienta dochádza z rôznych dôvodov. Na jednej strane to môže byť nedostatočná informovanosť o produkte, ku ktorej dochádza napríklad kvôli nedorozumeniam pri obchodných rokovaniach s finančným agentom, alebo podvodné správanie sprostredkovateľa poistenia.

Na druhej strane môže ísť o nepriaznivú finančnú situáciu klienta, ktorú mohla zapríčiniť strata zamestnania alebo dlhodobá práceneschopnosť, prípadne iné rodinné záležitosti.

Príčinou však môžu byť aj iné demografické faktory, ktoré poisťovne primárne nesledujú. Ich vplyv na správanie sa klienta nie je zjavný. Predpokladom je, že medzi tieto faktory patrí najmä prostredie, v ktorom poistník žije a pracuje – kraj resp. okres, priemerná mzda, výška nezamestnanosti, počet voľných pracovných miest v danom okrese, či kraji poistníka.

K najväčšiemu počtu zrušených zmlúv dochádza najmä počas prvého roka poistenia. Poisťovne každoročne sledujú a vyhodnocujú stornovanosť zmlúv a snažia sa efektívne znížiť mieru stornovanosti. Pri znižovaní rizika ukončenia zmluvy využívajú rôzne metódy. Medzi základné metódy patrí aj upomínanie klienta, komunikácia a starostlivosť o klienta prostredníctvom obchodných zástupcov a finančných agentov, prípadne možnosť dohody s klientom o znížení poistného.

1.2 Datamining

Datamining vznikol v devädesiatych rokoch dvadsiateho storočia, s rozvojom informačných technológií a s pribúdaním údajov v súboroch a databázach, spojením štatistických metód a umelej inteligencie. Odborná literatúra prekladá datamining ako dolovanie z dát. Taktiež je možné tento výraz preložiť ako hĺbkový prieskum údajov, ktorý nám ponúka možnosť využiť hodnotné informácie z minulosti na predikciu a plánovanie udalostí v budúcnosti.

„V posledných rokoch datamining priťahuje veľkú pozornosť informačnej spoločnosti, ale aj spoločnosti ako celok, vzhľadom k širokej dostupnosti veľkého množstva údajov a bezprostrednej potrebe nadobudnúť z týchto údajov cenné informácie a vedomosti. Takto získané informácie môžu byť použité v oblastiach od marketingových analýz, cez odhaľovanie podvodov až po udržanie si zákazníkov, riadenie výroby a vedecké skúmanie.“ (Han, 2006, s. 1)

1.3 Definícia dataminingu

V dostupnej literatúre je datamining definovaný ako „je proces výberu, prieskumu a modelovania na báze veľkých objemov údajov, s cieľom odhaliť predtým neznáme informačné vzory v údajoch, pre výhody v konkurenčnom prostredí.“ (<http://sas.com>, 2010) Ďalšiu inú definíciu uvádza Two Crows Corporation:

- „Datamining je proces, ktorý využíva rôzne druhy nástrojov analýzy dát na zistenie modelov a vzťahov medzi údajmi, ktoré môžu byť použité na utvorenie platných prognóz.“ (Two Crows Corporation, 2005, s.1)

Datamining využíva štatistické metódy na analýzu dát pre riešenie ekonomických problémov. Kým marketingový výskum vychádza z príležitosti na makroúrovni, datamining umožňuje objavovať možnosti, ktoré nie sú okamžite zrejmé, a môžu byť detekované pomocou štatistických metód a analýz. Marketingový výskum a datamining sú dve disciplíny, ktoré sú vzájomne komplementárne a pre spoločnosť dokážu priniesť benefity. (Chiu, 2008)

V literatúre sú s dataminingom spájané aj ďalšie odborné výrazy, ktoré sú veľmi podobné s týmto pojmom. Ide predovšetkým o dolovanie vedomostí z údajov

(knowledge mining), extrakcia znalostí (knowledge extraction), ale aj analýza dát, vzorov (data/pattern analysis). Ďalším výrazom je aj objavovanie poznatkov z údajov (Knowledge Discovery from Data, alebo KDD), niektorí autori však považujú datamining len za jednu časť v procese knowledge discovery from data. (Han, 2006, s.5)

1.4 Úlohy data miningu

Datamining je možné aplikovať na riešenie veľkého množstva problémov. Najskôr sa však musíme rozhodnúť, či cieľom pri vytváraní dataminingového modelu bude meranie alebo predikcia. Cieľové modely a následne aj úlohy sú rozdelené do dvoch kategórií – prediktívnych a deskriptívnych. Prediktívne modely vychádzajú zo štatistiky a matematiky, teda počítajú určité hodnoty. Deskriptívne modely vytvárajú pravidlá pomocou, ktorých sa predmety zoskupujú do popisných kategórií. (Parr Rud, 2001) Podľa druhu problému môžeme úlohy zoskupiť do nasledovných skupín, pričom prvé štyri skupiny patria k prediktívnym modelom a posledné dve skupiny úloh k deskriptívnym modelom:

- **opis** – vysvetlenie problému môžeme nájsť v opise konkrétnych vzorov a trendov
- **klasifikácia** – triedenie objektov do skupín na základe cieľovej premennej
- **odhad** (regresia) – je podobná klasifikácii, avšak cieľová premenná má numerickú hodnotu a nie kategorickú
- **predikcia** – predpoveď očakávaných hodnôt, pričom správnosť predpovede je možné určiť až v budúcnosti
- **zhlukovanie** – identifikuje skupiny objektov charakterizované určitými premennými
- **asociácia** – identifikácia produktov, ktoré sa často predávajú spoločne

1.5 Techniky využívané v dataminingu

Na nájdenie užitočných vzorov a trendov v dátach a pri vytváraní deskriptívnych a prediktívnych modelov sa využívajú štatistické, alebo iné analytické techniky na

spracovanie a analýzu dát.

Základom pre rozvoj rôznych techník dolovania dát sú štatistické ciele a typy dát, ktoré sú k dispozícii. Existujú dva spoločné štatistické ciele - analýza závislosti a analýza vzájomnej závislosti. Analýza závislosti sa využíva pri skúmaní vzájomných vzťahov medzi cieľovou premennou a prediktívnymi premennými. Analýza vzájomnej závislosti skúma súvislosti medzi nezávislými premennými, teda prediktívnymi. Predikcia v tomto kontexte znamená odhad možných hodnôt závislej premennej. Cieľová premenná sa tiež nazýva závislá premenná. (Chiu, 2008,)

V nasledujúcej časti popíšeme jednotlivé techniky, ktoré sa najčastejšie využívajú v dataminingu:

- Neurónové siete

Neurónové siete sú navrhnuté podobne ako neuróny v ľudskom mozgu. Dokážu efektívne spracovať obrovské množstvo údajov, ktoré obsahujú veľa prediktívnych premenných. Neurónová sieť sa skladá z vrstiev a každá vrstva obsahuje niekoľko uzlov. Prvá vrstva nazývaná aj vstupná (input layer) vrstva obsahuje uzly s prediktívnymi premennými, nasleduje niekoľko skrytých vrstiev (hidden layers). Jednotlivé uzly medzi vrstvami sú poprepájané. Poslednú vrstvu (output layer) tvoria výstupné uzly s cieľovými premennými. (Two Crows, 2005)

- Lineárna regresia

Podľa Olivie Parr Rud lineárna regresia patrí medzi štatistické metódy. Kvantifikuje závislosť medzi dvoma spojitými premennými – cieľovou a prediktívnou premennou. „V priebehu lineárnej regresie je hľadaná taká priamka prechádzajúca medzi jednotlivými bodmi, pre ktoré platí, že súčet druhých mocnín odchýlok od každého bodu je minimálna.“ (Parr Rud, 2001, s.10)

- Logistická regresia

Logistická regresia je podobná lineárnej regresnej analýze, avšak na rozdiel od lineárnej regresie využíva závislú premennú, ktorá má diskretný alebo kategorický charakter. (Parr Rud, 2001)

- Rozhodovacie stromy

„Cieľom rozhodovacieho stromu je sekvenčne rozdeliť údaje, tak aby sa maximalizovali rozdiely v závislej premennej. Skutočným účelom rozhodovacieho stromu je roztriediť údaje do odlišných skupín, či vetiev, ktoré vytvoria najsilnejšiu separáciu hodnôt závislej premennej.“ (Parr Rud, 2001, s. 14)

Rozhodovacie stromy je možné využiť pri zisťovaní skupín s určitým správaním.

- Asociačné pravidlá (Analýza nákupného koša)

Analýza nákupného koša je počítačový algoritmus, ktorý skúma množstvo transakcií s cieľom určiť, ktoré položky sú najčastejšie nakupované spoločne a poskytuje cenné informácie pre predajcu na uplatnenie lepšieho marketingu. (Fernandez, 2003)

- Bayesovské modely

„Bayesovské klasifikátory sú štatistické klasifikátory, ktoré predikujú pravdepodobnosti, s ktorými daný príklad patrí do tej – ktorej triedy. Pritom vychádzajú z určenia podmienených pravdepodobností jednotlivých hodnôt atribútov pre rôzne triedy.“(Paralič, 2003, s. 55)

Ďalšie techniky, ktoré uvádzame v nasledujúcom zozname nebudeme bližšie špecifikovať.

- Techniky najbližších susedov
- Zhlukovanie
- Metóda k-priemerov
- Analýza sekvencií
- Faktorová analýza
- Mechanizmy podporných vektorov
- Diskriminačná analýza

1.6 Dataminingové metódy

Na dosiahnutie optimálnych výsledkov v dataminingu je nevyhnutné postupovať podľa určitých pravidiel, postupností a krokov. Existuje viacero metód, ktoré sa využívajú v dataminingu. Medzi najpoužívanejšie metódy môžeme zaradiť CRISP-DM (Cross-Industry Standard Prodes for Data Mining) a SEMMA metóda vyvinutá SAS Institute.

CRISP-DM model sa skladá zo šiestich krokov a je to cyklický proces.

Skôr, ako sa začneme zaoberať samotnou prípravou modelu, musíme si ujasniť cieľ samotného modelovania ale aj to, za akým účelom model vytvárame. Následne je potrebné vypracovať plán a pri tímovej práci je nutné určiť zodpovedné osoby za jednotlivé časti projektu. Teda prvým krokom je **pochopenie problému** (business understanding)

Ďalším krokom je **pochopenie dát** (data understanding). „Tento krok môže obsahovať primárny zber údajov, popis údajov, ich prieskum a overovanie kvality údajov“ (Olson, 2008, s.9). Je vhodné vybrať potrebné údaje, ktoré sa môžu nachádzať vo viacerých databázach. Údaje rozdelíme do troch základných skupín: Demografické údaje – do tejto skupiny patria údaje o pohlaví, veku, rodinnom stave ale taktiež aj o príjme domácnosti a majetku. V skupine behaviorálnych informácií sa dozvieme o poisťovacích nárokoch výške platieb a podobne. Poslednou skupinou sú psychografické údaje, ktoré poskytujú informácie o životnom štýle, názoroch a hodnotách jednotlivca.

Nasleduje **príprava dát** (data preparation), kde sa uskutočňujú všetky aktivity zamerané na výber a čistenie údajov ako aj tabuliek, riadkov a pod., ktoré budú použité v konečnom súbore dát. Tento krok si vyžaduje vybrať premenné, ktoré budú najvhodnejšie reprezentovať a predikovať daný problém.

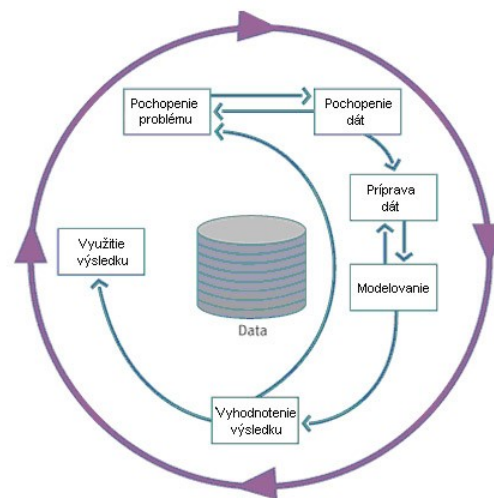
V procese **modelovania** (modeling) pomocou dataminingového softvéru aplikujeme jednotlivé metódy na konečný súbor údajov, pomocou ktorého vytvoríme niekoľko modelov, z ktorých následne vyberieme najvhodnejší.

Hodnotenie (evaluation). Modelové výsledky by mali byť hodnotené v rámci obchodných cieľov stanovených v prvej fáze (pochopenie problému). To povedie k

identifikácii ďalších potrieb (často prostredníctvom rozpoznávania vzorov), veľakrát je nutné vrátiť sa k predchádzajúcej fáze CRISP-DM (Olson, 2008, s.10).

Posledným krokom je **nasadenie** (deployment). Pozostáva z plánovania, kde a ako budú objavené poznatky využité. (Pal, 2005, s.8) Avšak je nevyhnutné sledovať možné zmeny vo vstupných údajoch, ktoré by mohli viesť k nesprávnym efektom z využívania takéhoto modelu.

Nasledujúci obrázok znázorňuje CRISP-DM proces:



Obrázok 1: CRISP-DM proces
Prameň: www2.fiit.stuba.sk

SEMMA (Fernandez, 2003, s.10)

- **Sample** – vyberie sa reprezentatívna vzorka údajov dostatočne veľká na to, aby obsahovala podstatné informácie a zároveň malá, aby manipulácia s ňou bola rýchla
- **Explore** – prieskum údajov slúži na odhalenie trendov a extrémnych hodnôt
- **Modify** – modifikovanie údajov, napr.: odstránenie nedefinovaných hodnôt, doplnenie popisov premenných, doplnenie nových informácií, vytvorenie zoskupení
- **Model** – tvorba konkrétnych modelov pre splnenie cieľa
- **Assess** – porovnanie a posúdenie vytvorených modelov

1.7 Datamining v poisťovníctve:

V poisťovníctve sa datamining uplatňuje pri riešení veľkého množstva problémov. Z marketingového hľadiska sa zameriava na optimalizáciu produktov a cien pomocou vhodnej segmentácie zákazníkov. Získavanie nových zákazníkov a riadenie marketingových kampaní na udržanie si súčasných klientov dokáže poisťovacím spoločnostiam ušetriť finančné a materiálové zdroje. Pri riadení rizika poisťovne využívajú datamining pri predikcii vlastností zákazníkov, ktorí s určitou pravdepodobnosťou zrušia zmluvu alebo nebudú platiť poisťné. Vďaka tomu je poisťovňa schopná efektívne riadiť svoje finančné zdroje. Aj keď mnoho rizikových faktorov ovplyvňujúcich cenu poznáme, medzi premennými môžu existovať aj neintuitívne väzby, ktoré je ťažké odhaliť bez uplatnenia sofistikovanejšej analýzy. (Guo, 2003) Odhaľovanie skrytých vzťahov, ktoré nevieme definovať, pomáha pri skúmaní podvodného správania sa.

1.8 Softvérová podpora – WEKA

„Pracovné prostredie WEKA je súborom algoritmov strojového učenia a nástrojov na spracovanie a úpravu údajov. Vďaka užívateľskému rozhraniu si užívateľ môže efektívne a rýchlo otestovať existujúce metódy. Podporuje celý proces dataminingu vrátane prípravy údajov, štatistického hodnotenia modelov a vizualizáciu vstupných údajov a výsledkov. Vďaka spoločnému užívateľskému rozhraniu užívatelia môžu porovnávať viaceré metódy a vybrať tú, ktorá je pre danú problematiku najvhodnejšia. Weka bola vyvinutá na novo-zélandskej univerzite University of Waikato a skratka Weka symbolizuje názov Waikato Environment for Knowledge Analysis.“(Witten, Frank, 2005, s. 365)

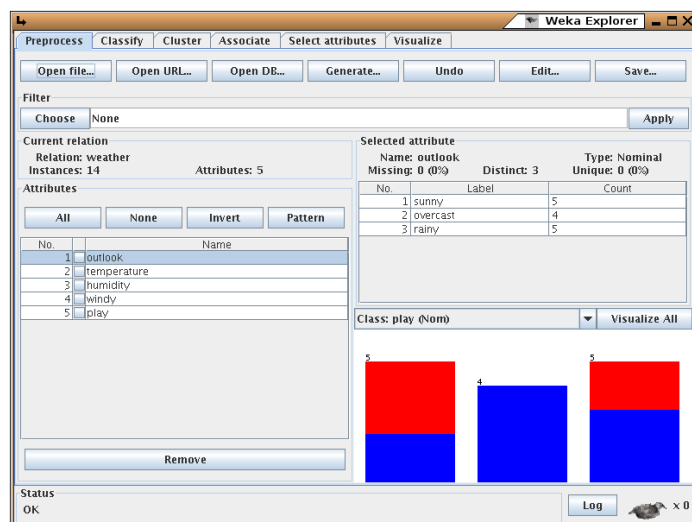
Výhody:

- Je distribuovaný pod GNU General Public License, teda slobodne šíriteľný (zadarmo). Vďaka tomu spoločnosti nemusia disponovať dostatkom voľných finančných zdrojov na zakúpenie licencie a pri rozšírení o nové funkcie nevznikajú žiadne dodatočné náklady.
- Veľkou výhodou je, že je spustiteľný na takmer všetkých platformách (Windows, Linux, Mac OS X).

- Obsahuje grafické užívateľské rozhranie, vďaka čomu je jeho používanie jednoduché.

WEKA podporuje niekoľko štandardných úloh dataminingu, konkrétne prieskum dát, zhľadávanie, klasifikáciu, asociáciu, vizualizáciu a funkciu výberu atribútov. Je navrhnutá tak, že je možné rýchlo a flexibilne otestovať dostupné metódy na súbore údajov.

Hlavné grafické užívateľské rozhranie je EXPLORER (Obr.3.). Umožňuje prístup ku všetkým možnostiam pomocou menu výberu a vyplnením formulára. Existuje šesť rôznych panelov, ktoré odpovedajú na rôzne úlohy dataminingu, ktoré WEKA podporuje.



Obrázok 2: Weka Explorer
Prameň: Vlastné spracovanie

1.9 Charakteristika Poist'ovne a produktové portfólio

Poist'ovňa pôsobiaca na slovenskom poisťovacom trhu vznikla takmer pred desiatimi rokmi. Jej činnosť je zameraná na uzatváranie poisťných zmlúv v oblasti životného poistenia. Predmetom činnosti podľa obchodného registra je vykonávanie poisťovacej činnosti pre poisťný druh životného poistenia podľa zákona č. 95/2002 Z. z. o poisťovníctve, konkrétne ide o poistenie pre prípad smrti, poistenie pre prípad dožitia, poistenie pre prípad smrti a/alebo dožitia, poistenie spojené s investičným fondom, alebo poistenie pre prípad úrazu či choroby, ak je pripoistením k hlavnému poisteniu.

Už počas krátkej existencie na trhu sa Poist'ovni podarilo získať stabilné postavenie na trhu životného poistenia, taktiež sa jej podarilo obsadiť významné miesto v hodnotení ekonomického časopisu TREND a získala titul Poist'ovňa roka.

Poist'ovňa svojim klientom ponúka množstvo poisťných produktov, ktoré neustále zlepšuje a prispôsobuje aktuálnym podmienkam na poisťnom trhu.

Produkt 1A

Poistenie pre prípad smrti, smrti následkom úrazu, trvalých následkov úrazu a dožitia ponúka aj produkt 1A. Poistné je platené v pravidelných intervaloch, a to: mesačne, štvrťročne, polročne alebo ročne. Taktiež je určené pre osoby od 18-65 rokov na minimálne 5 ročnú poisťnú dobu. Výhodou je, že výnos je zdanený len na konci poisťnej doby.

Produkt 1B

Produkt 1B je určený pre deti vo veku od 0-14 rokov. Minimálna poisťná doba je tiež 5 rokov. Poistenie sa vzťahuje pre prípad trvalých následkov úrazu, závažných ochorení a dožitia sa 19. roku. Poistné je platené jednorázovo. Disponibilné právo má poisťník, ale zhodnotené finančné prostriedky po dovŕšení 19. roku dostane poistený.

Produkt 2A

Ide sa o jednorázovo platené investičné životné poistenie, ktoré poskytuje poisťnú ochranu pre prípad smrti z akýchkoľvek príčin a dožitia sa konca dohodnutého obdobia. Produkt je určený osobám vo veku 6 týždňov až 60 rokov a je uzatváraný

minimálne na 10 rokov s možnosťou mimoriadnych vkladov alebo výberov. Výnos je optimalizovaný fixáciou hodnoty fondu. Výhodou je, že ak nastane poistná udalosť - teda smrť poisteného, pozostalým bude vyplatená buď garantovaná poistná suma, alebo aktuálna kapitálová hodnota poistnej zmluvy zvýšená o 5 % (závisí to od toho, ktorá hodnota je vyššia).

Produkt 2B

Produkt 2B je určený klientom, ktorým nevyhovuje jednorázovo platené poistné. Môžu si vybrať či poistné budú platiť mesačne, štvrťročne, polročne alebo ročne. Tiež si môže vybrať pomer medzi investičnou a rizikovou zložkou poistenia. Indexácia zabezpečí ochranu finančných prostriedkov pred infláciou. Ostatné parametre poistenia sú totožné s produktom 2A.

Produkt 3A

Produkt 3A je úrazovo orientované poistenie. Klient má možnosť poistiť sa pre prípad smrti, smrti následkom úrazu a trvalých následkov úrazu. Je možné ho využiť aj ako pripoistenie k iným produktom, ktoré nepokrývajú riziká spojené s úrazom. Vstupný vek poisteného je 18-64 rokov, platenie prebieha inkasom mesačne alebo ročne.

2 Tvorba modelov rizika s využitím softvéru WEKA

Po teoretickom vymedzení problematiky sa teraz zameriame na praktickú aplikáciu. Najskôr opíšeme vstupné údaje a neskôr tvorbu modelov.

2.1 Štruktúra údajov

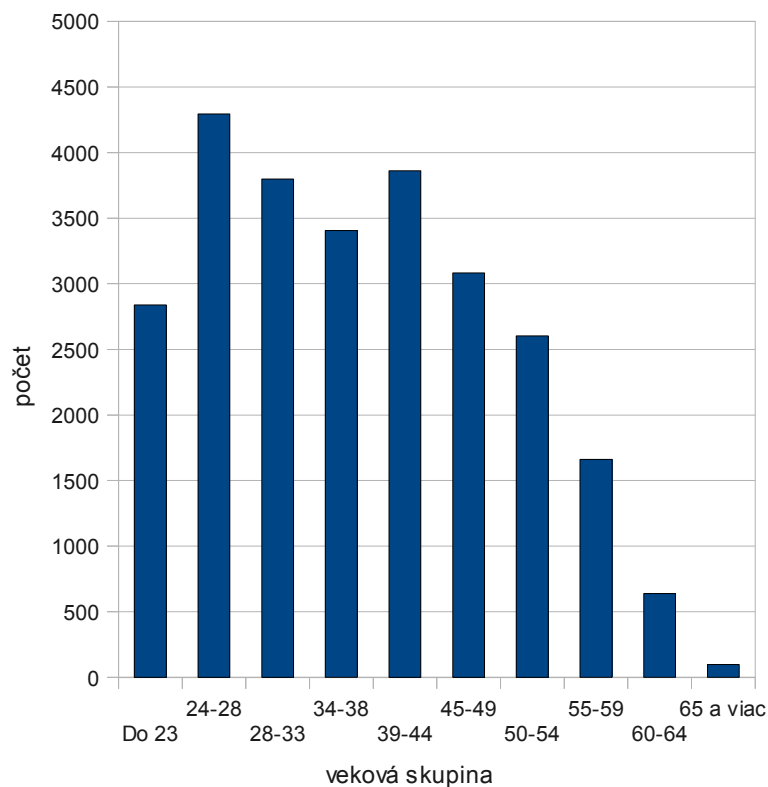
Keďže Poist'ovňa stanovuje mieru stornovanosti pre každý produkt osobitne, aj my budeme v tejto časti popisovať údaje podľa typu poistenia. Rozanalyzujeme databázu s údajmi a jednotlivými charakteristikami. Načrtujeme štruktúru premenných, ktoré budú v ďalšej kapitole vstupovať do dataminingových modelov.

Základnú charakteristiku jednotlivých produktov sme priblížili v predchádzajúcej kapitole. Teraz sa bližšie pozrieme na kvantitatívne a niektoré štatistické ukazovatele, ktoré je možné identifikovať pomocou softvérového nástroja Weka.

Produkt 1A

Produkt 1A sa od predchádzajúceho poistenia odlišuje spôsobom platenia. Klient si môže vybrať mesačné, štvrťročné, polročné alebo ročné platby poistného. Informácie potrebné na analýzu sa nachádzali v približne 50 000 záznamoch o zmluvách. Pomer aktívnych a zrušených zmlúv v tomto prípade bol 47:53. Poistnú zmluvu uzavrelo 52% mužov a 48% žien. Využitím filtra „Discretize“ sme vek poistených osôb rozdelili do siedmich intervalov. Vekové zloženie bolo pomerne rovnomerné. Najviac zmlúv klienti uzatvorili vo veku 25-31 rokov (19,5%), 44-51 rokov (18,4%). V skupine od 58 rokov to však boli len 3%. Až 40% klientov využilo možnosť poistiť sa na minimálnu dobu 5 rokov, pričom najčastejší spôsob platenia si vybrali mesačne (96%).

Analýza vstupných údajov ukázala, že najčastejšou príčinou zrušenia zmluvy bolo nezaplatenie bežného následného poistenia 49%. Odkupom bolo zrušených 27% zmlúv. Najviac zrušených zmlúv bolo v Košickom kraji, konkrétne 18,6%, za ním nasleduje Bratislavský a Banskobystrický kraj (12,9%). Takmer 11% všetkých zrušených zmlúv bolo ukončených v okrese Košice, v Bratislave bolo zrušených 9% zmlúv. Nasledujúci graf zobrazuje vekovú štruktúru klientov, ktorí zrušili poistnú zmluvu pri produkte 1A.



Garf 1: Veková štruktúra klientov, ktorí zrušili zmluvu pri produkte 1A

Prameň: Vlastné spracovanie

Produkt 1B

Posledné poistenie zo skupiny kapitálového životného poistenia je produkt 1B. Od predchádzajúcich produktov sa odlišuje tým, že poistník a poistený nie je tá istá osoba. V analýze budeme vychádzať z informácií o poistníkoch, ktorí dané poistenie uzatvárali a majú vplyv aj na zrušenie zmluvy. Ako sme už spomínali v predchádzajúcej kapitole, toto poistenie je určené pre deti od narodenia po dovŕšenie 14-teho roku.

Vzorka údajov obsahovala len 1300 záznamov. Pri tomto produkte je vysoký počet aktívnych zmlúv až 94%. Poistenie uzatvárajú častejšie ženy ako muži, ale pohlavie poistených osôb je takmer rovnaké (Príloha 1). Z geografického hľadiska najviac zmlúv bolo uzavretých v Bratislavskom kraji – takmer 18% zmlúv, za ním nasleduje Banskobystrický kraj, kde bolo uzavretých 15% zmlúv. Najmenej zmlúv

bolo uzatvorených v Trenčianskom kraji. Filter Discretize nám opäť rozdelil vek poistníka do 10 intervalov. Z výstupu sme zistili, že vo vekovej skupine od 28 do 36 rokov bolo uzatvorených 24% poistných zmlúv. 18% poistiek bolo uzavretých vo veku od 36 do 43 rokov. Zaujímavé je, že 0,75% zmlúv bolo uzavretých vo veku nad 80 rokov.

Najčastejším dôvodom ukončenia zmluvy bol odkup (78%). Takmer 10% zmlúv bolo zrušených do dvoch mesiacov od uzavretia poistnej zmluvy zo strany poistníka. Zmluvu ukončilo 48 žien a 35 mužov. Pri analýze pohlavia poisteného sme zistili, že u chlapcov bolo zrušených 46 zmlúv a u dievčat 37. Priemerný vek poistených pri zrušených zmluvách bol 6,9 roka. Z vekového hľadiska poistníkov najviac zrušených zmlúv bolo vo veku do 45 rokov - a to až 66%. Z toho 26% bolo zrušených vo veku od 35 do 40 rokov. Analýza zrušených zmlúv podľa krajov poistníka ukázala, že najviac zmlúv bolo zrušených v Žilinskom kraji – 16,87%, tesne za ním to boli kraje Banskobystrický a Bratislavský – 15,66%. V Nitrianskom kraji zrušilo poistku 7,23% poistníkov. Z okresov sa zmluvy za sledované obdobie najčastejšie rušili v okrese Bratislava a Košice (8,4%), okres Žilina 7,2% zmlúv a Veľký Krtíš 6%.

Produkt 2A

Poistenie 2A patrí do skupiny produktov investičného životného poistenia. V tomto type poistenia klient investuje časť finančných prostriedkov do podielových fondov, pričom môže ísť o konzervatívne, zmiešané alebo progresívne fondy a druhá časť je určená na zabezpečenie poistenia.

Z poskytnutých údajov sme pomocou záložky preprocess zistili, že v produkte 2A máme k dispozícii 93% aktívnych zmlúv a teda len 7% zrušených. Jednorázovo platené poistenie sa rozhodlo v sledovanom období uzatvoriť 56% žien. Veková štruktúra klientov ukázala, že najviac zmlúv bolo uzatvorených vo veku 48-54 rokov. Možnosť zníženia daňového základu o nezdaniiteľnú časť na životné poistenie v tomto produkte využilo až 43% klientov. Z geografického hľadiska bolo najviac zmlúv uzavretých v:

- Košický kraj – 299
- Banskobystrický kraj – 286

- Bratislavský kraj – 254
- Žilinský kraj – 183
- Trnavský kraj – 176
- Trenčiansky kraj – 167
- Nitriansky kraj – 125
- Prešovský kraj – 120

V dvoch prípadoch bol poistený členom iného štátu.

Preskúmaním zrušených zmlúv sme zistili, že skoro 70% zmlúv bolo ukončených z dôvodu odkupu, 14% zmlúv bolo ukončených poistníkom do dvoch mesiacov. Zmluvy v prípade pohlavia poisteného rušili rovnako muži aj ženy, u poistníkov viac zmlúv ukončili ženy (57%). K najväčšiemu počtu zrušených zmlúv došlo v Banskobystrickom kraji – až $\frac{1}{4}$ zmlúv, pričom 5% bolo zrušených v okrese Banská Bystrica. Na druhom mieste bol Košický kraj, kde klienti zrušili 18% zmlúv a 13% zo všetkých zrušených zmlúv bolo v okrese Košice.

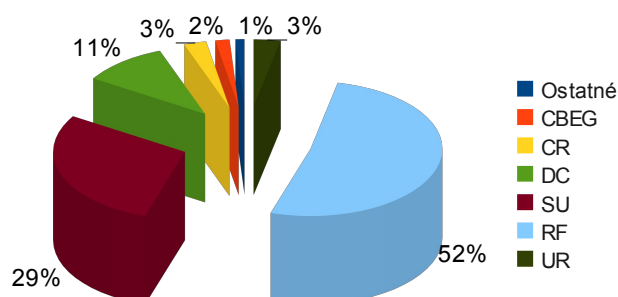
Produkt 2B

Produkt 2B je druhým poistením v skupine investičného poistenia, pričom poistné je platené bežne. Poistník má na výber z mesačného, štvrťročného, polročného alebo ročného spôsobu platenia. Toto poistenie je medzi klientmi poisťovne populárnejšie ako produkt 2A. Vo vzorke údajov máme k dispozícii viac ako 35 000 záznamov. Percento zrušených zmlúv v tomto prípade je 34%.

Poistnú zmluvu uzatvorilo viac mužov ako žien (59,5%). Priemerný vstupný vek poistených osôb bol 34,31rokov a smerodajná odchýlka 11,88, pričom 3% poistených klientov boli osoby mladšie ako 18 rokov. Mesačnú frekvenciu platenia poistného si zvolilo 98% zákazníkov. Rovnako ako v predchádzajúcich prípadoch aj pri tomto poistení najväčší počet zmlúv bol uzatvorený v Banskobystrickom a Košickom kraji (približne 16%). Naopak - najmenej poistných zmlúv bolo uzatvorených v Bratislavskom kraji – len 8%.

Ako sme už spomenuli - zrušených zmlúv z celkového počtu bolo 34%. Najčastejším spôsobom zrušenia zmluvy bolo nezaplatenie bežného následného

poistného, a to v 51% prípadov a 29% zrušených zmlúv zapríčinil odkup. Všetky spôsoby rušenia zmluvy pri tomto produkte znázorňuje nasledujúci graf.¹



Gráf 2: Spôsoby zrušenia zmlúv
Prameň: Vlastné spracovanie

Pri skúmaní zrušených zmlúv podľa krajov sme zistili, že najviac zrušených zmlúv bolo v Košickom a Banskobystrickom kraji a taktiež aj v Trenčianskom kraji. Z okresov to boli: Košice, Prievidza, Bratislava, Dunajská Streda, Trebišov, Lučenec.

Produkt 3A

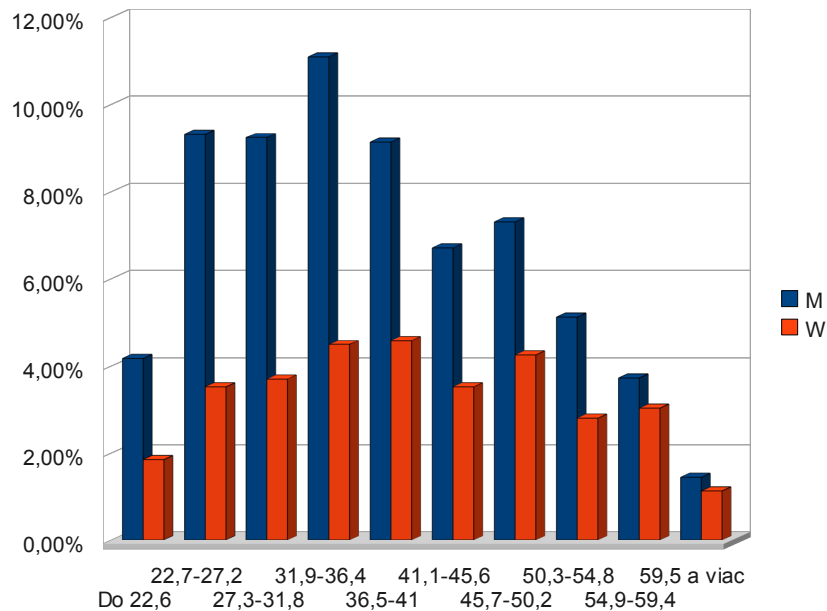
Posledným analyzovaným produktom je rizikovo orientované poistenie. Špecifikom tohto poistenia je, že sa uzatvára len na Poist'ovňou stanovenú poistnú sumu na hlavné krytie. Poist'ovňa ponúka dva varianty poistnej sumy, Teda poistná suma je vo všetkých prípadoch rovnaká. Ďalším špecifikom je, že poistná doba sa dojednáva len na prvé obdobie, pričom sa po zaplatení následného poistného toto obdobie predĺži. Klient si môže opäť vybrať z dvoch variantov, a to ročné alebo mesačné poistné.

Štruktúra údajov pri tomto produkte je nasledovná:

Databáza obsahovala skoro 10 000 záznamov, z toho aktívnych zmlúv je 63%. 89% klientov si vybralo mesačnú frekvenciu platenia poistného. Pri produkte 3B si

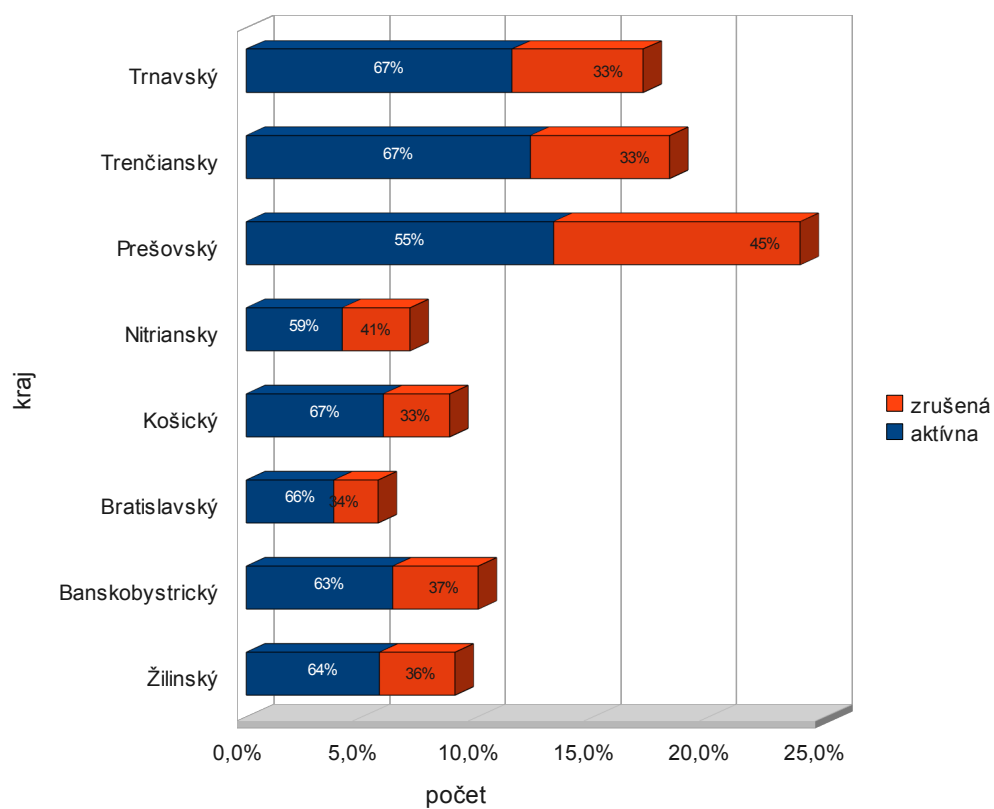
¹ RF-nezaplatenie následného bežného poistného, SU- odkup, CBEG- storno od počiatku, CR- zrušenie do 2m, DC-zrušenie do 2m. Pre nedodanie podkladov, UR-zrušenie do 2m, neovplyvniteľné

zmluvu uzatvorilo viac mužov ako žien (67% mužov). Priemerný vek poistených osôb je 38,4 roka, minimálny vek je 18 rokov a maximálny 65. Rozdelením veku do skupín pomocou filtra Discretize sme zistili, že najviac zmlúv bolo uzatvorených vo veku od 32 do 36 rokov (15,5%). Ostatné údaje sú zobrazené v grafe.



Garf 3: Veková štruktúra poistencov pri produkte 3A
Prameň: Vlastné spracovanie

Analýza zrušených zmlúv ukázala, že k zrušeniu z dôvodu dožitia sa konca poistného obdobia bolo ukončených až 88% zmlúv, 5% zmlúv bolo zrušených stornom od začiatku. Nasledujúci graf zobrazuje aktívne a zrušené zmluvy rozdelené podľa krajov.



Garf 4: Analýza aktívnych a zrušených zmlúv podľa krajov
Prameň: Vlastné spracovanie

2.2 Tvorba dataminingových modelov

V predchádzajúcich častiach sme sa venovali charakteristike Poist'ovne a jej produktov, ďalej sme opísali štruktúru vstupných údajov. Teraz sa budeme venovať samotnej tvorbe dataminingových modelov, pričom budeme vychádzať z pôvodnej analýzy Ing. Hromadu a budeme modelovať tie modely, ktoré v pôvodnej analýze boli vyhodnotené ako najlepšie.

Produkt 1A

Aj pri tvorbe modelu v prípade tohto produktu sme sa snažili čo najviac priblížiť modelom urobeným v konkurenčnom nástroji SAS EM.

Celý data set bol pôvodne rozdelený na tri množiny: tréningovú, validačnú a testovaciu. Weka však nemá nástroj na rozdelenie data setu na tri subdata sety. Celý súbor dát sme teda rozdelili na dve množiny – tréningovú a validačnú v pomere 60:40. Vzhľadom na to, že Weka vyhodnotila cieľovú premennú ako numerickú museli sme ju transformovať na nominálnu hodnotu pomocou filtra Numeric to nominal. Následne sme odstránili atribúty, ktoré neboli vhodné pre model rozhodovacieho stromu, ktorý sa javil ako najoptimálnejší. Pri tvorbe rozhodovacieho stromu sme využili algoritmus J48 (C4.5) s faktorom spoľahlivosti 0,05. Z výstupu vyplýva, že sme dosiahli rovnaké výsledky ako boli v SAS EM. Nasledujúca tabuľka porovnáva vybrané ukazovatele.

	WEKA		SAS EM	
	tréningová m.	validačná m.	tréningová m.	validačná m.
ROC index	0,67	0,66	0,68	0,67
RMSE	0,47	0,47	0,47	0,48

Tabuľka 1: Vybrané ukazovatele pri produkte 1B

Prameň: Vlastné spracovanie

Produkt 1B

Pri produkte 1B sme disponovali databázou údajov, v ktorej bolo len 1 300 záznamov, preto pre tvorbu modelu sme túto databázu rozdelili na dve časti. Tréningová množina obsahovala 60% záznamov a validačná zvyšných 40%. Vo Weke rozdelenie data setu umožňuje filter Remove percentage, ktorý nájdeme v záložke

preprocess. Premenné, ktoré najviac vplyvajú na cieľovú premennú (teda to, či je zmluva zrušená) sme prevzali z pôvodnej analýzy, čím sme zabezpečili porovnateľné výsledky. Vstupnými premennými, ktoré vchádzali do modelu boli:

- kraj poistníka
- okres poistníka
- pohlavie poistníka
- vstupný vek poistníka
- technická úroková miera
- a iné logaritmované premenné.

Skúmanie informácií o poistníkovi je pri tomto poistení nevyhnutné vzhľadom na to, že poistník má právo disponovať s poistnou zmluvou.

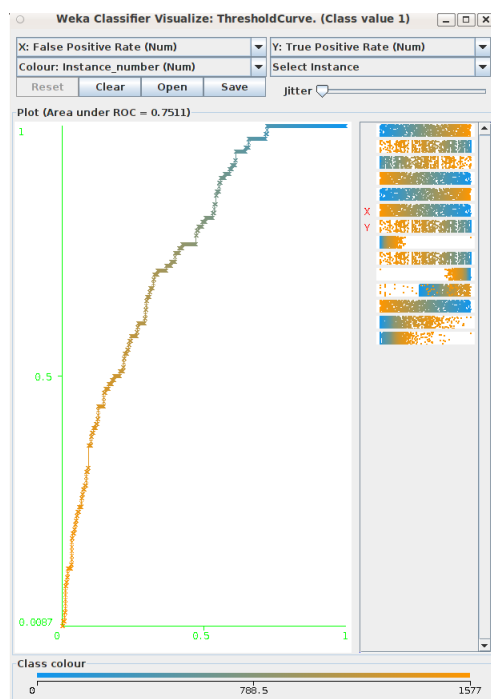
V pôvodnej analýze bol ako najvhodnejší model označený model logistickej regresie. V záložke Classify sme si preto aj my zvolili klasifikátor logistickú regresiu. Hodnoty ROC indexu a RMSE na tréningovej množine boli porovnateľné s pôvodnou analýzou. ROC krivka je znázornená v prílohe 2.

Keďže databáza obsahovala málo záznamov a je potrebné overiť model pri väčšom počte poistných zmlúv, tento model nebudeme ďalej hodnotiť.

Produkt 2A

Po rozdelení datasetu na tréningovú a validačnú množinu sme zistili, že validačná množina neobsahuje žiadnu zrušenú zmluvu. Z toho dôvodu sme pri tvorbe modelov pracovali s celou vzorkou údajov. Cieľovú premennú v tomto prípade najviac ovplyvňujú anualizovaná hodnota, okres, nezamestnanosť v kraji, vstupný vek poisteného a aj poistníka. Analýza, ktorá bola urobená v SAS EM ukázala, že najvhodnejším modelom je rozhodovací strom. Po vytvorení rozhodovacieho stromu využitím algoritmu J48 a nastavením faktoru spoľahlivosti na 0,05 sme zistili, že veľkosť rozhodovacieho stromu bola len jedna vrstva, čo potvrdilo aj výsledok pôvodného modelovania. Ďalším modelom, ktorý sme tvorili bola logistická regresia. Hodnota ROC indexu sa síce líšila o 8%, tento rozdiel mohol byť spôsobený tým, že databáza nebola rozdelená na tréningovú a validačnú množinu. Avšak pri analýze v

SAS EM došlo k rôznym výsledkom ROC indexu pri tréningovej a validačnej množine. Hodnota RMSE bola takmer totožná. Obrázok zobrazuje ROC krivku pre produkt 2A.



Obrázok 3: ROC krivka pre produkt 2A

Prameň: Výstup z aplikácie WEKA

Táto analýza však nemá vypovedaciu hodnotu práve pre nízky počet zrušených poisťných zmlúv. Tvorbu modelu a jeho overenie by bolo vhodné zopakovať pri väčšej databáze. Avšak výsledky môžu ovplyvniť aj iné faktory (nové legislatívne podmienky, vývoj trhu...). Vtedy bude nutné vytvoriť úplne nový model.

Produkt 2B

Pri tvorbe modelu pre poistenie 2B sme začali transformáciou cieľovej premennej z numerickej na nominálnu hodnotu. Následne sme využili filter Normalize pre znormovanie numerických premenných. Vybrali sme premenné, ktoré podľa predchádzajúcej analýzy vstupovali do modelov. Využitím filtra AddExpression sme vytvorili nové atribúty, ktoré zlogaritmovali pôvodné premenné. V záložke Classify sme si vybrali opäť funkciu - logistickú regresiu, ktorá bola označená ako najvhodnejšia pre tvorbu modelu. Dáta sme rozdelili pomocou test option- percentage split na tréningovú a validačnú množinu v pomere 60:40. Výsledok modelovania však

dosahoval iné hodnoty ROC indexu ako v predchádzajúcich modeloch. Rozdiel bol až 13%. Hodnota zle klasifikovaných prípadov bola tiež oveľa vyššia.

Dôvodov pre rôzne výsledky je pravdepodobne viac. Jedným z dôvodov môže byť využitie filtra Standardize, ktorý štandardizuje všetky numerické premenné nielen niektoré, ako to bolo v prípade pôvodného modelu. Rovnaké výsledky sme však dosiahli aj po znormovaní premenných. Ďalším dôvodom môže byť aj to, že dataset sme rozdelili len na dve množiny a nie na tri ako to bolo v predchádzajúcom modeli.

Produkt 3A

Tak ako aj pri predchádzajúcich produktoch aj tu sme museli v prvom rade transformovať cieľovú premennú na nominálnu. Pomocou filtra Discretize sme jednotlivé premenné rozdelili do intervalov a premenné sme znormovali. V predchádzajúcich analýzach ako významné premenné boli určené: doba trvania poistnej zmluvy, pohlavie poistníka, vstupný vek. Z geografických faktorov to sú kraj a okres poistníka. Taktiež sme vytvorili aj nové premenné zlogaritmovaním pôvodných. Následne sme spustili tvorbu modelu s využitím funkcie Logistic, teda logistickú regresiu. Tak ako v predchádzajúcom prípade, aj tu sme dosiahli rozličné výsledky. Hodnota ROC indexu bola 0,76 - čo je o 0,07 menej ako v už vytvorenom modeli. Hodnota RMSE bola o niečo vyššia. Nami dosiahnutá hodnota bola 0,24 oproti pôvodnej 0,19. Graf ROC indexu je zobrazený v prílohe č. 3.

Záver

Datamining sa stáva veľmi zaujímavou oblasťou, ktorá sa pomaly dostáva do povedomia spoločnosti. Pre úspešnú tvorbu dataminingových modelov je nevyhnutné nájsť aj vhodný softvérový nástroj, ktorý by uľahčoval tvorbu modelov. Na druhej strane je potrebné, aby náklady na takýto nástroj boli pre spoločnosť akceptovateľné.

Cieľom tejto práce bola portácia prediktívnych modelov medzi dvoma programovými prostrediami - z komerčného prostredia programovej sady SAS do cenovo výhodnejšieho prostredia voľne šíriteľného programového vybavenia WEKA a analytických nástrojov, ktoré poskytuje.

Najskôr sme sa zamerali na štruktúru dát, kde WEKA poskytla plnohodnotné pracovné prostredie na spracovanie a analýzu vstupných údajov. Dokázali sme získať takmer rovnaké informácie ako pri použití uzla Stat Explore, ktorý je súčasťou programu SAS Enterprise Miner. Pre grafické spracovanie údajov sme použili tabuľkový kalkulátor Calc z kancelárskeho balíka OpenOffice.org, ktorý je taktiež voľne šíriteľný. Weka poskytuje aj grafické spracovanie údajov, avšak nie je možné ich uložiť ako výstupnú zostavu. Pri samotnej tvorbe modelov sme dospeli k menším rozdielom pri výsledkoch, čo mohlo byť spôsobené použitím iných nástrojov na úpravu údajov. Nevýhodou je, že WEKA neposkytuje možnosť skúmať charakteristiky ako sú šikmosť a špicatosť na vstupných dátach. Ďalšou nevýhodou je, že dataset dokáže rozdeliť len na dve množiny. Problém môže nastať pri modelovaní s väčšou databázou, pretože Weka automaticky načíta údaje do pamäte.

Aj napriek nevýhodám a menším problémom, s ktorými sme sa stretli, považujeme tento nástroj za vhodnú alternatívu dataminingového nástroja SAS EM. Odporúčame ju využívať malým podnikom, ktoré majú na jednej strane menšiu databázu údajov a na druhej strane nedisponujú dostatočnými finančnými prostriedkami na zakúpenie licencie SAS. Cieľ práce považujeme za splnený.

Zoznam bibliografických odkazov

CAO, L., et al. 2009. Data Mining for Business Applications. New York: Springer Science+Business Media, LLC, 2009. 302 p. ISBN: 978-0-387-79419-8

FERNANDEZ, G. 2003. Data Mining Using SAS Applications. [s.l.] Chapman & Hall/CRC, 2003. 349 p. ISBN 1-58488-345-6

GUO, L. [online]. 2003. Applying Data Mining Techniques in Property-Casualty Insurance . 2003. [cit. 2011-01-21]. Dostupné na internete:
<http://www.casact.org/dare/index.cfm?abstrID=5263&fuseaction=view>

HAN, J., KAMBER, M. 2006. Data Mining: Concepts and Techniques. Second Edition. San Francisco: Morgan Kaufmann Publishers. 2006. 745 p. ISBN 13: 978-1-55860-901-3

CHIU, S., TAVELLA, D. 2008. Data Mining and Market Intelligence for Optimal Marketing Returns . First Edition. [s.l.] Elsevier Inc.. 2008. 269 p. ISBN: 978-0-7506-8234-3

OLSON, D., L.DELEN, D. 2008. Advanced Data Mining Techniques. [s.l.] Springer-Verlag. 2008 177 p. ISBN: 978-3-540-76916-3

PAL, R. N., JAIN, L. 2005. Advanced Techniques in Knowledge Discovery and Data Mining. London: Springer-Verlag. 2005. 253 p. ISBN 1852338679

PARALIČ, J. 2003. Objavovanie znalostí v databázach . Elfa: Košice. 2003. 88 s. ISBN 80-89066-60-7

PARR RUD, O. 2001. Data Mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computer Press. 2001. 322 s. ISBN 80-7226-577-6

TWO CROWS CORPORATION. 2005. Introduction to Data Mining and Knowledge Discovery. Third Edition. 2005. 36 p. ISBN: 1-892095-02-5

WITTEN, H.I. FRANK, E. 2005. Data Mining : Practical Machine Learning Tools and Techniques. Second Edition. [s.l.] Morgan Kaufmann Publishers. 2005. 525 p. ISBN: 0-12-088407-0.

SAS Institute Inc.. 2010 Data mining: Hĺbkový prieskum údajov. [online]. [cit. 2010-12-25]. Dostupné na internete:
<http://www.sas.com/offices/europe/slovakia/solutions/crm/dmining.html>

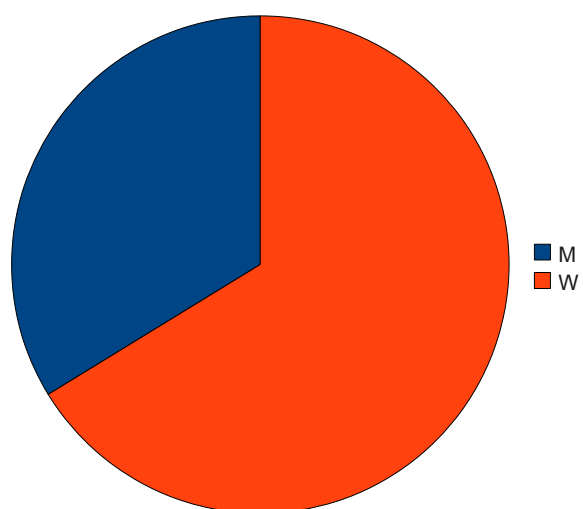
Prílohy

Príloha 1: Pohlavie poistníkov pri produkte 1B

Príloha 2: ROC krivka pre produkt 1B

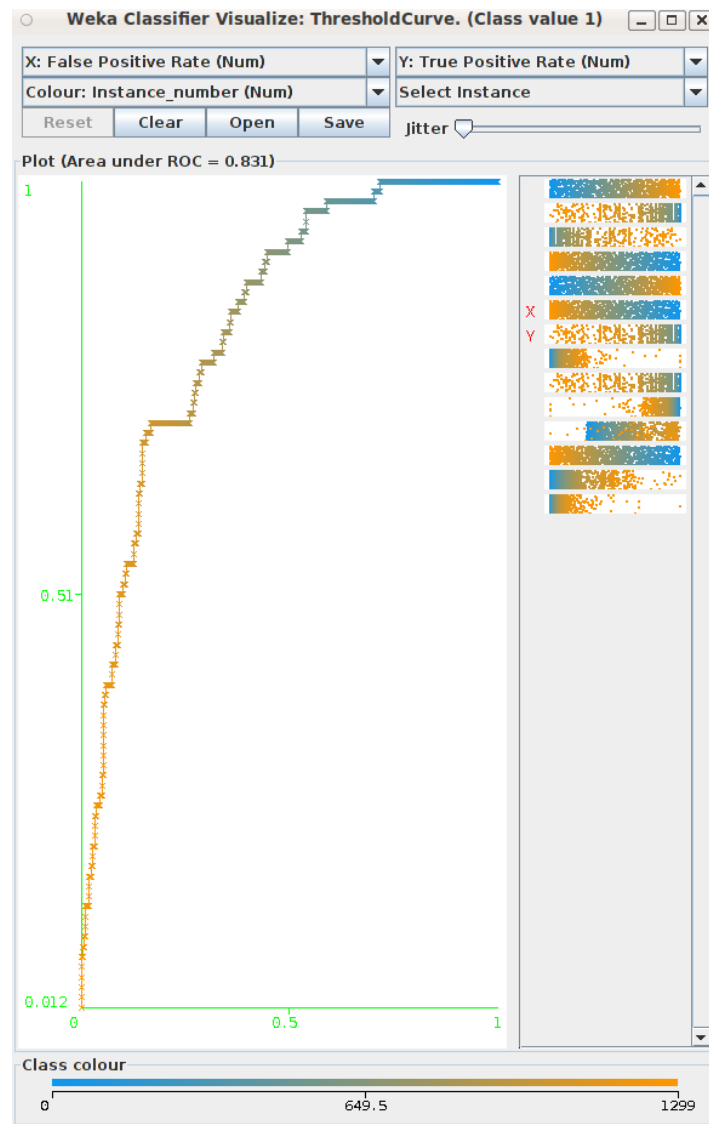
Príloha 3: ROC krivka pre produkt 3A

Príloha1: Pohlavie poistníkov pri produkte 1B



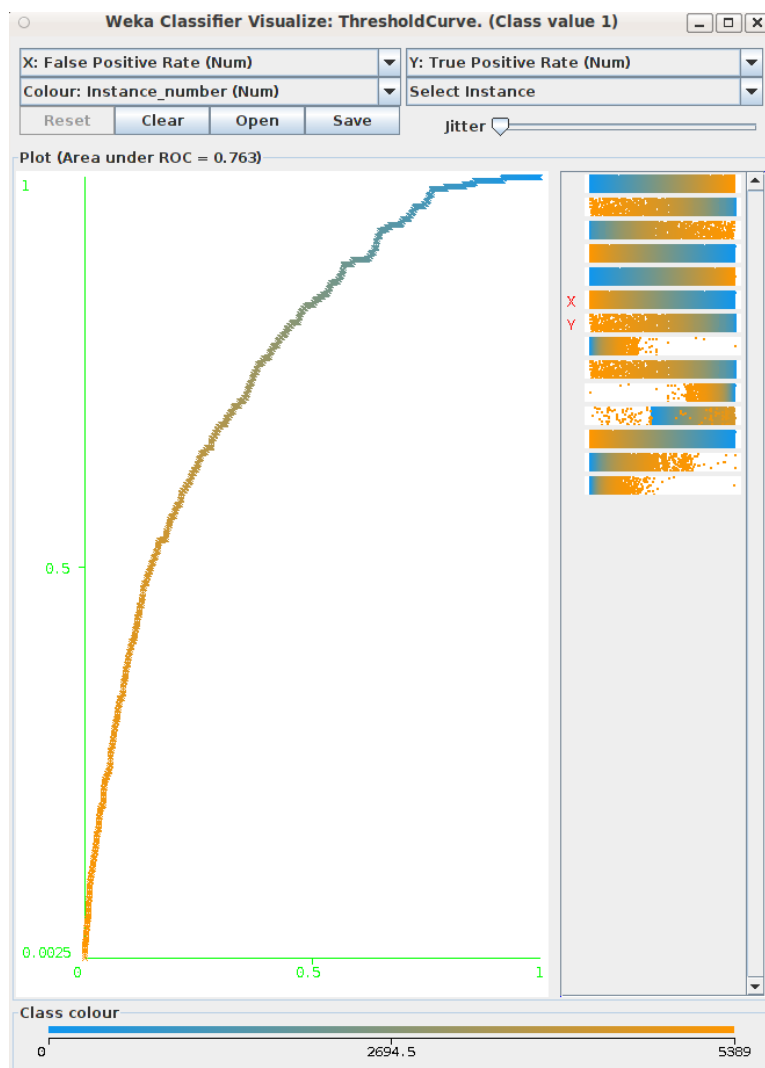
Prameň: Vlastné spracovanie

Príloha 2: ROC krivka pre produkt 1B



Prameň: Výstup z aplikácie WEKA

Príloha 3: ROC krivka pre produkt 3A



Prameň: Výstup z aplikácie WEKA