Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informatiky a kvantitativních metod

Data mining: Analýza klientů retailového bankovnictví

Použití statistických metod na údaje o chování bankovních klientů pro analýzu a klasifikaci

Diplomová práce

Autor: Norbert Snášel
Studijní obor: Informační management-magisterský navazující

Vedoucí práce: Mgr. Jan Draessler, Ph.D.

Hradec Králové Duben 2016

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a s použitím uvedené literatury.

V Hradci Králové dne 25.4.2016 Norbert Snášel

Poděkování:

Děkuji vedoucímu diplomové práce Mgr. Janu Draesslerovi, Ph.D. za metodické vedení práce, pomoc a vřelé rady pro napsání mé diplomové práce.

Anotace

Diplomová práce se zabývá analýzou klientů retailového bankovnictví v České republice. První část popisuje přípravu dat získaných z internetového kalkulátoru bankovních poplatků a použité statistické metody. Zmiňuje se také o použitém software SPSS Statistics a SPSS Modeler a jejich možnostech. Druhá část zpracovává dvě klasifikační data miningové úlohy. Diplomová práce ověřuje stanovenou hypotézu o využívání okrajových bankovních služeb a pokouší se nalézt kvalitní klasifikační modely. V první části nabízí rady pro lepší a snáze zpracovatelnější formy internetových kalkulátorů. Čtenář získá teoretické a praktické informace o shlukové a diskriminační analýze a jejich aplikaci na bankovní data. V závěru práce jsou shrnuty použité postupy a získané výsledky. Cílem je použití vícerozměrných statistických metod k získání cenných informací o bankovních klientech.

Annotation

Title: Data mining: Analysis of retail banking clients

Diploma thesis analyses clients of retail banking in the Czech Republic. First part describes preparation of data gained from online banking charges calculator and used statistical methods. It also mentions used software SPSS Statistics and SPSS Modeler and their possibilities. Second part processes two classification data mining tasks. Diploma thesis verifies determined hypothesis about using infrequent banking services and tries to find quality classification models. In first part offer advice for better and easily utilized forms of internet calculators. Reader will acquire theoretical and practical information about cluster and discriminant analysis and their application on banking data. In the end are summarized used approaches and obtained results. The objective is to use multidimensional statistical methods for obtaining valuable information about banking clients.

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Použití statistických potažmo data miningových nástrojů je dnes nutností pro střední a velké podniky, které se chtějí udržet na vrcholu svého trhu. Ze všech oblastí, ve kterých se data miningu využívá, je nejčastější použití v komerční sféře. Bankovnictví je tedy logicky oblastí s častým využitím data miningu. Výhodou je, že na rozdíl od jiných finančních institucí, disponují bankovní společnosti detailními a rozsáhlými daty o svých klientech. Jedná se o informace nejen bankovního, ale také geografického a demografického charakteru. Hlavní otázkou, kterou se tato práce zabývá, zda je možné z dat získaných internetovým kalkulátorem bankovních poplatků analyzovat chování klientů bank, případně lze-li tato data použít pro vytvoření klasifikačních modelů a nalézt nejlepší možný model pomocí nejvhodnějších statistických metod.

Cílem práce je zanalyzovat získaná data o klientech retailového bankovnictví v České republice. Prvním krokem je zpracování původních dat, exportovaných z databáze do tabulek Excelu, a převedení do analyzovatelného datasetu použitelného v SPSS Modeler. Dalším krokem je výběr vhodných statistických a data miningových metod, popis jejich charakteristik, předpokladů, výhod a nevýhod. Třetím krokem je použití vybraných metod pro získání užitečných informací o bankovních klientech. Konkrétně jde o analýzu skupin klientů, ověření stanovené hypotézy, vytvoření modelu pro predikci využívání okrajových služeb a interpretace získaných výsledků. Posledním krokem je návrh možností využití výsledků pro management bank. Cílem práce je také návrh úprav internetového kalkulátoru, pro zvýšení kvality získaných dat po technické a logické stránce.

2 Data mining v business sféře

Data mining se objevuje jako samostatné odvětví na začátku devadesátých let. Hlavním důvodem bylo navýšení obsahu skladovaných dat. Původně se data mining zabýval pouze strukturovanými daty. V dnešní době se také využívá k analýzám textů nebo webového prostředí. Dobývání znalostí z databází lze definovat jako komplexní získávání neznámých, na první pohled nezřejmých a případně užitečných informací z dat. Celkový proces nezahrnuje pouze aplikaci statistické analýzy, ale začíná výběrem, porozuměním, zhodnocením a přípravou dat. Pokračuje výběrem vhodné metody, její aplikací a následnou interpretací získaných výsledků. Mezi základní typy úloh patří klasifikace, predikce, deskripce a takzvané hledání nuggetů. V data miningové oblasti existují metodiky, které si kladou za cíl sjednotit postup data miningových úloh, především zefektivnit a zrychlit zpracování a snížit celkové náklady. Mezi nejznámější patří metodika 5A od firmy SPSS, metodika SEMMA od firmy SAS a metodika CRISP-DM. Pro kvalitní analýzu je důležité rozumět analyzovaným datům. Díky odborným znalostem lze provést kvalitní přípravu dat a jednodušeji interpretovat výsledky. Metody data miningu lze aplikovat v mnoha oblastech. Mezi nejčastější patří komerční sféra, věda, lékařství a průmysl. Celkový objem trhu s data mining softwarovými nástroji je odhadován na 10 miliard dolarů. [5] [10]

Významnou oblastí komerční sféry data miningu je bankovnictví. V bankovnictví lze data mining použít pro mnoho účelů, mezi nejčastější patří cílený marketing, management vztahu se zákazníky (CRM), analýzy nákupního košíku, detekce podvodů nebo odhad rizikovosti klienta při půjčce. Banky mají rozsáhlé datové sklady, které není možné běžnými prostředky analyzovat. Data také často obsahují informace demografického a geografického charakteru, jako jsou věk a pohlaví klientů nebo místa výběrů z bankomatů. S nástupem internetového bankovnictví došlo k dalšímu navýšení objemu dat. Bankovnictví je silně konkurenčním trhem, kde zákazníci mohou volně bez velkých postihů přecházet mezi společnostmi. V konečném důsledku banky využíváním výše zmíněných analýz mohou získávat nové zákazníky, udržet si stávající zákazníky, zlepšovat nabízené služby a šetřit výdaje. V prostředí českého bankovního trhu

byla například analyzována stabilita bankovnictví pomocí shlukové a diskriminační analýzy (Černohorská, Černohorský a Teplý, 2007) nebo provedena analýza trendů v oblasti uspokojování klientů bank (Belas, Cipovova a Demjan, 2014). [4] [8] [9] [21]

Častou otázkou v bankovní oblasti je, zda pro data mining použít outsourcing nebo insourcing. Outsourcing je pro společnost finančně výhodnější, problémem pak ale může být ztráta bankovního pohledu na data. Řešením je spoluúčast zaměstnance banky na analýze v roli experta na data. Pokud se společnost rozhodne pro insourcing data miningu, pak hlavní otázkou je do jaké míry. Možným řešením je vyčlenění skupiny pracovníků, kteří se místo svých běžných pracovních pozic dočasně přesunou. Dalším problémem je rozhodnutí o umístění pracovníků. Běžně lze vybrat IT oddělení nebo obchodní oddělení. Výhodou umístění do IT oddělení je přístup ke všem dostupným hardwarovým a softwarovým prostředkům, mohou ale chybět expertní znalosti. Mezi výhody umístění do obchodního oddělení patří dostupnost obchodních znalostí, naopak ale tato oddělení běžně nedisponují potřebným software. V praxi pak záleží zejména na velikosti banky, objemu dostupných dat, počtu klientů a analýz a dalších parametrech. [6] [21]

3 Popis a příprava dat

Příprava dat je důležitým a časově často nejnáročnějším krokem data miningové úlohy. Připravená data se jednodušeji analyzují a výstupy metod lze jednodušeji interpretovat. Příprava dat může zahrnovat úpravu strukturovaných dat, odvození nových proměnných, snížení počtu případů, proměnných a další. Častým úkonem je nahrazení chybějících hodnot. Pro přípravu dat byly použity programy Microsoft Excel 2013, SPSS Statistics 23 a SPSS Modeler 17. Výběr software vycházel z dostupnosti na fyzických a virtuálních učebnách FIM UHK, předchozích zkušenostech a dostupnosti dokumentace. [5]

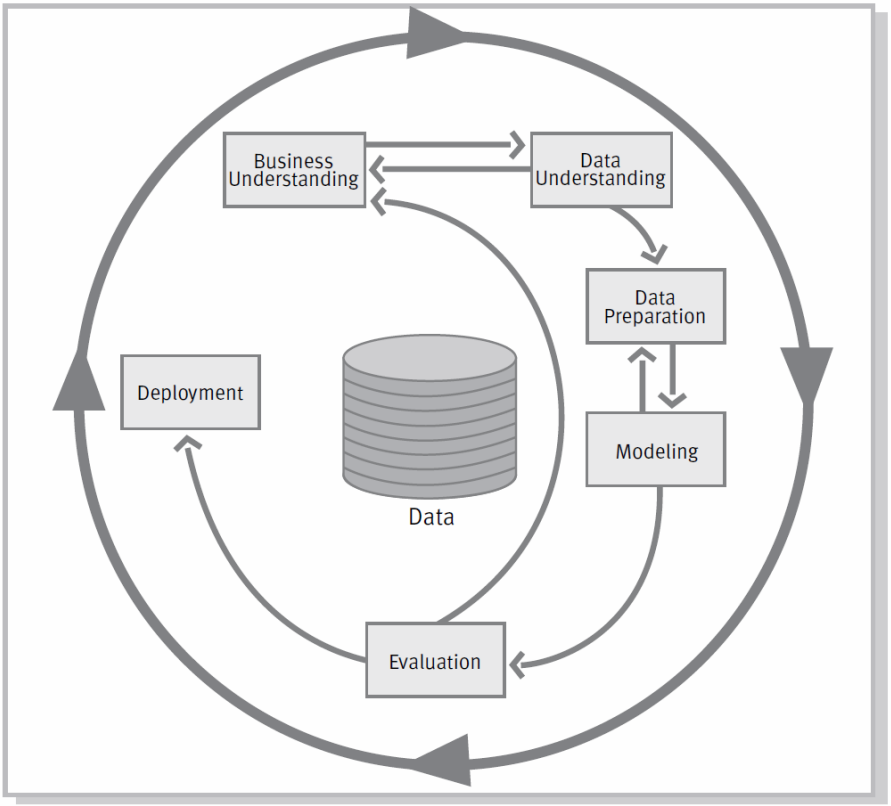
3.1 Původ dat

Data použitá pro tuto diplomovou práci pochází z kalkulátoru bankovních poplatků na portálu www.bankovnipoplatky.com. Portál se zabývá problematikou bankovních služeb a poplatků na českém trhu. Mezi jeho hlavní cíle patří informování bankovních klientů, zvyšování kvality vztahů mezi klienty a bankami a snižování bankovních poplatků. V rámci své činnosti portál vyhlašuje každoroční anketu o nejabsurdnější bankovní poplatek, hodnocení retailových bank a mimo jiné také poskytuje bankovním klientům kalkulátor, který srovnává ceny bankovních účtů dostupných v České republice. Kalkulátor je určen pro fyzické osoby a obsahuje celkem 29 otázek. Některé otázky obsahují další podotázky, uživatel tedy může zadat až 54 údajů + 4 povinné kontaktní údaje. Kvůli původu dat z internetového dotazníku lze očekávat určitý náklon směrem k uživatelům internetovému bankovnictví. Podobně také demograficky tato data odpovídají klientům nižšího věku s přístupem k internetu. Nakonec je třeba zmínit, že otázky kalkulátoru byly zvoleny právě s ohledem na porovnání cen bankovních účtů, takže některé oblasti využívání bankovních služeb jsou vynechány a jiné jsou rozebrány velmi podrobně. [3]

3.2 Metodika CRISP-DM

Postup zpracování práce odpovídá metodice CRISP-DM (Cross Industry Standard Process for Data Mining). Tato metodika zajišťuje ověřený rámec pro zpracování data miningové úlohy a je používána častěji než metodika SEMMA od firmy SAS. Historie metodiky CRISP-DM sahá do roku 1996. Projekt vznikal pod programem evropské unie ESPRIT a mezi hlavní členy patřilo pět společností, konkrétně SPSS, Teradata, Daimler AG, NCR Corporation a OHRA pojišťovna. První verze metodiky byla představena v roce 1999 a publikována v roce 2000. Metodika je také současně prosazována společností IBM, tvůrcem rodiny programů SPSS a byla implementována do SPSS Modeler. [5] [7] [14]

CRISP-DM rozděluje data miningové úlohy do šesti fází. Pořadí zpracování jednotlivých fází není pevně dáno, často se v průběhu úloh k jednotlivým fázím podle potřeby navrací. Celkový proces je zobrazen na obrázku č. 1. [7] [14]

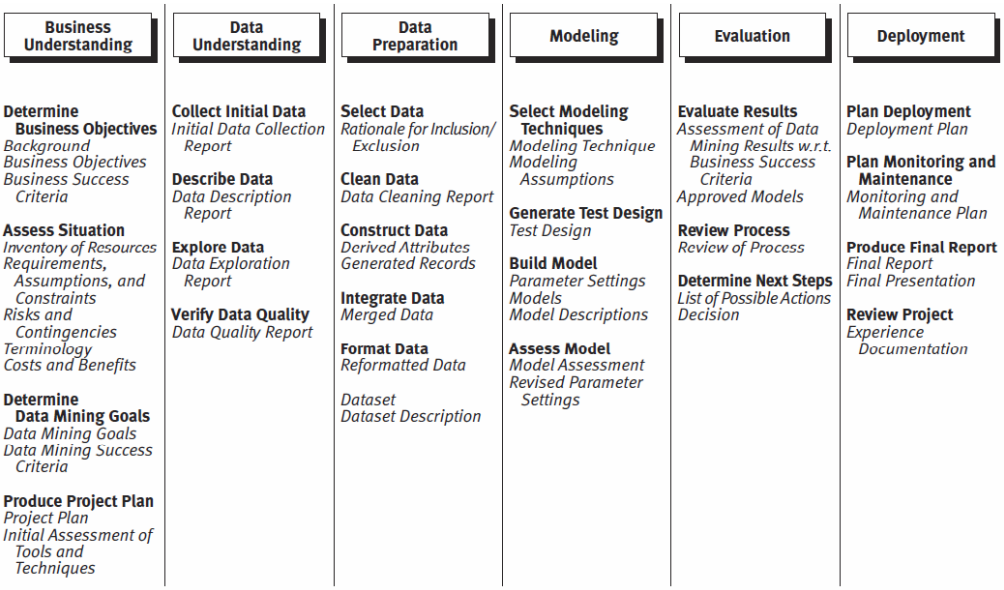


Obr. 1 Diagram metodiky CRISP-DM
Zdroj: Převzato z [7]

První fází je porozumění datům z hlediska řešené problematiky. Tato fáze se zaměřuje na formulaci data miningové úlohy, navrhnutí prvotního plánu úlohy, zhodnocení situace, pochopení cílů a požadavků manažerů (vychází se z předpokladu, že v praxi jsou data miningové úlohy obvykle provedeny na základě požadavku manažerů). Druhou fází je sběr a porozumění datům. Zde se nejprve sjednotí dostupná data a poté zkoumají popisné statistiky proměnných, vztahy mezi proměnnými a formulují se hypotézy nebo vybírají zajímavé podskupiny vyskytující se v datech. Nakonec se hodnotí kvalita získaných dat a případné nedostatky. Třetí fáze slouží k přípravě dat do výsledného datasetu, což zahrnuje například export dat z databáze nebo jiného datového skladu, pročištění, sloučení, výběr a vytvoření nových proměnných, použití transformací a nahrazení chybějících hodnot. [7] [14]

Čtvrtou fází je modelování. V této fázi se vybírají použité statistické metody a jejich parametry jsou optimálně nastaveny. Po prvotním modelování se často vrací zpět k předchozí fázi. Mezi obvyklé důvody patří odhalení nedostatků počáteční přípravy dat z výsledků prvotních analýz. Objevuje se potřeba náhrady chybějících hodnot, použití vhodnějších transformací proměnných nebo odstranění objevených odlehlých hodnot a případů. Pátou fází je vyhodnocení výsledků. Po modelování se vyberou nejkvalitnější získané modely. Tyto modely jsou zhodnoceny a validovány i z manažerského pohledu a případně vybrány pro použití v praxi. Získané modely samy o sobě laikům mnohé neřeknou, proto jsou získané znalosti často vizualizovány pomocí tabulek a grafů. [7] [14]

Poslední šestou fází je použití výsledků a modelů v praxi. Aplikace získaných modelů je pak často úkolem zadavatelů, v podobě manažerů firmy nebo externích zákazníků. Zákazník musí pochopit výstupy z data miningové úlohy tak, aby byl schopný efektivně využívat získané znalosti. Poslední fáze může také zahrnovat hardwarovou a softwarovou aplikaci nebo sepsání závěrečné zprávy. Jednotlivé fáze se skládají z dalších kroků popsaných na obrázku č. 2. [7] [14]



Obr. 2 Kroky jednotlivých fází metodiky CRISP–DM
Zdroj: Převzato z [7]

Nakonec je třeba zmínit, že pořadí fází neodpovídá pořadí v této práci právě z důvodu, že se k některým fázím bylo třeba vrátit. [7] [14]

3.3 Příprava dat

První částí je základní příprava dat po technické stránce do datasetu, který umožní analýzu ve vybraném software. Druhou částí je úprava dat, například pomocí transformací proměnných pro dosažení normality, vytvoření nových proměnných nebo nahrazení chybějících hodnot. Prvotní příprava dat byla provedena v Microsoft Excel a v SPSS Statistics. Druhá část byla provedena v SPSS Modeler. Tato kapitola odpovídá fázím jedna, dva a tři metodiky CRISP-DM. [5]

V prvotní části přípravy dat bylo provedeno přes dvacet dílčích kroků. Data byla poskytnuta ve třech souborech formátu Excel. Nejprve byly všechny soubory sjednoceny, výsledný soubor nakonec obsahoval zhruba 90 000 záznamů. Řádky prázdné, na první pohled chybné, s proměnnými obsahující písmeno *O* místo nul a řádky obsahující text v numerických proměnných byly odstraněny. Vyřazeny byly také řádky obsahující velké množství chyb způsobené exportem z databáze do Excel souborů. Otázky a podotázky kalkulátoru byly poté považovány za jednotlivé proměnné a jednotlivé záznamy o využití kalkulátoru

za jednotlivé případy. Následovala filtrace případů podle IP adres a Hash hodnot, které jsou kalkulátorem generovány pro odhalení duplicit nebo IP adres s podezřele vysokým počtem případů. Některé IP adresy se v datech objevovaly více než tisíckrát a pokaždé se stejnými hodnotami. Maximální počet případů z jedné IP adresy byl proto nastaven na 250 vzhledem k typické nejvyšší používané agregaci internetového připojení 1:50 a sběru dat po dobu pěti let. Pro zjednodušení bylo odstraněno zhruba sto případů z roku 2009 a stejný počet případů se záporným měsíčním zůstatkem. Protože je téma diplomové práce zaměřené na analýzu běžných fyzických bankovních klientů, byla data filtrována po stránce rozmezí hodnot jednotlivých proměnných tak, aby odpovídaly běžnému využití. Například byly odstraněny případy s hodnotou proměnné *Minimální měsíční obrat* vyšší než 1 000 000 Kč nebo případy s hodnotou proměnné *Výběr z ATM vlastní banky* vyšší než 30. Mezi další provedené úpravy patří sjednocení názvů bank a spolu s proměnnou *Frekvence výpisu* následné překódování názvů na numerické indexy. Překódování proměnné *Frekvence výpisu* je zobrazeno v tabulce č. 1.

Tabulka 1 Překódování proměnné Frekvence výpisu

Frekvence výpisu	Překódované indexy
Týdně	1
Měsíčně	2
Čtvrtletně	3
Ročně	4

Zdroj: Vlastní zpracování

Proměnná *Datum* byla zredukována pouze na rok odeslání. Konkrétní čas je v kontextu analýzy nevýznamný, protože z dat nelze jednotlivé případy přiřadit konkrétním uživatelům nebo jejich skupinám. Následně byly sjednoceny formáty buněk, které způsobovaly občasné zobrazení numerických hodnot jako datum nebo jako jiné nenumerické formáty. Pro snížení dimenze dat a kvůli zbytečné podrobnosti byly vybrané proměnné sloučeny. Jednalo se o proměnné ze skupin *Platby - jednorázové, Platby - trvalé příkazy (TPÚ)* a *Platby - povolení k inkasu*

(včetně SIPO). Spojení jednotlivých proměnných vycházelo z nahrazení původních proměnných, které rozlišovaly platby podle toho, zda byly zaslány do vlastní nebo cizí banky, jednou proměnnou obsahující součet plateb do vlastní a cizí banky. Lze předpokládat, že běžný uživatel nerozlišuje příkazy do vlastní a cizí banky a z pohledu analýzy je tento rozdíl zanedbatelný. V dalším dílčím kroku byly odstraněny proměnné *IP, Hash, Služby, které jste v seznamu nenašli* a *Pořadí*. Proměnné *IP, Hash* a *Pořadí* jsou z pohledu statistické analýzy bezvýznamné a proměnná *Služby, které jste v seznamu nenašli* je prakticky nezpracovatelná.

V dalším kroku byla data očištěna od nekonzistencí. Pro získání kvalitních a přesných výsledků je třeba, aby data neobsahovala logické nesprávnosti. Jde například o případy, kde uživatel uvedl, že nevybírá hotovost z bankomatů cizích bank, u následující otázky však uvedl, že z těchto bankomatů určitou částku vybírá. Takové případy byly z datasetu odstraněny stejně jako všechny další, ve kterých byly nalezeny podobné nekonzistence. Chybějící hodnoty u některých proměnných byly nahrazeny s ohledem na tendenci lidí vyplňovat kalkulátor co nejrychleji. Například pokud uživatel u otázky *Počet výběrů z ATM vlastní banky* vyplnil hodnotu větší než jedna a otázku *Počet výběrů z ATM cizí banky* nevyplnil lze logicky předpokládat, že kvůli rychlejšímu vyplnění kalkulátoru druhou otázku vynechal právě proto, že by odpověď byla nula.

Pro další úpravy byl použit program SPSS Statistics 23 protože obsahuje potřebné předdefinované funkce. Nejprve byly v záložce Variable View nastaveny parametry proměnných, konkrétně jejich zkratky, počet desetinných míst pro numerické proměnné, datové typy a další. U kategoriálních proměnných byly nastaveny popisky skupin, takže místo numerických hodnot se ve výstupech analýz rovnou zobrazují například jména bank. Z důvodu technických chyb a pokusů o sabotování kalkulátoru se v původních datech vyskytovaly duplicitní řádky, které byly odstraněny pomocí funkce pro detekci duplicit. Z datasetu byly vyloučeny případy, ve kterých bylo vyplněno méně než 15% proměnných. Pomocí funkce pro detekci podezřelých případů bylo odstraněno zhruba padesát případů. Po provedení zmíněných úprav byl dataset připraven pro analýzu. Z původních 90 000 případů nakonec zůstalo necelých 32 000. Z tabulky č. 2 je zřejmé, že případů, potažmo využití kalkulátoru rok od roku ubývá. Důvodem mohou být

níže zmíněné nedostatky a také fakt, že některé banky a portály na svých stránkách nabízí obdobné kalkulátory.

Tabulka 2 Počet využití kalkulátoru v jednotlivých rocích

	Frequency	Percent	Valid Percent	Cumulative Percent
2010	11030	33,5	33,5	33,5
2011	12084	36,7	36,7	70,1
2012	4973	15,1	15,1	85,2
2013	2945	8,9	8,9	94,1
2014	1932	5,9	5,9	100,0
Total	32964	100,0	100,0	

Zdroj: Vlastní zpracování

Další kroky přípravy dat byly provedeny v programu SPSS Modeler 17. Vzhledem k povaze dat a výběru analýz byly vytvořeny nové kategoriální a numerické proměnné. Použity byly uzly Derive, které slouží k výpočtům a tvorbě nových proměnných. Numerické proměnné *Uhrada.celkem, TPU.celkem*, a *SIPO.celkem* byly vypočteny součtem hodnot všech forem dané platby. Do datasetu byly přidány kategoriální proměnné *Banky podle velikosti aktiv, Banky podle počtu poboček* a *Banky podle počtu bankomatů*. Rozdělení bank do jednotlivých skupin bylo provedeno podle parametrů zobrazených v tabulce č. 3. Počty poboček a bankomatů byly získány z portálu www.kurzy.cz. [15] [16]

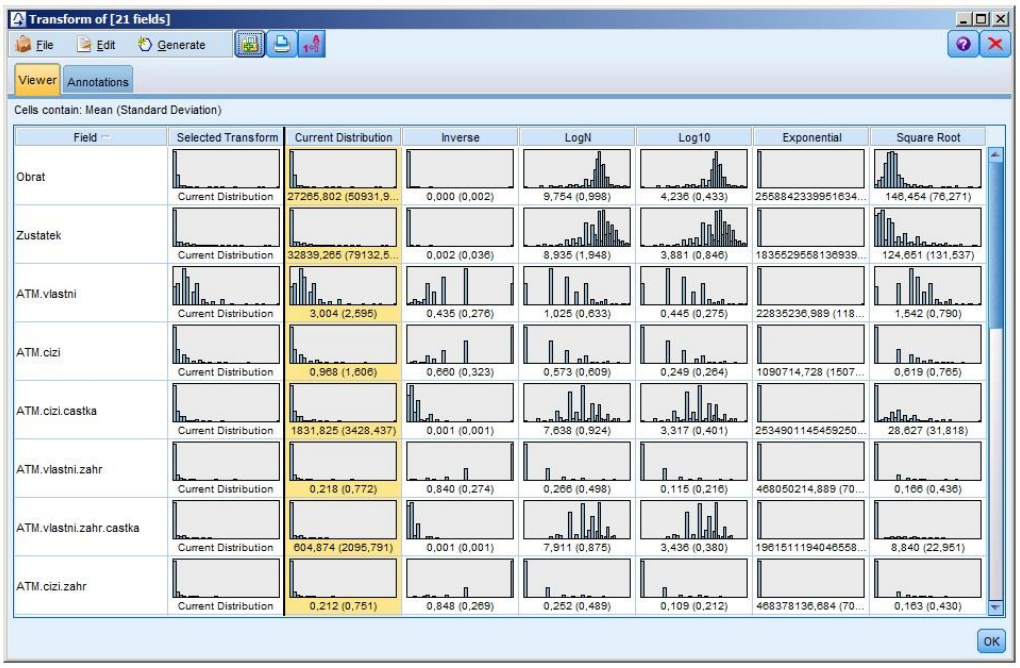
Tabulka 3 Tvorba nových proměnných

Proměnná/Hodnota	Banka.aktiva	Banka.podle.pobocek	Banka.podle.bankomatu
0	500+ mld. Kč	100+ poboček	100+ bankomatů
1	0 - 500 mld. Kč	10 - 99 poboček	1 - 99 bankomatů
2	X	0 - 9 poboček	0 bankomatů

Zdroj: Vlastní zpracování

Vybrané proměnné byly z analýz vyřazeny z důvodu vysoce odlišných apriorních pravděpodobností nebo nízké variace. Ze stejného důvodu byly některé numerické proměnné převedeny na dichotomické, které klienty rozdělují skupin podle toho, zda využívají danou službu. Jednalo se o proměnné *Cashback, Výběr na pobočce, Vklad na pobočce, Nadměrný vklad*, *Vklad přes bankomat* a *Výběr z bankomatu v zahraničí*.

Uzel Transform byl použit pro výběr optimálních transformací vybraných proměnných, tak aby zvýšily jejich normalitu a zároveň odstranily některé vysoké korelace. Uzel nabízí na výběr z pěti transformací, konkrétně inverzní, exponenciální, odmocninovou a dvě logaritmické (se základem logaritmu o hodnotě 10 a přirozeným logaritmem). Výstup zachycený na obrázku č. 3 zobrazuje histogramy transformovaných proměnných, jejich průměry a v závorkách směrodatné odchylky. Pro použitá data byly použity zejména logaritmické a odmocninové transformace. [15]

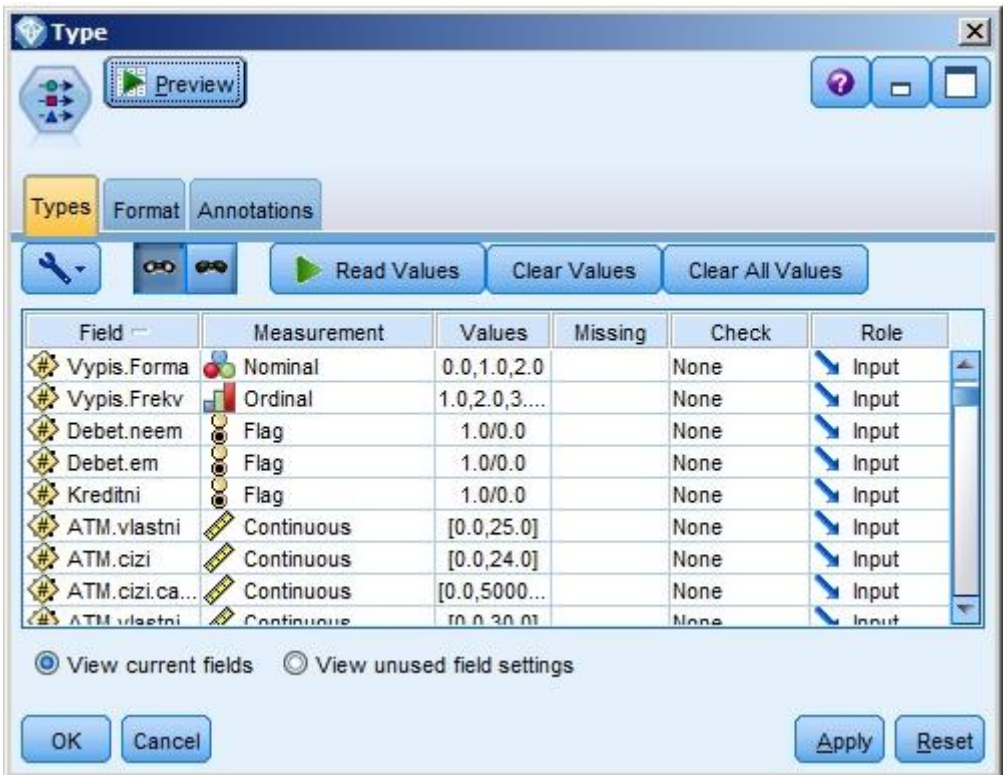


Obr. 3 Výstup uzlu Transform
Zdroj: Vlastní zpracování

Po prvotních analýzách došlo k situaci, kdy pro některé metody nebyl dostatek případů. Některé analýzy používají metodu Listwise, která do analýzy

zahrnuje pouze případy, které mají všechny hodnoty použitých nezávislých proměnných vyplněny. Proto byl použit uzel Filler, který umožňuje nahrazení chybějících hodnot. Chybějící hodnoty byly nahrazeny hodnotou nula, protože šlo o proměnné otázek, kde dochází k přeskočení vyplnění hodnoty nula kvůli tendenci lidí vyplnit dotazník co nejrychleji. [15]

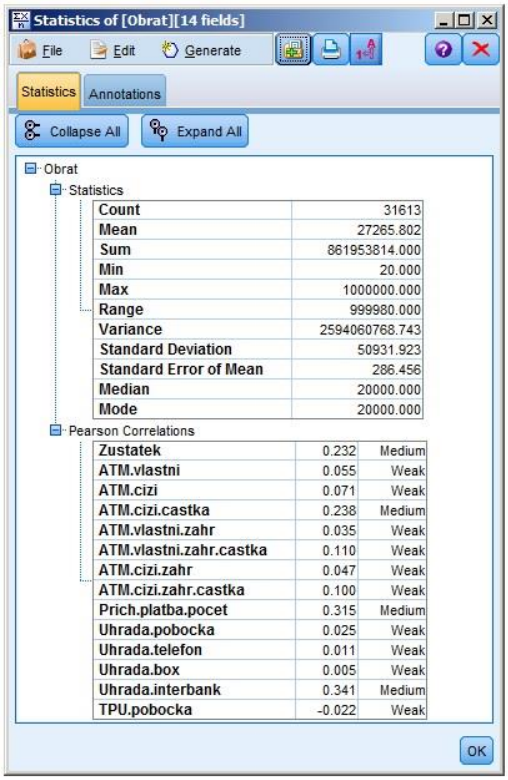
Pomocí uzlu Filter byly vyřazeny proměnné služeb, které naprostá většina klientů nikdy nepoužila a které byly vyplněny jen v několika případech. Jde o proměnné *Příjem mincí – kolikrát, Příjem mincí - Počet mincí jeden úkon, Výměna bankovek – kolikrát, Výměna bankovek - Počet bankovek jeden úkon, Příjem mincí - průměrná vkládaná částka* a *Výměna bankovek - průměrná vkládaná částka.* Nakonec byly v uzlu Type načteny hodnoty a nastaveny parametry proměnných jak lze vidět na obrázku č. 4. [15]



Obr. 4 Nastavení parametrů proměnných v uzlu Type
Zdroj: Vlastní zpracování

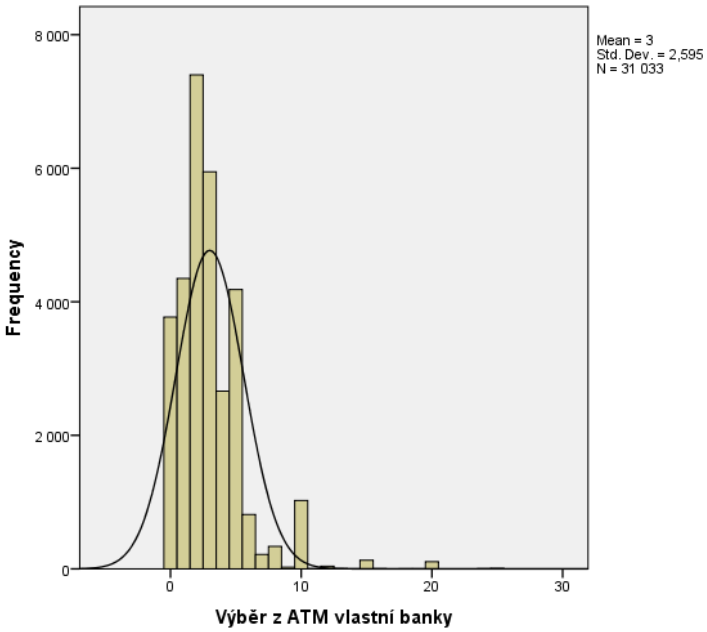
3.4 Explorace dat

Explorace dat je v práci rozdělena na analýzy jednotlivých proměnných (jednorozměrné) a na analýzy vztahů mezi více proměnnými (vícerozměrné). Pro jednotlivé numerické proměnné byly vygenerovány popisné statistiky jako počet, průměr, medián nebo šikmost. Ke zjištění popisných statistik a korelací mezi numerickými proměnnými byl použit uzel Statistics. Ten umožňuje nastavení mezních hodnot korelací tak, aby bylo možné vysoké korelace ve výstupu jednodušeji odhalit. Tento přístup je pohodlnější než hledání vysokých korelací v rozsáhlé korelační matici. Výstup uzlu je zobrazen na obrázku č. 5. [1] [15]



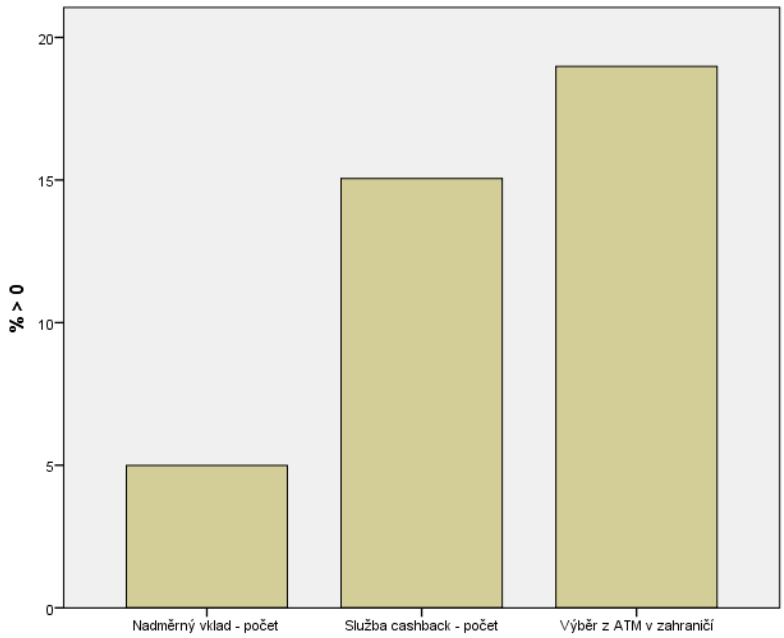
Obr. 5 Výstup uzlu Statistics
Zdroj: Vlastní zpracování

Dalším krokem je zobrazení histogramů, grafů, které zobrazují rozdělení četností hodnot proměnné. Histogram napoví, nakolik se rozdělení proměnné blíží normálnímu rozdělení, jak je zřejmé z obrázku č. 6. Pro jednodušší interpretaci lze graf proložit Gaussovou křivkou. Pro kategoriální proměnné byly vygenerovány tabulky a grafy četností. [1]



Obr. 6 Histogram vybrané proměnné
Zdroj: Vlastní zpracování

Na obrázku č. 7 jsou vidět procentuální četnosti využití okrajových služeb. Zobrazeny jsou procenta klientů, kteří na otázku odpověděli a službu použili alespoň jednou. Nejčastěji využívanou okrajovou službou je Výběr z bankomatu v zahraničí (vlastní nebo cizí banky).



Obr. 7 Frekvence využití okrajových služeb
Zdroj: Vlastní zpracování

Analýzy vztahů mezi více proměnnými lze rozdělit do tří kategorií. Jedná se analýzy vztahů dvou numerických proměnných, numerické a kategoriální proměnné nebo dvou kategoriálních proměnných. Pro kombinace dvou numerických proměnných byly použity korelační koeficienty a korelační diagramy. Pro analýzu vztahu dvou kategoriálních proměnných byl použit Chi-square test. Pro analýzu vztahů mezi numerickou a kategoriální proměnnou byla použita jednorozměrná analýza rozptylu. [1]

Mezi nejvýznamnější nalezené závislosti patří vztah proměnných *Přímé bankovnictví* a *Internetbanking*. Pokud má banka malé množství poboček, je vysoce pravděpodobné, že její klienti používají internetové bankovnictví. Tento vztah je také posilněn faktem, že data pochází z internetového kalkulátoru. Proto bylo rozhodnuto, že v analýzách bude používána pouze proměnná *Přímé bankovnictví*. Vysoce korelované jsou proměnné z kategorie plateb, zejména proměnné týkající se využití telebankingu a plateb na pobočkách bank. Vysoké korelace se vyskytují mezi proměnnými celkového počtu plateb a plateb přes internetbanking. Zde se opět projevuje vliv původu dat. Další významný vztah je mezi proměnnými výběr a vklad na účet na pobočce. Vysoké korelace se vyskytují mezi proměnnými týkající se výběrů z bankomatu v zahraničí. Ukazuje se, že klienti, kteří vybírají hotovost v zahraničí, nerozlišují mezi bankomaty vlastní a cizí banky. [1]

3.5 Rady pro úpravu kalkulátoru

Prvotní příprava dat do datasetu byla komplikována některými technickými nedostatky kalkulátoru a nekvalitním exportem dat z databáze do Excel souborů. Tyto chyby zkomplikovaly přípravu dat a bylo by vhodné je do budoucna odstranit. Hlavním problémem je, že údaje zadávané do kalkulátoru nejsou korektně validovány. Kalkulátor upozorňuje uživatele, aby hodnoty u vybraných otázek zadával číslovkou. V praxi je ale vždy vhodnější implementovat validaci zadávaných údajů nejen po stránce datového typu, ale často i rozmezí hodnot tak, aby například nebylo možné zadat do počtu výběru z bankomatu zápornou nebo nelogicky vysokou hodnotu. Před použitím kalkulátoru je vhodné uživatele upozornit, že nevyplněné hodnoty u otázek jsou kalkulátorem zaznamenány jako

chybějící hodnoty a ne hodnota nula. Problém je častý u otázek, ve kterých uživatel zadává počet trvalých příkazů přes internetbanking, po telefonu a na pobočce. Uživatel zadá počet příkazů přes internetbanking a následující otázky ponechá nevyplněné. V takových případech lze předpokládat, že uživatel přeskočil vyplnění dalších dvou otázek, protože tyto služby nevyužívá. Řešením je u takových otázek předvýběr výchozí hodnoty nula. Uživatel poté hodnotu může změnit, tímto lze dosáhnout snížení počtu chybějících hodnot. Problémem může být také skutečnost, že klient některou službu využívá velmi zřídka a údaj proto vynechá. Dalším velkým problémem je výskyt nekonzistencí v datech. Případy ve kterých uživatel zadá, že nevlastní kreditní nebo debetní kartu a poté uvede, že vybírá hotovost z bankomatu, se v datech vyskytují velmi často. Tento problém by bylo možné řešit zpřístupněním navazujících otázek až poté, co na předchozí otázku uživatel odpoví kladně nebo vyplní hodnotu větší než nula. Často se v datech objevují duplicitní řádky způsobené dvojitým odesláním kalkulátoru nebo chybou při exportu databáze do Excel souborů.

Výše zmíněné problémy s opakováním případů se stejnými IP adresami a stejnými nebo velice podobnými hodnotami proměnných jsou způsobeny částečně náhodně a částečně se dají považovat za pokus o sabotáž dostupnosti kalkulátoru, sbíraných dat nebo z důvodu testování vlastních programátorských dovedností. Z těchto důvodů je vhodné implementovat kvalitnější ochrany před takovými útoky. Je sice pravda že běžného uživatele tyto sabotáže kromě útoků snižující dostupnost kalkulátoru nijak nepoškozují, ale v případě následné analýzy dat dochází ke zdlouhavé přípravě. Původní data obsahovala názvy bank zaznamenané několika způsoby. Doporučil bych zvolit stejný způsob jako u typů účtů, kdy je do databáze ukládána numerická hodnota a zobrazení jednotlivých názvů je řešeno na aplikační úrovni.

Případný dotazník lze také rozšířit o demografické otázky. V kontextu kalkulátoru bankovních poplatků tyto údaje nejsou třeba, v obecnějším dotazníku však mohou v následné analýze přinést užitečné informace. Mezi možné otázky patří pohlaví nebo věková skupina (často vhodnější než konkrétní věk, lidé tak na otázku odpoví častěji a ulehčí se příprava dat).

4 Popis a výběr analýz

Pro zpracování vybraných úloh byly použity dvě vícerozměrné statistické metody, konkrétně shluková a diskriminační analýza. Volba metod vycházela z dostupnosti v použitém statistickém programu, povahy a kvality použitých dat, jejich dimenzi, typu proměnných a počtu případů. SPSS Modeler nabízí širokou škálu uzlů analýz, v případě shlukové analýzy například nabízí tři odlišné metody. V této kapitole jsou teoreticky popsány použité analýzy. Teoretické pochopení principů metod je klíčové pro správnou interpretaci výsledků. [5] [15]

4.1 Shluková analýza

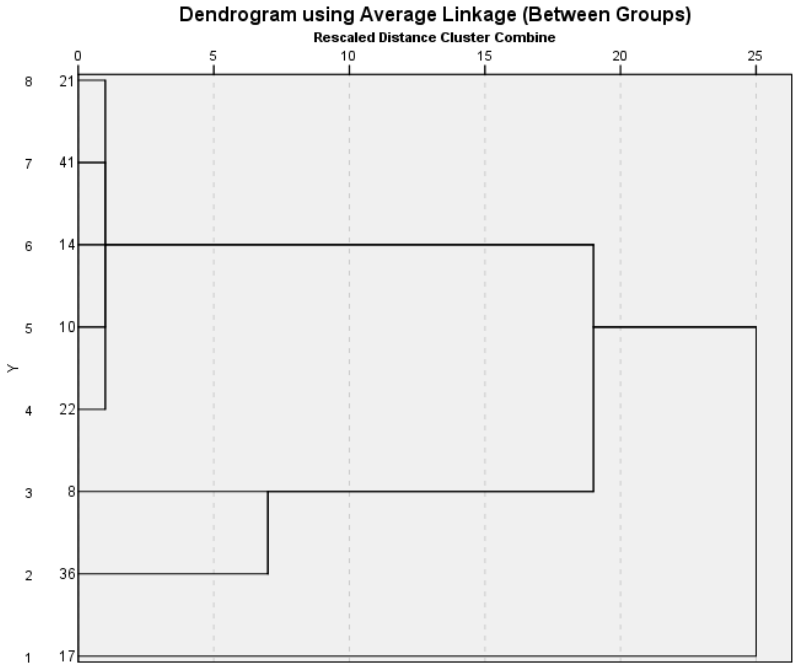
Shluková analýza slouží k nalezení přirozeně se vyskytujících skupin, zvaných shluky, ve kterých jsou zahrnuty případy navzájem velmi podobné a velice se lišící od případů z ostatních skupin. Tyto skupiny lze poté charakterizovat a odlišit od ostatních. Nalezené shluky lze dále použít jako podskupiny pro použití v dalších statistických metodách. Používá se v mnoha oblastech, především biologii, medicíně a ekonomii. Mezi využití shlukové analýzy patří například zkoumání skupin klientů internetového bankovnictví (Mann a Sahni, 2012). Předpokladem je možnost měřit vzdálenosti mezi jednotlivými případy. Mezi další hlavní předpoklady patří nezávislost zvolených proměnných, normální rozdělení numerických proměnných a multinomické rozdělení kategoriálních proměnných. Z podstaty metody plyne, že nepoužívá závislou proměnnou. Kvalitní model obsahuje malý počet dobře interpretovatelných shluků. [5] [13] [18] [19]

Mezi základní způsoby shlukování patří hierarchické a nehierarchické metody. Hierarchické shlukování funguje na principu rekurzivního spojování shluků, dokud nevznikne pouze jeden shluk obsahující všechny případy. Proces začíná definováním počátečního shluku pro každý malý shluk vzniklý v prvním kroku. Poté jsou všechny shluky porovnány a pár shluků s nejmenší vzdáleností je spojen do jednoho. Tento postup se opakuje, dokud nejsou všechny shluky spojeny do jediného (jedná se o podobný proces jako u rozhodovacích stromů, pouze v obráceném pořadí). Tímto postupem lze jednoduše porovnat více modelů s odlišným počtem shluků. Mezi metody hierarchického shlukování patří metoda

nejbližšího a nejvzdálenějšího souseda, metoda průměrné vzdálenosti nebo metoda těžiště. Nehierarchické metody jako například K-Means, které jsou popsány dále, používají charakteristické výchozí případy, jako základ počátečních shluků, ke kterým se podle podobnosti přiřazují další případy. [12] [13] [19] [23]

Shluková analýza je citlivá na rozdíl v měřítkách proměnných, proměnné s větší směrodatnou odchylkou více ovlivňují zvolenou míru podobnosti. Některé metody proto umožnují při procesu shlukování provést standardizaci proměnných, u jiných je potřeba proměnné dopředu standardizovat. Mezi často používaný způsob standardizace patří normování pomocí Z-skóre. Shluková analýza je také citlivá na odlehlé případy. Některé metody umožňují nastavení určitého procenta případů, které budou v rámci analýzy považovány za odlehlé. Pro úspěšné provedení shlukové analýzy je třeba pečlivě vybrat proměnné tak, aby splňovaly uvedené předpoklady. Analýzou korelací vypočtených podle Pearsonova korelačního koeficientu byly v této práci numerické proměnné vybírány tak, aby se mezi nimi nevyskytovaly korelace větší než 0,5. [5] [13] [19]

Počet shluků je určen na začátku analýzy nebo je dopředu neznám a optimální počet vyplyne ze samotné analýzy. Algoritmy pro výběr počtu shluků v praxi používají heuristické procedury nebo formální testy. Nejjednodušším způsobem je určení počtu shluků pomocí dendrogramu, grafu, který ukazuje postupný rozklad datasetu do shluků a následné zvolení maximální míry rozkladu jak lze vidět na obrázku č. 8. [5] [12] [13]



Obr. 8 Dendrogram
Zdroj: Vlastní zpracování

Další volbou je výběr způsobu, kterým bude určována míra podobnosti případů. Mezi základní kategorie patří korelační míry, míry asociace a nejčastěji používané míry vzdálenosti. Mezi běžně používané míry vzdálenosti patří Hammingova, Euklidovská a Čebyševova vzdálenost. Euklidovská vzdálenost určuje vzdálenost dvou shluků pomocí rovné čáry. Lze ji použít pouze pro numerické proměnné a není vhodná pro situace, kdy proměnné nejsou standardizované. SPSS dále nabízí měření vzdálenosti pomocí logaritmu věrohodnostní funkce. Vzdálenost mezi dvěma shluky je vázána na pokles v logaritmické pravděpodobnosti, tak jak jsou spojeny do jednoho shluku. [5] [19]

Pro posouzení kvality modelu se v SPSS Modeler používá statistika zvaná silueta. Její výchozí hodnota je nula a nabývá rozmezí od mínus jedné pro nekvalitní modely do jedné pro modely vysoce kvalitní. Hodnota siluety nižší než nula znamená, že jednotlivé případy mají průměrnou vzdálenost mezi případy přiřazeného shluku větší než je minimální průměrná vzdálenost ke středu jiného shluku. Takové modely jsou z principu špatné a ihned zamítnuty. Hodnota siluety kombinuje koncept soudržnosti, který upřednostňuje modely s vysokou soudržností shluků a koncept odlišení, který upřednostňuje modely, u kterých

se shluky velice odlišují. Celková hodnota siluety je vypočtena jako průměr hodnot pro všechny případy. Hodnota pro konkrétní případ se vypočítá podle vzorce

$$\frac{(B - A)}{max(A, B)}$$

kde A je vzdálenost případu od středu shluku do kterého patří a B je minimální vzdálenost případu ke středu všech ostatních shluků. Siluetu lze vypočítat pro všechny případy nebo pro konkrétní shluky. Běžně se za ucházející považují modely s hodnotou siluety vyšší než 0,26. Kvalitní modely dosahují hodnot siluety přes 0,5. SPSS Modeler nabízí tři metody shlukové analýzy, K-Means, Kohonen a dvoukrokovou metodu. [12] [13] [19]

4.1.1 Metoda Kohonen

Kohonen metoda vychází z principu neuronových sítí. Známá je též pod názvy Knet nebo samo-organizující se mapa. Využívá dvouvrstvou síť, ve které jsou všechny neurony vstupní vrstvy propojeny se všemi neurony výstupní vrstvy. V průběhu algoritmu pak jednotlivé neurony soutěží o jednotlivé případy. Na počátku jsou váhy neuronů náhodné, po přidělení případu k neuronu je jeho váha a váhy okolních neuronů aktualizovány. Proces se opakuje, dokud nejsou změny vah dostatečně malé. Metodu lze také použít pro redukci dimenze. [13]

4.1.2 Metoda K-Means

K-Means metoda umožňuje použití velkého množství dat a vyžaduje dopředu určit počet shluků. Metoda umožňuje nastavení výchozích středů shluků. Pro shlukování případů jsou na výběr dvě metody, aktualizace středů shluků nebo pouze klasifikace. Metoda umožňuje uložit informace o příslušnosti případů do shluků, jejich vzdálenosti ke středu shluku a výsledné středy shluků. K-Means metoda je vhodná pro numerické proměnné, protože používá Euklidovskou vzdálenost. Nevýhodou je potřeba standardizace numerických proměnných před analýzou a nutnost dobrého odhadu počtu shluků, nepřesný odhad vede k zavádějícím výsledkům. Výsledek shlukování může být ovlivněn pořadím případů, proto je vhodné analýzu vícekrát opakovat s odlišným pořadím pro ověření stability modelu. [13] [19]

4.1.3 Dvoukroková metoda

Dvoukroková metoda byla vytvořena pro analýzu velkých datasetů a umožňuje použití kategoriálních i numerických proměnných. Další výhodou je větší odolnost vůči porušení předpokladu o normalitě a nezávislosti proměnných než u ostatních metod. Metoda používá aglomerativní hierarchickou shlukovací metodu, protože její použití kooperuje dobře s možností výběru vhodného počtu shluků pomocí zvoleného kritérionu. Dvoukroková metoda nepovoluje chybějící hodnoty. Takové případy jsou z analýzy odstraněny podle principu Listwise, který byl zmíněn v přípravě dat. Jméno metody je odvozeno z použitého principu, který se skládá ze dvou základních kroků. SPSS Modeler nabízí i použití volitelného mezikroku. [13] [23]

Prvním krokem je rozdělení případů do velkého množství malých shluků. Tento krok funguje na principu sekvenčního shlukování. Implementaci metody v SPSS Modeler lze popsat následovně. Algoritmus nejdříve prochází po jednom jednotlivé případy a na základě zvolené míry podobnosti rozhoduje, zda mají být zařazeny do existujících shluků nebo je třeba vytvořit nový shluk. Tento proces je implementován pomocí takzvaného CF stromu (cluster feature tree). CF strom se skládá z uzlů několika úrovní. Souhrny případů v listech stromu představují konečné malé shluky. Nelistové uzly stromu směrují nové případy do správných listových uzlů. Listové uzly jsou ve stromu charakterizovány počtem zařazených případů, dále průměrem a variací numerických proměnných a skupinou kategoriálních proměnných. [13] [23]

Po dosažení listového uzlu pak algoritmus rozhodne, zda bude případ zařazen do existujícího uzlu nebo bude vytvořen nový uzel. Pokud je případ dostatečně podobný ostatním případům v uzlu, pak je absorbován a charakteristiky uzlu jsou přepočítány. V opačném případě je vytvořen nový uzel obsahující daný případ. V situaci kdy CF strom přeroste za povolenou velikost je přestavěn na základě původního CF stromu zvýšením míry vzdálenosti. Přestavěný CF strom je menší a jeho uzly pojmou více případů. Tento proces pokračuje, dokud stromem neprojdou všechny případy. Jednotlivé uzly tedy neobsahují konkrétní případy ale pouze souhrnné vlastnosti všech případů daného

uzlu. Díky těmto vlastnostem je velikost CF stromu mnohem menší než původní data a lze jej efektivněji uložit v paměti. Struktura vytvořeného CF stromu může být závislá na pořadí vstupů. Proto je vhodné před průběhem analýzy případy náhodně seřadit. [13] [23]

Volitelný krok analýzy slouží k vyřazení odlehlých případů. Odlehlé případy nezapadají dobře ani do jednoho z existujících shluků. Před přestavěním CF stromu procedura zkontroluje možné odlehlé případy a z analýzy je dočasně vyřadí. Po sestavení výsledného CF stromu se zkontroluje, jestli lze některé odlehlé případy zařadit zpět bez nutnosti zvětšení stromu. [13] [23]

Druhým krokem je samotné shlukování, které používá malé shluky z prvního kroku. Algoritmus je poté spojí do požadovaného počtu shluků. Skutečnost že počet malých shluků z prvního kroku je mnohokrát menší, než počet původních případů znamená, že druhý krok je výpočetně velmi rychlý a efektivní. Kvalita shlukování je závislá na počtu malých shluků vzniklých v prvním kroku. Více malých shluků zpřesní výsledný model, ale zpomalí zpracování druhého kroku. [13] [23]

4.2 Diskriminační analýza

Diskriminační analýza slouží ke klasifikaci závislé kategoriální proměnné pomocí nezávislých numerických proměnných. Používá se v mnoha oblastech, především v medicíně a komerční sféře. Mezi praktické aplikace patří například analýza potenciálních členů Evropské unie (Tevdovski a Trpkova, 2010) nebo výdělečnost zahraničních bank (Shanthi a Shobana, 2010). [5] [11] [24] [26]

Za autora diskriminační analýzy je považován sir Ronald Aylmer Fisher (1890-1962). Mezi základní předpoklady patří nezávislost případů, vícerozměrné normální rozdělení numerických nezávislých proměnných a shoda kovariančních matic jednotlivých skupin závislé proměnné. Závislá proměnná je ideálně nominální a jména skupin jsou překódována na numerické hodnoty. Dalším předpokladem je, že jednotlivé případy patří vždy pouze do jedné skupiny závislé proměnné. [5] [11]

Cílem diskriminační analýzy je nalezení lineární kombinace p závislých proměnných, která dokáže přiřazovat prvky do skupin závislé proměnné lépe než jakákoliv jiná kombinace podle vzorce

$$Y = b^T x$$

kde b^T je vektor parametrů který umožní dosáhnout co nejvyšší variability mezi skupinami a nejnižší variability uvnitř skupin, což zajistí nejlepší možné odlišení skupin závislé proměnné. K takovému stavu se lze přiblížit při maximálním podílu meziskupinové a vnitroskupinové variability podle vzorce

$$F = \frac{b^T B b}{b^T E b}$$

kde B je matice meziskupinové variability, E je matice vnitroskupinové variability a b je vektor, jehož prvky maximalizují diskriminační kritérium. Hodnota F se nazývá Fisherovo diskriminační kritérium. Vektor b se vypočítá použitím parciální derivace diskriminačního kritéria a položením prvků vektoru b rovné nule. Z tohoto postupu vyjde několik řešení v podobě vlastních čísel a největší z nich odpovídá vektoru b. Konkrétní hodnoty vektoru b, které tvoří ideální lineární kombinaci, jsou poté dopočítány. Lineární kombinace Y_1 je považována za první diskriminant, graficky jde o projekci bodů odpovídací jednotlivým případům na přímku. Umístění bodů na přímce určuje jejich diskriminační skóre. Právě v situacích kdy je třeba tří a více diskriminantů dochází k projekci bodů do vícerozměrného prostoru. Pro závislé proměnné se dvěma skupinami stačí pro popsání celkové variability jediný diskriminant. Počet potřebných diskriminantů pro závislé proměnné s více než dvěma kategoriemi je roven počtu skupin sníženým o jedna nebo počtu nezávislých proměnných. Použije se nižší z těchto dvou hodnot. [5] [11]

Ideálním stavem je získání diskriminantu, který skupiny odlišuje co nejvíce. Po vypočtení diskriminantů a dosazení známých hodnot do rovnice lineární kombinace Y získáme diskriminační skóre jednotlivých případů. Poté se vypočítá průměrné diskriminační skóre v jednotlivých skupinách. Čím více se průměrná skóre odlišují, tím lépe diskriminanty odlišují jednotlivé skupiny. Diskriminační analýzu negativně ovlivňuje existence korelací mezi nezávislými proměnnými.

Korelované proměnné mají podobně velké koeficienty a jejich znaménka se shodují. [5] [11] [17]

Klasifikace pomocí diskriminační analýzy vychází z funkce původních proměnných. Diskriminační funkce má v skupinách závislé proměnné střední hodnoty s určitou vzdáleností. Po výpočtu středu těchto dvou středních hodnot lze zařazovat případy do skupin podle vypočtené hodnoty diskriminační funkce. Pokud je hodnota blíže průměrné hodnotě jedné skupiny, potom je do ní zařazena a naopak. Pro situace kdy se apriorní pravděpodobnosti závislé proměnné velice odlišují, dochází k častějšímu zařazení případů do skupiny s vyšší apriorní pravděpodobností. Tuto chybu lze minimalizovat zahrnutím apriorních pravděpodobností do výpočtu diskriminačního kritéria použitím Bayesova kritéria. Pro zjištění přesnosti klasifikace lze zvolit ze dvou hlavních přístupů, rozdělení dat na trénovací a testovací data nebo odhad pomocí redistribuce. Pro první přístup je třeba velký vzorek dat. Trénovací část dat je použita pro výpočet diskriminačního kritéria, které je následně použito na testovací část dat. Další možností je použití křížové validace, která náhodně zvolí část dat pro výpočet diskriminačního kritéria, čímž lze dosáhnout maximální nestrannosti. [5] [11]

SPSS Modeler umožňuje nastavit řadu parametrů pro diskriminační analýzu. Hlavním krokem je výběr metody analýzy. SPSS Modeler nabízí celkem dvě volby. První nabízenou metodou je Enter, která používá všechny zvolené nezávislé proměnné. Druhou metodou je Stepwise, která začíná pouze s konstantou a postupně přidává do modelu ty proměnné, které zvyšují jeho kvalitu. Tato metoda zahrnuje výběr statistiky, podle které se budou proměnné do modelu zařazovat a vyřazovat. V prvním kroku jsou způsobilé proměnné s vyšší úrovní zařazení vybrány před proměnnými s nižší úrovní zařazení. Pořadí vstupu proměnných se stejnou sudou úrovní zařazení je rozhodnuto podle pořadí zadaného v analýze. Pořadí vstupu proměnných se stejnou lichou úrovní zařazení je rozhodnuto podle hodnoty zvolené statistiky. Proměnná s nejlepší hodnotou je vybrána první. Když je první fáze dokončena, jsou všechny vybrané proměnné s úrovní zařazení jedna, otestovány pro vyřazení. Proměnná je vyřazena, pokud její F-hodnota pro vyřazení je nižší než zvolená F-hodnota. Pokud je více proměnných způsobilých pro vyřazení, jsou vyřazeny ty proměnné, které ponechají nejlepší

hodnotu vybrané statistiky pro zbývající proměnné. Vyřazování proměnných pokračuje, dokud nejsou odstraněny všechny proměnné splňující podmínky pro vyřazení. Sekvenční vstup proměnných poté pokračuje stejně, jako bylo popsáno výše kromě toho, že po každém kroku jsou proměnné s úrovní zařazení jedna také kontrolovány pro vyřazení, jak bylo zmíněno výše. Proměnné s úrovní zařazení nula do modelu nevstupují nikdy, ačkoliv jsou pro ně některé statistky zobrazeny. [13] [17]

4.3 Ostatní metody

Prvotní analýzy zahrnovaly použití i dalších statistických a data miningových metod. Vybrané metody, které byly následně zamítnuty, jsou popsány zde spolu s důvody zamítnutí. Tím však výčet metod nekončí. Použití lineární regrese bylo zamítnuto po exploraci dat. Pro tuto analýzu je potřeba, aby zvolené nezávislé proměnné byly co nejméně korelované. Explorace ukázala vysoké korelace mezi numerickými proměnnými z jednotlivých kategorií kalkulátoru, pro analýzu by proto byl nedostatek vhodných nezávislých proměnných. [5]

4.3.1 Neuronové sítě

Neuronové sítě lze použít pro klasifikaci, predikci a shlukování. V historii nejdříve došlo k matematickému popisu neuronu a poté se objevily snahy o použití soustavy neuronů (neuronových sítí) pro klasifikaci. Metoda vychází z principů fungování lidského mozku. Mozek obsahuje miliardy mezi sebou propojených nervových buněk zvané neurony. Neurony se skládají ze synapsí, které spojují jednotlivé neurony a pomocí kterých se přenáší signály. Jednotlivé neurony mají takzvanou prahovou hodnotu. Při překročení této hodnoty je neuron sepnut a přes jeho synapse začne posílat signály. Z těchto principů vychází i samotná metoda. Použití neuronových sítí je vhodné v situacích, kdy data obsahují velké množství numerických proměnných (na rozdíl od rozhodovacích stromů). Lze je také použít v kombinaci s jinými metodami, například diskriminační analýzou pro výzkum klientů vlastnící kreditní kartu (Yazıcı, 2011) nebo predikci rizikových bank (Swicegood a Clark, 2011). Kategoriální proměnné s třemi a více kategoriemi

je vhodné pro jednoduchost binarizovat. Topologie sítě vychází z počtu proměnných a trénovacích dat. [5] [6] [25] [27]

Jednotlivé neurony mají tzv. aktivační funkci, která při sepnutí přenese na výstup hodnotu mínus jedna nebo jedna. Aktivační funkce se skládá z kombinační funkce, která spojí vstupní hodnoty a transferní funkce, která vypočítá výstupní hodnotu z hodnoty kombinační funkce. V dnešní době se používají vícevrstvé sítě, mezi nejčastější typ patří vícevrstvý perceptron. Principiálně zde nejsou mezi sebou propojeny jednotlivé neurony stejné úrovně, ale neuron patřící do jedné úrovně je propojen se všemi neurony další úrovně. Pro zjednodušení se používá síť s jednou skrytou úrovní. Pro trénování sítě se nejčastěji používá metody zpětného šíření, kdy jsou informace nejdříve poslány z vstupní úrovně do výstupní, a při učení sítě se postupuje od výstupní úrovně ke vstupní. Často lze zastaralou síť aktualizovat použitím nových trénovacích dat. Učící proces je zastaven při dosažení zadaného počtu opakování nebo při dosažení zvolené úrovně chybové funkce. [5] [6]

Mezi výhody neuronových sítí patří jednoduchá automatizace a schopnost reprezentovat libovolné spojité funkce. Mezi nevýhody patří výběr topologie, složitost interpretace výstupů nebo riziko přeučení sítě. Způsob jakým vytvořit síť s nejlepší možnou topologií je závislý na volbě počtu neuronů ve skryté vrstvě. Velké množství způsobuje zdlouhavé a komplikované učení sítě a přináší riziko přeučení, kdy síť rozpoznává příliš detailní a irelevantní souvislosti. Problematika interpretace vychází ze skutečnosti, že výsledek analýzy je závislý na topologii sítě a vazbách mezi neurony. Síť se v podstatě chová jako černá skříňka, do které není vidět. I přes nemožnost zjistit jak byl výstup získán, má použití neuronových sítí často vysokou praktickou hodnotu. Pro jednodušší interpretaci výstupů je vhodné použití dostatečně srozumitelných vstupních dat a jasný cíl modelování. Z důvodu složité interpretace výstupů byla po prvotních analýzách tato metoda zamítnuta. Při klasifikaci také docházelo k zařazování naprosté většiny případů do skupin s vyšší apriorní pravděpodobností. [5] [6]

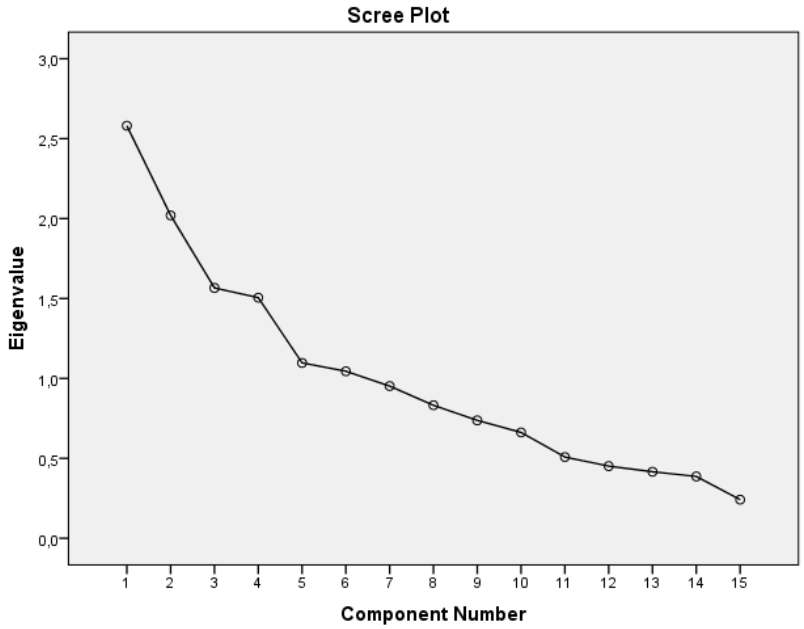
4.3.2 Logistická regrese

Logistická regrese slouží pro klasifikaci. Vychází z klasické regresní analýzy, závislá proměnná je ale kategoriální. Závislá proměnná může být dichotomická nebo multinomiální. V případě dichotomické proměnné musí být její datový typ nastaven na flag. Svými vlastnostmi se podobá diskriminační analýze, proto jsou často používány společně, například pro predikci rizikových úvěrů (Memić, 2015). Na rozdíl od diskriminační analýzy však logistická regrese umožňuje použití nezávislých kategoriálních proměnných. Mezi předpoklady patří nezávislost jednotlivých případů a vzájemně se nepřekrývající kategorie závislé proměnné. Mezi výhody patří častá vysoká přesnost klasifikace a nižší požadavky na normalitu numerických proměnných. Mezi nevýhody patří nutnost odstranění multikolinearit. Podobně jako u neuronových sítí při klasifikaci docházelo k neúměrnému zařazování případů do skupiny s vyšší apriorní pravděpodobností. Proto byla použita diskriminační analýza. [13] [17] [20]

4.3.3 Faktorová analýza

Faktorová analýza se používá jako jedna z metod pro snížení dimenze dat. Cílem takových metod je výpočet menšího počtu nových proměnných, které budou reprezentovat původní proměnné. Mezi další použití patří explorace dat, například odhalování multikolinearit. Využití má také pro odhalování skrytých korelací a kombinací proměnných neviditelných na první pohled, které ale mohou významně ovlivňovat závislosti jiných proměnných. Nové proměnné, zvané faktory, jsou vyjádřené jako lineární kombinace původních proměnných. Ideální faktory objasňují co největší část variace původních proměnných, je jich nízký počet a jsou málo korelované. V případě faktorové analýzy je obzvlášť důležité dodržení předpokladů a dispozice odborných znalostí z dané oblasti. [12] [13]

Před analýzou je třeba zvolit počet faktorů. Mezi základní přístupy patří zvolení konkrétního počtu nebo nastavení hodnoty vlastního čísla, které použije pouze faktory s vyšší než zadanou hodnotou. V praxi se často používá hodnota jedna. Graf na obrázku č. 9, zachycující pokles hodnot vlastních čísel faktorů oproti počtu faktorů, se nazývá scree plot graf. [12] [13]

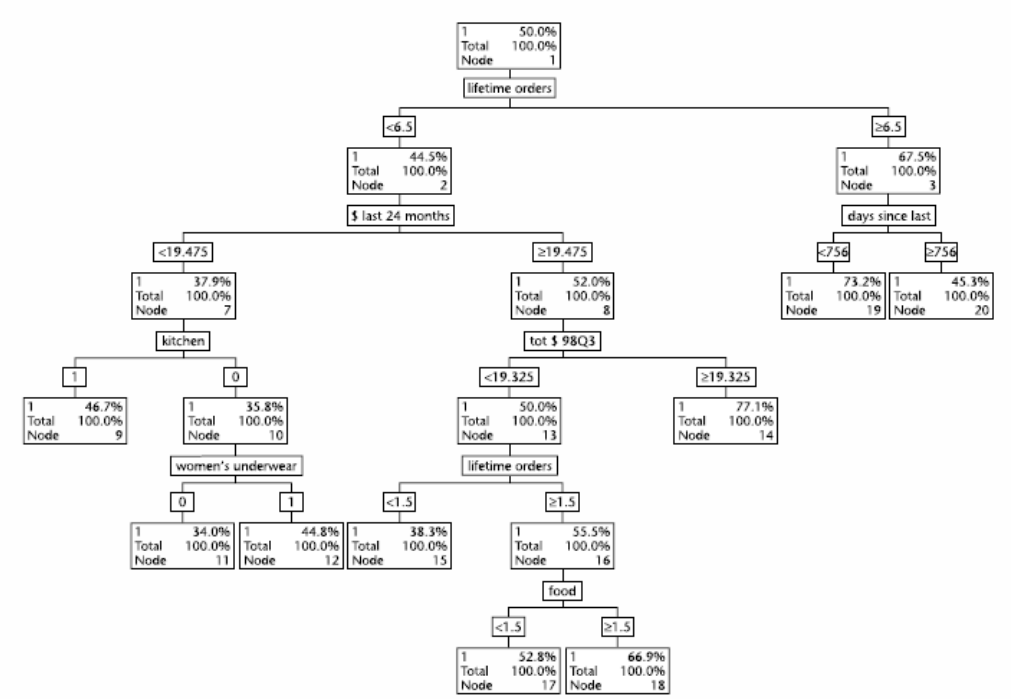


Obr. 9 Scree plot graf
Zdroj: Vlastní zpracování

K základním předpokladům patří použití numerických proměnných a nezávislost jednotlivých případů. V praxi faktorová analýza vychází z korelační matice. Kovarianční matici lze použít, pouze pokud mají všechny proměnné stejné měřítko. Na vypočítané faktory se dále používají rotace neboli maticové transformace. Ortogonální rotace jako Varimax a Orthomax slouží k odstranění korelací mezi faktory, šikmé rotace naopak umožňují existenci korelací mezi faktory. Mezi základní metody odvození faktorů patří metoda hlavních komponent, odhady komunalit a odhady specifických rozptylů. Metoda byla zamítnuta pro nejednoznačnost vypočtených faktorů, použití v kombinaci se shlukovou analýzou navíc nepřineslo zvýšení hodnot siluety modelů. [12] [13]

4.3.4 Rozhodovací stromy

Rozhodovací stromy slouží ke klasifikaci a predikci. Metoda vychází z přístupu „rozděl a panuj“. Původní data jsou dělena do skupin, které odpovídají jednotlivým uzlům stromu od shora dolů, dokud nejsou jednotlivé větve dostatečně podrobné. Rozhodovací strom lze vidět na obrázku č. 10. [5] [6]



Obr. 10 Příklad rozhodovacího stromu
Zdroj: Převzato z [6]

Na začátku jsou veškeré případy zařazeny do jedné skupiny a na konci jsou rozděleny do uzlů. Ideálním řešením je menší strom s vysokou přesností klasifikace. Základem je, aby další úroveň uzlů vždy rozdělila předchozí data na menší, odlišné části. Mezi časté způsoby dělení patří entropie, informační zisk nebo Gini index. V praxi je potřeba výsledný strom prořezávat (zmenšit jeho velikost) jelikož při dosažení maximální přesnosti klasifikace by byl výsledný strom příliš velký. V podstatě jde o výběr nejlepšího podstromu, nahrazením nelistových uzlů listovými. Menší strom lze také lépe interpretovat. Kvalitu výsledného stromu lze hodnotit především podle jeho klasifikační schopnosti, dále podle počtu případů v jednotlivých listových uzlech nebo poměrem případů

v uzlech. Mezi možné použití patří detekce podvodů s kreditními kartami (Sahin a Duman, 2015). Odlišnou variantou jsou regresní stromy, které slouží k predikci. Uzly stromu obsahují průměrné hodnoty numerických proměnných. Růst stromu vychází ze směrodatných odchylek a jejich redukce. V dnešní době se používají algoritmy C.5, CaRT (classification and regression tree) a CHAID (Chi-square Automatic Interaction Detection). [5] [6] [28]

Mezi výhody rozhodovacích stromů patří vhodnost pro kategoriální data a možnost použití trénovacích dat s chybějícími hodnotami. Rozhodovací stromy také umějí využít možnou prediktivní hodnotu chybějících hodnot. Další výhodou je jednoduchá interpretace. Jelikož strom v principu vyjadřuje pravidla, lze výsledek jednoduše vyjádřit v přirozeném jazyce nebo jako databázový příkaz. Mezi nevýhody patří nutnost diskretizace numerických proměnných. Nejčastěji se používá binarizace, rozdělení do dvou skupin. Protože prvotní analýzy ukázaly, že získané modely byly nekvalitní, bylo použití rozhodovacích stromů zamítnuto. [5] [6]

5 Vybrané analýzy business dat

Analýzy dat byly provedeny v programu SPSS Modeler 17 od firmy IBM. Tento nástroj umožňuje jednodušší nastavení parametrů metod, tvorbu dodatečných transformací, generování grafů a další užitečné funkce. Každý uzel plní specifickou funkci a podle použití se dělí do kategorií. Patří sem uzly sloužící jako zdroje dat, dále uzly pro výběr a změnu pořadí datových záznamů, uzly pro úpravu a výběr proměnných, pro vizualizaci dat, tvorbu modelů a uzly pro analýzu a export výstupů. Jako zdroj byl použit dataset připravený v SPSS Statistics. Z pohledu analýzy je zajímavé zkoumat určité skupiny klientů, proto kromě analýzy celého datasetu byly zkoumány i podskupiny klientů, kteří vlastní kreditní kartu, kteří nepoužívají přímé bankovnictví a klienti malých bank. Tato část odpovídá čtvrté fázi metodiky CRISP-DM. Celkem byly provedeny dvě úlohy. [15]

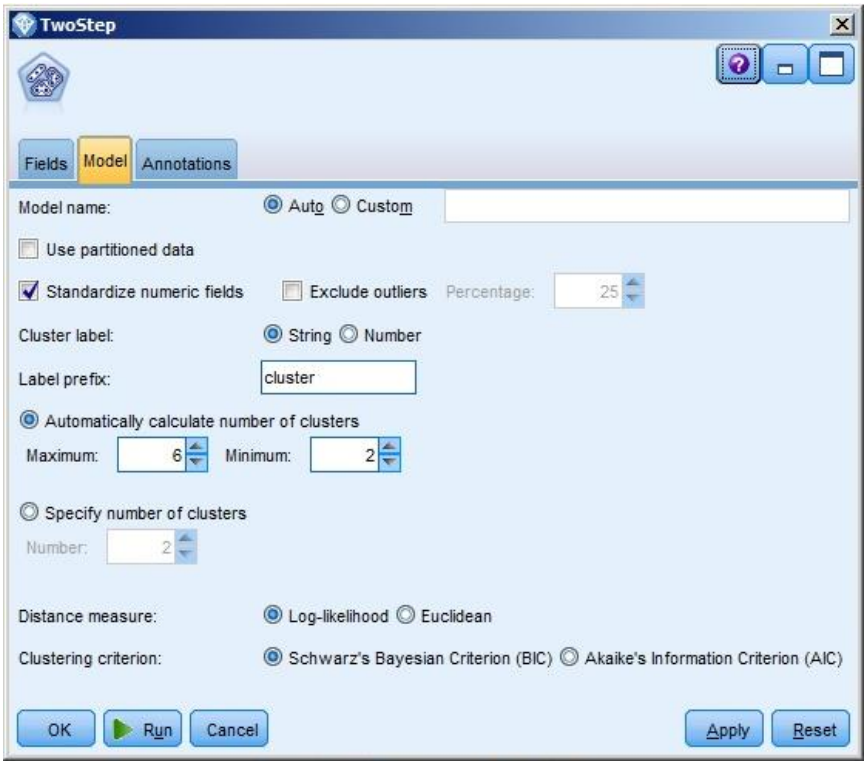
5.1 Analýza skupin bankovních klientů

Cílem úlohy je identifikace přirozeně se vyskytujících skupin klientů retailového bankovnictví. První krok úlohy je výběr proměnných. Protože skupiny nejsou dopředu známé nelze vybrat závislou proměnnou. Právě proto byla zvolena shluková analýza. Volba kategoriálních proměnných byla provedena s ohledem na apriorní pravděpodobnosti a jejich předpokládanou popisnou schopnost. Numerické proměnné byly vybrány s ohledem na splnění předpokladů o normalitě a neexistenci velkých korelací. Pro své výhody byla zvolena dvoukroková metoda, která je vhodná pro velké datasety a je méně citlivá na proměnné, které nemají normální rozdělení, než například metoda K-Means. Přesto byly vybrané numerické proměnné před analýzou transformovány. [13] [15]

Dvoukroková metoda umožňuje standardizovat numerické proměnné na stejnou míru o průměru nula a rozptylu jedna. Tato byla funkce použita, protože použité numerické proměnné mají odlišné míry. Další nastavení metody zahrnuje volbu použití volitelného mezikroku pro odstranění odlehlých případů. Procentuální hodnota určuje velikost malých shluků, které jsou podezřelé z obsažení odlehlých případů. Pro tuto úlohu bylo zvoleno rozmezí počtu shluků

od dvou do šesti. Větší počet by bylo obtížné popsat, protože rozdíly mezi shluky by byly zanedbatelné. Došlo by také k zvýšení počtu malých shluků, které sice reprezentují existující skupiny, praktické využití takové informace je však nízké. [13] [15]

Pro nalezení optimálního počtu shluků SPSS Modeler nabízí dva informační kritériony, AIC (Akaike information criterion) a BIC (Bayesian information criterion). Pro analýzu byl zvolen druhý jmenovaný, protože se mu dokumentace SPSS věnuje podrobněji. Algoritmus nejprve spočítá BIC pro každý povolený počet shluků, který poté použije pro počáteční odhad. V druhé fázi je počáteční odhad upřesněn nalezením největšího relativního nárůstu vzdálenosti mezi dvěma nejbližšími shluky v každé fázi hierarchického shlukování. Tento postup se opakuje, dokud nejsou vypočteny všechny poměry minimální vzdálenosti. Nakonec se porovnají dva největší získané poměry. Pokud je největší poměr 1,15x větší než druhý největší poměr, potom je vybrán model s největším poměrem jako nejlepší řešení, jinak je z těchto dvou modelů vybrán ten, s vyšším počtem shluků jako optimální. Z důvodu použití kategoriálních proměnných byl pro měření vzdálenosti zvolen logaritmus věrohodnostní funkce. Výsledné nastavení parametrů metody je zřejmé z obrázku č. 11. [5] [12] [13] [15]



Obr. 11 Nastavení parametrů dvoukrokové metody
Zdroj: Vlastní zpracování

5.2 Klasifikace klientů podle využívání okrajových služeb

Tato úloha si klade za cíl ověření pravdivosti zvolené hypotézy. Pro ověření hypotézy byla zvolena hladina významnosti $\alpha = 0{,}05$. Hypotéza je formulována následovně:

H_0: *Využívání okrajových bankovních služeb jako Cashback (výběr hotovosti na pokladně při platbě kartou) není závislé na četnosti využití ostatních bankovních služeb.*

H_1: *Využívání okrajových bankovních služeb jako Cashback (výběr hotovosti na pokladně při platbě kartou) je závislé na četnosti využití ostatních bankovních služeb.*

Pokud bude nulová hypotéza vyvrácena a přijata hypotéza alternativní, bude další částí úlohy nalezení klasifikačního modelu, který bude nejpřesněji klasifikovat klienty využívající okrajové služby. Za okrajové služby jsou v úloze

považovány Cashback, nadměrný vklad a výběr z bankomatu v zahraničí (z vlastní nebo cizí banky). Pro tuto úlohu byla použita diskriminační analýza, protože data obsahují velké množství numerických proměnných. Výběr proměnných pro tuto analýzu byl volen s ohledem na jejich relativní významnosti a nízké korelace. [11]

V SPSS Modeler uzlu diskriminační analýzy je třeba nastavit několik parametrů. Prvním je výběr metody. Pro tuto byla zvolena metoda Stepwise, protože poskytuje detailnější výstupy. Metoda dále nabízí volbu, zda bude použita matice vnitroskupinových nebo meziskupinových kovariancí. Použití matice vnitroskupinových variací vychází ze splnění předpokladu o shodě kovariančních matic. Pro ověření tohoto předpokladu byl použit Boxův M test. Nulová hypotéza Boxova M testu

$$H_0: \Sigma_1 = \cdots = \Sigma_g$$

říká, že se kovarianční matice shodují, alternativní hypotéza říká, že se dané matice nehodují. Pokud se matice neshodují, je potřeba pro klasifikaci použít kvadratickou diskriminační funkci. Kvadratická diskriminační funkce se používá, když se matice velmi odlišují a pokud jsou použita rozsáhlá data. [5] [11] [13]

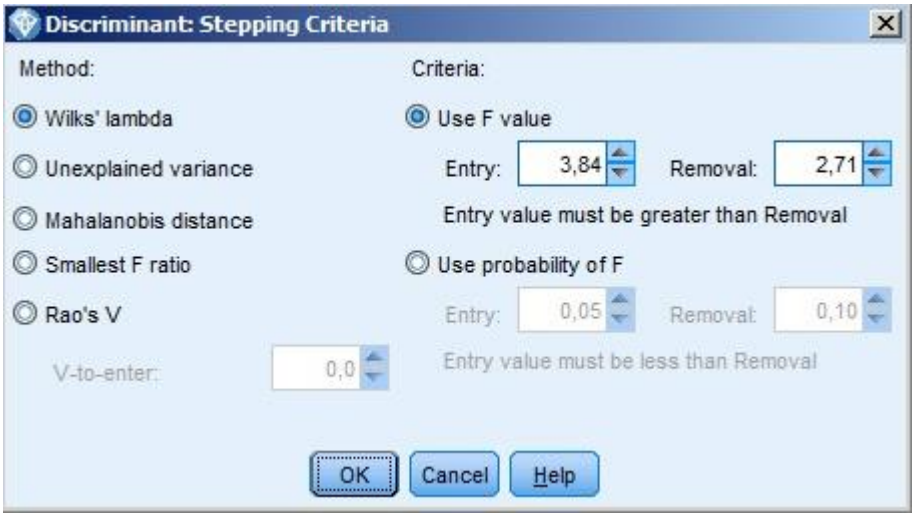
Dalším nastavením je volba, zdali budou klasifikační koeficienty uvažovat apriorní pravděpodobnosti. Jelikož apriorní pravděpodobnosti skupin závislých proměnných použitých v úloze se velmi odlišují, byla zvolena možnost, která uvažuje apriorní pravděpodobnosti. Skutečný počet případů ve skupinách poté určuje dřívější pravděpodobnosti, takže klasifikační koeficienty jsou upraveny, aby zvýšili pravděpodobnost příslušnosti ve větší skupině oproti menší. Nastavení parametrů uzlu diskriminační analýzy je vidět na obrázku č. 12. [13] [15]

Obr. 12 Nastavení parametrů diskriminační analýzy
Zdroj: Vlastní zpracování

Nastavení metody Stepwise, zobrazené na obrázku č. 13, vyžaduje upřesnění použité statistiky a kritérií pro zařazování a vyřazování proměnných. SPSS Modeler nabízí pět odlišných statistik. Pro úlohu bylo zvoleno Wilksovo lambda. Jedná se o statistiku, testující shodu skupinových průměrů. Nulová hypotéza

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

říká, že se průměry shodují. V každém kroku je vybrána proměnná, která nejvíce minimalizuje celkové Wilksovo lambda. Pro zařazení a vyřazení proměnných metoda nabízí na výběr mezi nastavením kritických hodnot a hladin významnosti. Pro úlohu byla vybrána první možnost. Pokud je F-hodnota proměnné vyšší, než zadaná vstupní F-hodnota, pak je zařazena do modelu. Naopak pokud je F-hodnota proměnné nižší, než zadaná F-hodnota pro vyřazení, pak je z modelu vyřazena. Nastavením kritických hodnot lze regulovat počet proměnných v modelu. [15] [22]

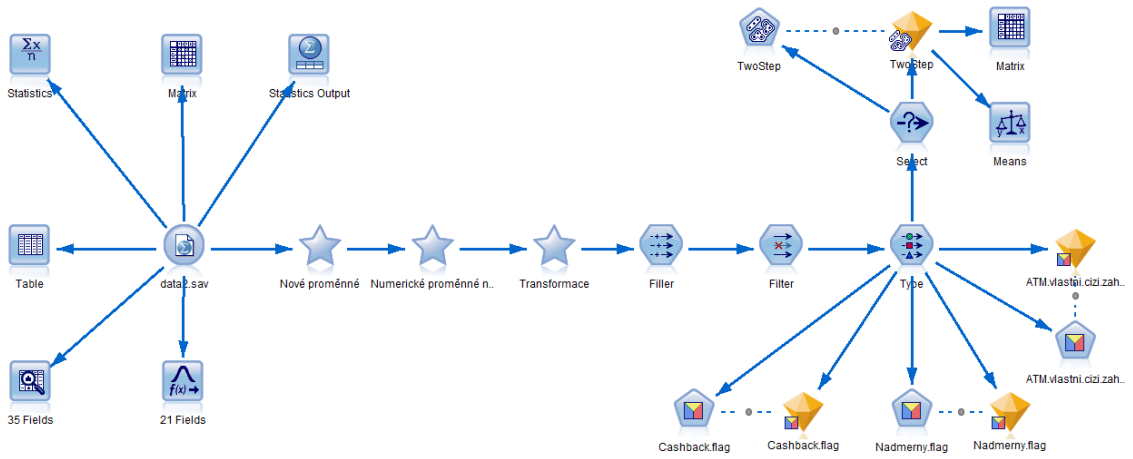


Obr. 13 Výběr metody pro Stepwise
Zdroj: Vlastní zpracování

Volitelné výstupy analýzy jsou rozděleny do tří kategorií. První kategorie zahrnuje popisné statistiky, matice a koeficienty funkcí. Klasifikační kategorie umožňuje do výstupu zahrnout grafy nebo výsledek křížové validace. Poslední kategorie umožňuje zobrazení podrobných informací o průběhu metody Stepwise. Pro získání maximálního množství informací o modelu byly použity všechny volitelné výstupy. [13] [15]

6 Interpretace výstupů analýz

Interpretace výstupů odpovídá páté fázi metodiky CRISP-DM. Vybrané modely a výstupy jsou hodnoceny a analyzovány. Modely popsané v jednotlivých úlohách byly vybrány po mnoha opakování a testování odlišných nastavení parametrů metod a použitých proměnných. Výstupy metod jsou v proudu obsaženy v takzvaných nuggetech. V praxi vznikají proudy, které obsahují například dvacet navazujících Derive uzlů. Řešením je sjednocení uzlů podobných funkcí do souhrnných uzlů ve tvaru hvězdy. Proud použitý pro tuto práci obsahuje tři takové uzly, které slouží pro transformace proměnných, tvorbu nových proměnných a převedení vybraných numerických proměnných na dichotomické. Výsledný proud s výstupy obou úloh je zobrazen na obrázku č. 14. [13] [15]



Obr. 14 Výsledný proud pro obě úlohy
Zdroj: Vlastní zpracování

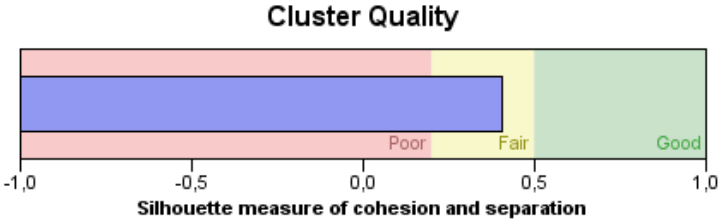
6.1 *Analýza skupin bankovních klientů*

V této úloze byl analyzován celý dataset i podskupiny zmíněné v přípravě dat. Popsaný model analyzuje právě podskupinu bankovních klientů vlastnící kreditní kartu. Důvodem je zejména fakt, že modely v této podskupině vycházely kvalitněji. Celkem bylo použito zhruba 4500 případů. Vybrané výstupy jsou níže podrobně popsány. Na obrázku č. 15 je souhrn základních informací o počtu shluků, použitých proměnných a výsledná hodnota siluety modelu. Celkem bylo

použito devět proměnných a na základě BIC kritérionu byl vybrán model s pěti shluky. Hodnota siluety modelu vyšla 0,41.

Model Summary

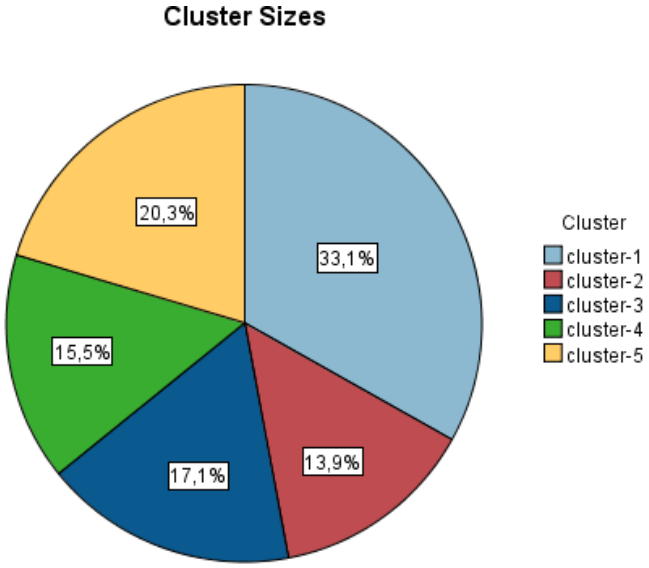
Algorithm	TwoStep
Inputs	9
Clusters	5



Obr. 15 Souhrn informací o modelu a hodnota siluety
Zdroj: Vlastní zpracování

Na obrázku č. 16 jsou znázorněny relativní velikosti shluků. Poměr mezi největším a nejmenším shlukem je necelých 1 : 2,5.

Size of Smallest Cluster	645 (13,9%)
Size of Largest Cluster	1533 (33,1%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2,38



Obr. 16 Velikosti shluků
Zdroj: Vlastní zpracování

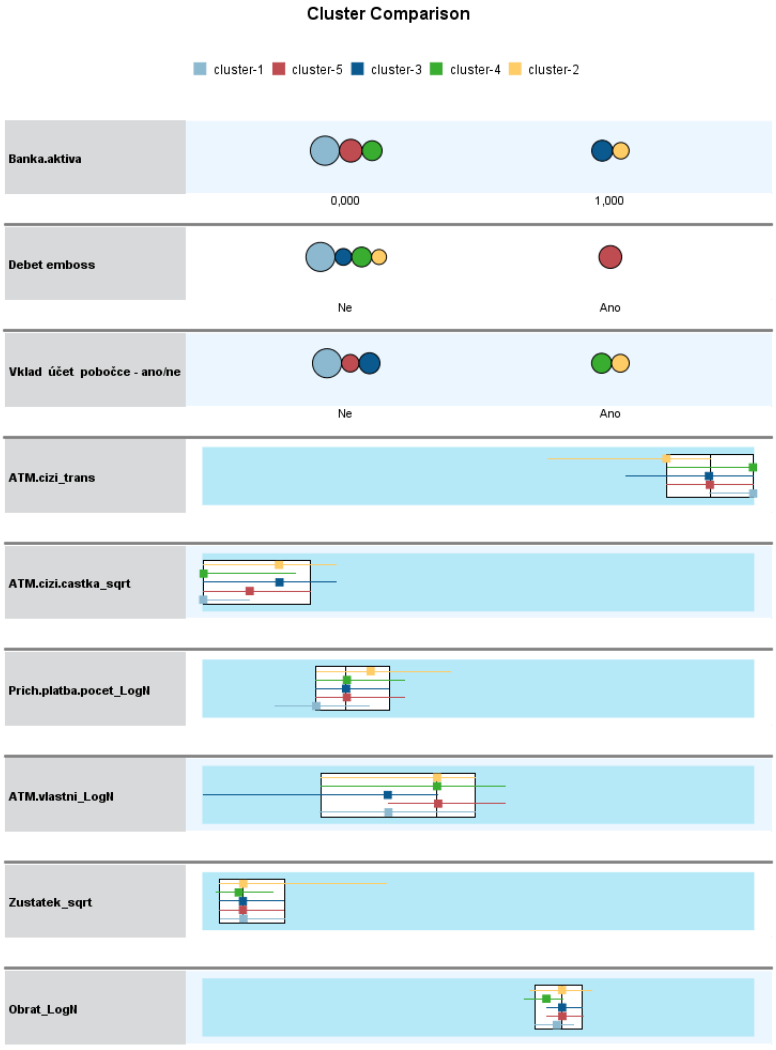
V tabulce č. 4 jsou zobrazeny vypočtené významnosti použitých proměnných. Jako nejvýznamnější se ukázaly kategoriální proměnné. Mezi významné numerické proměnné patří *ATM.cizi_trans* a *Prich.platba.pocet_LogN*.

Tabulka 4 Významnosti nezávislých proměnných

Proměnná	Významnost
Banka.aktiva	1
Vklad účet pobočce - ano/ne	1
Debet emboss	1
ATM.cizi_trans	0,54
ATM.cizi.castka_sqrt	0,3
Prich.platba.pocet_LogN	0,25
ATM.vlastni_LogN	0,2
Zustatek_sqrt	0,19
Obrat_LogN	0,09

Zdroj: Vlastní zpracování

Výstup dále umožňuje zobrazení podrobnějších popisů a vizualizaci vzdáleností mediánů proměnných pro jednotlivé shluky. Hodnoty lze vidět na obrázku č. 17. U numerických proměnných barevný čtverec symbolizuje hodnotu mediánu pro konkrétní shluk. Je patrné, že čím více se odlišují hodnoty mediánů, tím je proměnná významnější. [13] [15]



Obr. 17 Grafické porovnání proměnných a výsledných shluků

Zdroj: Vlastní zpracování

Průměrné hodnoty proměnných v jednotlivých shlucích, získané pomocí uzlu Means, jsou v tabulce č. 5. Tabulka č. 6 zobrazuje procentuální hodnoty využití okrajových služeb pro konkrétní shluky, získané pomocí uzlu Matrix. [13] [15]

Tabulka 5 Hodnoty proměnných pro jednotlivé shluky

Shluk	1	2	3	4	5
Velikost	33,1 %	14 %	17,1 %	15,5 %	20,3 %
Banka.aktiva	Velké (100%)	Malé (76,3 %)	Malé (100%)	Velké (100%)	Velké (100%)
Vklad účet pobočce - ano/ne	Ne (100%)	Ano (88,1%)	Ne (100%)	Ano (100%)	Ne (59,3%)
Debet emboss	Ne (100%)	Ne (63,6%)	Ne (64,1%)	Ne (100%)	Ano (100%)
ATM.cizi	0,5	1,9	1,4	0,9	1
ATM.cizi.castka	809	2939	2283	1121	1525
Prich.platba.pocet	3,2	8,9	4,4	5,4	4,8
ATM.vlastni	2,8	3,5	2,2	3	4
Zustatek	24980	89665	19332	13674	17969
Obrat	21732	46301	24762	18819	26941

Zdroj: Vlastní zpracování

Tabulka 6 Četnosti využití okrajových služeb

Shluk	1	2	3	4	5
Cashback	5,7%	17,8%	10,1%	15,3%	24,5%
Výběr v zahraničí	5,8%	36,4%	16,8%	18,5%	33,1%
Nadměrný vklad	1%	16,8%	1%	16,6%	5,3%

Zdroj: Vlastní zpracování

Jednotlivé shluky jsou popsány s ohledem na použité proměnné, jejich celkovou a relativní významnost a průměrné hodnoty proměnných v jednotlivých shlucích.

První, největší shluk reprezentuje skupinu klientů velkých bank, kteří mají účet pouze pro potřeby kreditní karty. Tito klienti nenavštěvují pobočku své banky a nemají ani další nekreditní karty. Vybírají částky do 1 000 Kč z bankomatů cizí

banky. Průměrný měsíční obrat a zůstatek se pohybují okolo 25 000 Kč. Mají nejnižší počet příchozích plateb a téměř nikdy nevyužívají okrajové služby.

Druhý, nejmenší shluk popisuje skupinu klientů převážně malých bank, kteří navštěvují pobočku své banky. Z bankomatu vybírají často a vysoké částky, průměrně 3 000 Kč. Tato skupina má nejvyšší počet příchozích plateb, zůstatek a obrat. Zhruba třetina klientů má také debetní embosovanou kartu. Klienti této skupiny často využívají okrajové služby.

Do třetího shluku patří klienti malých bank, kteří nenavštěvují pobočku své banky. V jedné třetině případů tito klienti mají spolu s kreditní kartou také debetní embosovanou kartu. Z bankomatů vlastní banky nevybírají často a z bankomatů cizích bank vybírají částky přes 2 000 Kč. Zůstatek i obrat se pohybují okolo 20 000 Kč. Tito klienti zřídka používají okrajové služby.

Čtvrtý shluk zahrnuje klienty velkých bank, kteří navštěvují pravidelně pobočku pro vklad peněz na účet. Klienti mají nízký měsíční obrat a zůstatek pohybující se okolo 15 000 Kč. Peníze vybírají především z bankomatů vlastní i cizí banky. Průměrný výběr je zhruba 1 000 Kč. Klienti nadprůměrně využívají okrajových služeb.

Pátý shluk obsahuje klienty velkých bank, na rozdíl od prvního shluku však tito klienti v téměř polovině případů navštěvují pobočku své banky pro vklad peněz na účet. Všichni klienti mají k účtu embosovanou debetní kartu. Ze všech skupin nejčastěji vybírají částky okolo 1 500 Kč z bankomatu své banky. Téměř ve čtvrtině případů klienti využívají službu Cashback.

6.2 Klasifikace klientů podle využívání okrajových služeb

V této úloze byl použit celý dataset, analýza podskupin neukázala významné rozdíly. Některé výstupy byly shledány nevýznamné, a proto zde jejich popis není. První tabulka obsahuje souhrn počtu použitých a vyřazených případů. Výstup také obsahuje tabulku s průměry a směrodatnými odchylkami proměnných v jednotlivých skupinách závislé proměnné. Tabulky č. 7-9 ukazují výsledky testů ANOVA pro jednotlivé závislé proměnné. Tento test je proveden před vznikem modelu. Nejvýznamnější údaje v těchto tabulkách jsou sloupce s hodnotami Wilksova lambda a P-hodnotami. Čím nižší je hodnota Wilksova lambda pro danou proměnnou, tím významnější je její vztah se závislou proměnnou. Pro P-hodnoty platí, že pokud je nižší než stanovená hladina významnosti, zamítá se nulová hypotéza a proměnná je statisticky významná. [11] [13]

Analýza ukázala, že šest nezávislých proměnných je statisticky významných na zvolené hladině významnosti 0,05 ve všech třech modelech. U těchto proměnných byla prokázána významná závislost četnosti využití daných bankovních služeb na využití zvolených okrajových služeb. Proto byla zamítnuta definovaná nulová hypotéza a přijata hypotéza alternativní.

Tabulka 7 Test shody skupinových průměrů pro Cashback

	Wilks' Lambda	F	df1	df2	Sig.
Obrat_LogN	1,000	5,229	1	13945	,022
ATM.vlastni_LogN	,997	41,526	1	13945	,000
ATM.cizi.castka_sqrt	,979	304,711	1	13945	,000
Prich.platba.pocet_LogN	,992	108,695	1	13945	,000
TPU.pobocka_LogN	,999	17,197	1	13945	,000
SIPO.celkem_LogN	,981	272,188	1	13945	,000
Uhrada.telefon_LogN	1,000	4,803	1	13945	,028
Uhrada.box_LogN	1,000	2,609	1	13945	,106
Zustatek_LogN	1,000	1,719	1	13945	,190
ATM.cizi_trans	,994	86,406	1	13945	,000

Zdroj: Vlastní zpracování

Tabulka 8 Test shody skupinových průměrů pro nadměrný vklad

	Wilks' Lambda	F	df1	df2	Sig.
Obrat_LogN	1,000	,386	1	14159	,535
ATM.vlastni_LogN	,999	20,024	1	14159	,000
ATM.cizi.castka_sqrt	,989	158,344	1	14159	,000
Prich.platba.pocet_LogN	,984	228,001	1	14159	,000
TPU.pobocka_LogN	,998	27,550	1	14159	,000
SIPO.celkem_LogN	,991	126,465	1	14159	,000
Uhrada.telefon_LogN	1,000	5,765	1	14159	,016
Uhrada.box_LogN	,999	7,159	1	14159	,007
Zustatek_LogN	,997	39,780	1	14159	,000
ATM.cizi_trans	,992	115,382	1	14159	,000

Zdroj: Vlastní zpracování

Tabulka 9 Test shody skupinových průměrů pro výběr v zahraničí

	Wilks' Lambda	F	df1	df2	Sig.
Obrat_LogN	,994	101,265	1	15966	,000
ATM.vlastni_LogN	,983	282,473	1	15966	,000
ATM.cizi.castka_sqrt	,929	1217,505	1	15966	,000
Prich.platba.pocet_LogN	,978	351,457	1	15966	,000
TPU.pobocka_LogN	,998	31,949	1	15966	,000
SIPO.celkem_LogN	,982	288,651	1	15966	,000
Uhrada.telefon_LogN	1,000	,854	1	15966	,355
Uhrada.box_LogN	1,000	2,120	1	15966	,145
Zustatek_LogN	,996	60,504	1	15966	,000
ATM.cizi_trans	,954	770,134	1	15966	,000

Zdroj: Vlastní zpracování

Další částí úlohy bylo nalezení nejpřesnějšího modelu pro klasifikaci klientů do skupin. Jako závislá proměnná byla zvolena okrajová služba Cashback. Nejdříve byla ověřena platnost předpokladu o shodě kovariančních matic nezávislých proměnných napříč skupinami závislé proměnné pomocí Boxova M-testu. Výsledek testu je v tabulce č. 10. Výsledná P-hodnota $p<0,001$ zamítla nulovou hypotézu. Test tedy tento předpoklad nepotvrdil, použití meziskupinových matic ale nepřineslo zlepšení. Pro úlohu proto byla nakonec použita matice vnitroskupinových kovariancí. [11] [12] [13]

Tabulka 10 Výsledek Boxova M testu

Box's M		1534,478
F	Approx.	23,192
	df1	66
	df2	24625162,521
	Sig.	,000

Zdroj: Vlastní zpracování

Metoda Stepwise z deseti původních proměnných vybrala pro model osm. V tabulce č. 11 lze vidět, jak postupně byly proměnné vybírány. V každém z osmi kroků byla vybrána proměnná, která nejvíce snížila celkové Wilksovo lambda. [13]

Tabulka 11 Průběh metody stepwise

		Wilks' Lambda							
						Exact F			
Step	Entered	Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	ATM.cizi.castka_sqrt	,979	1	1	13945,	304,711	1	13945,0	,000
2	SIPO.celkem_LogN	,963	2	1	13945,	266,907	2	13944,0	,000
3	Obrat_LogN	,959	3	1	13945,	199,322	3	13943,0	,000
4	Prich.platba.pocet_LogN	,955	4	1	13945,	162,815	4	13942,0	,000
5	ATM.vlastni_LogN	,952	5	1	13945,	139,471	5	13941,0	,000
6	TPU.pobocka_LogN	,952	6	1	13945,	118,055	6	13940,0	,000
7	Uhrada.telefon_LogN	,951	7	1	13945,	101,828	7	13939,0	,000
8	ATM.cizi_trans	,951	8	1	13945,	89,607	8	13938,0	,000

Zdroj: Vlastní zpracování

Tabulka č. 12 standardizovaných koeficientů zobrazuje hodnoty, které byly použity k vypočítání diskriminačního skóre pro jednotlivé případy. Postup výpočtu je stejný jako např. u lineární regrese. Rozdělení skóre pro jednotlivé funkce je nastaveno tak, aby se průměr rovnal nule a směrodatná odchylka se rovnala jedné. Z těchto hodnot lze dále odvodit významnosti proměnných. Čím větší absolutní hodnotu má koeficient tím významnější vliv má příslušná proměnná. [17] [27]

Tabulka 12 Standardizované koeficienty diskriminační funkce

	Function
	1
Obrat_LogN	-,388
ATM.vlastni_LogN	,249
ATM.cizi.castka_sqrt	,602
Prich.platba.pocet_LogN	,249
TPU.pobocka_LogN	-,130
SIPO.celkem_LogN	,586
Uhrada.telefon_LogN	,079
ATM.cizi_trans	-,085

Zdroj: Vlastní zpracování

Tabulka č. 13 obsahuje vnitroskupinové korelace všech nezávislých proměnných s kanonickou funkcí. Tyto hodnoty nejsou na rozdíl od předchozí tabulky ovlivněny kolinearitou takže přesně ukazují proměnné s vysokou diskriminační schopností. [17] [27]

Tabulka 13 Tabulka korelací s kanonickou funkcí

	Function
	1
ATM.cizi.castka_sqrt	,652
SIPO.celkem_LogN	,616
Prich.platba.pocet_LogN	,389
ATM.cizi_trans	-,347
ATM.vlastni_LogN	,241
TPU.pobocka_LogN	-,155
Obrat_LogN	-,085
Uhrada.telefon_LogN	,082
Uhrada.box_LogN[a]	-,039
Zustatek_LogN[a]	,009

Zdroj: Vlastní zpracování

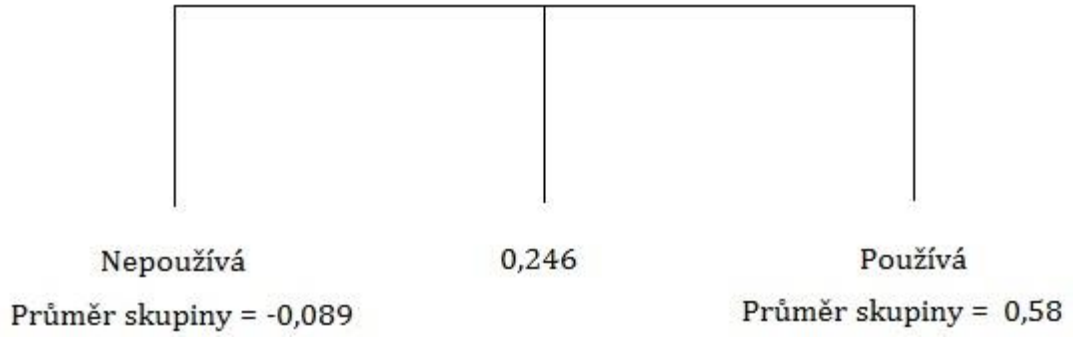
V tabulce č. 14 jsou zobrazeny průměry hodnot diskriminační funkce (zvané také centroidy) v daných skupinách. Samotné zařazení případů do skupin poté závisí na tom, jestli je diskriminační skóre daného případu blíže hodnotě 0,58 pro zařazení do skupiny klientů kteří službu Cashback využívají nebo hodnotě -0,089 pro zařazení do skupiny klientů kteří Cashback nevyužívají. V praxi se

pro klasifikaci vypočítá průměr těchto dvou hodnot. V tomto modelu průměr vychází 0,246. Obrázek č. 18 ukazuje, že případy s hodnotou diskriminačního kritéria nad 0,246 jsou zařazeny do skupiny Ano a nižší do skupiny Ne. [17] [27]

Tabulka 14 Středy skupin závislé proměnné

	Function
Cashback.flag	1
Ne	-,089
Ano	,580

Zdroj: Vlastní zpracování



Obr. 18 Klasifikace podle průměrné hodnoty
Zdroj: Vlastní zpracování

Tabulka č. 15 ukazuje charakteristiky získané diskriminační funkce. Kanonická korelace měří vztah mezi diskriminačními skóre a skupinami. Hodnota kanonické korelace je 0,221. Čím větší je získané vlastní číslo, tím více variace závislé proměnné lze funkcí vysvětlit. Hodnota vlastního čísla vypočtené funkce je 0,051. [17] [27]

Tabulka 15 Vlastnosti diskriminační funkce

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,051[a]	100,0	100,0	,221

Zdroj: Vlastní zpracování

Tabulka č. 16 ukazuje hodnotu Wilksova lambda pro získanou funkci, které ukazuje část celkové variace nevysvětlenou rozdíly mezi skupinami. Hodnota Chi-square statistiky dosáhla hodnoty 699,2. Statistika testuje, zda se kanonická korelace diskriminační funkce rovná nule. Nulová hypotéza tvrdí, že funkce nemá diskriminační schopnost. Dále je zobrazen počet stupňů volnosti a výsledná P-hodnota. Protože P-hodnota vyšla 0,000 lze nulovou hypotézu zamítnout, takže funkci a potažmo model lze považovat za významnou pro diskriminaci. [17] [27]

Tabulka 16 Test diskriminační funkce

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,951	699,2	8	,000

Zdroj: Vlastní zpracování

Přesnost klasifikace modelu Cashback je v tabulce č. 17. Pro ukázku jsou zobrazeny i klasifikační tabulky č. 18-19 pro modely nadměrný vklad a výběr v zahraničí. Křížová validace vyšla ve všech třech modelech téměř stejně. Model lépe zařazuje klienty, kteří službu Cashback nepoužívají. Celková klasifikační přesnost modelu vypočtená pomocí váženého průměru se rovná 66%. [2]

Tabulka 17 Přesnost klasifikace modelu Cashback

			Predicted Group Membership		
		Cashback.flag	Ne	Ano	Total
Original	Count	Ne	8073	4048	12121
		Ano	688	1166	1854
		Ungrouped cases	1500	2565	4065
	%	Ne	66,6	33,4	100,0
		Ano	37,1	62,9	100,0
		Ungrouped cases	36,9	63,1	100,0
Cross-validated[b]	Count	Ne	8067	4054	12121
		Ano	693	1161	1854
	%	Ne	66,6	33,4	100,0
		Ano	37,4	62,6	100,0

Zdroj: Vlastní zpracování

Tabulka 18 Přesnost klasifikace modelu nadměrný vklad

			Predicted Group Membership		
		Nadmerny.flag	Ano	Ne	Total
Original	Count	Ano	369	228	597
		Ne	3841	9723	13564
		Ungrouped cases	1837	1990	3827
	%	Ano	61,8	38,2	100,0
		Ne	28,3	71,7	100,0
		Ungrouped cases	48,0	52,0	100,0
Cross-validated[b]	Count	Ano	364	233	597
		Ne	3845	9719	13564
	%	Ano	61,0	39,0	100,0
		Ne	28,3	71,7	100,0

Zdroj: Vlastní zpracování

Tabulka 19 Přesnost klasifikace modelu výběr v zahraničí

			Predicted Group Membership		
		ATM.vlastni.cizi.zahr.flag	Ano	Ne	Total
Original	Count	Ano	1930	881	2811
		Ne	3868	9289	13157
		Ungrouped cases	1449	571	2020
	%	Ano	68,7	31,3	100,0
		Ne	29,4	70,6	100,0
		Ungrouped cases	71,7	28,3	100,0
Cross-validated[b]	Count	Ano	1928	883	2811
		Ne	3869	9288	13157
	%	Ano	68,6	31,4	100,0
		Ne	29,4	70,6	100,0

. Zdroj: Vlastní zpracování

7 Shrnutí výsledků

Celkem ve dvou úlohách byla zanalyzována poskytnutá data o bankovních klientech. Postup zpracování úloh se řídil metodikou CRISP-DM, která se ukázala velmi užitečná. Často bylo třeba vrátit se k předcházejícím fázím, především k přípravě dat.

Ve třetí kapitole byly poskytnuty rady pro zkvalitnění internetových kalkulátorů. Zmíněné rady vychází zejména ze zkušeností získaných při přípravě dat.

V úloze Analýza skupin bankovních klientů byla zvolena dvoukroková shluková analýza. Analyzováni byli klienti, kteří vlastní kreditní kartu. Výsledným modelem byly případy rozděleny do pěti shluků. Kvalita modelu daná siluetou dosáhla hodnoty 0,4. Vzhledem k tomu, že data jsou původem z internetového kalkulátoru nakloněna směrem k uživatelům internetového bankovnictví, lze tuto hodnotu považovat za dostatečně vysokou. Celkem bylo použito devět proměnných, tři dichotomické a šest numerických. Jako nejvýznamnější se ukázaly dichotomické proměnné. Tři shluky obsahují klienty bank takzvané velké trojky, do které patří Česká spořitelna, ČSOB (Československá obchodní banka) a Komerční banka. Celkově se nedá říci, že by klienti vlastnící kreditní kartu preferovali větší banky.

V úloze Klasifikace klientů podle využívání okrajových služeb byla nejdříve vyvrácena zvolená nulová hypotéza na hladině významnosti 5%. Pro vytvoření klasifikačního modelu zvolena závislá proměnná Cashback. Získaný model dosáhl přesnosti klasifikace 66%. Stejnou přesnost modelu ukazuje i použití křížové validace. Model lze použít v bankovním prostředí, zejména pro cílenou reklamu. Mnoho potenciálních klientů o určitých okrajových službách neví, přitom existuje vysoká šance, že o ně budou mít zájem. Přesnost klasifikačních modelů pro okrajové služby nadměrný vklad na pobočce a výběr z bankomatu v zahraničí dosáhla 71,2% respektive 70,2%. Modely přesněji klasifikují klienty, kteří okrajové služby nevyužívají, rozdíl je však malý.

8 Závěry a doporučení

Získané modely vychází z dat, které jsou nakloněna klientům používající internet, takže trénovací data nereflektují skutečnou populaci bankovních klientů. Přesto však mohou přinést užitečné informace o určitých skupinách, například o klientech vlastnící kreditní kartu. Většina numerických proměnných i přes použití transformací nesplňuje kompletně předpoklady normality. V případě shlukové analýzy, používající princip Listwise bylo potřeba kvůli nedostatku kompletních případů nahradit chybějící hodnoty. Vhodné je zvážení zmíněných rad ohledně úpravy existujícího kalkulátoru a případně návrhu dotazníku, který by byl konkrétně zaměřen na získání dat reflektující skutečnou populaci klientů retailového bankovnictví. Tím by se razantně urychlila příprava dat a odpadly problémy s interpretací chybějících hodnot. Pokud by dotazník obsahoval i základní demografické údaje pak by data umožňovala i některé další analýzy.

Výsledek shlukové analýzy je uspokojivý. Hodnota siluety téměř dosáhla hodnoty 0,5. Významnosti jednotlivých proměnných ukazují, že u numerických proměnných nejsou průměrné hodnoty ve shlucích velice rozdílné, což způsobilo nárůst významnosti dichotomických proměnných. Problémem je zejména původ dostupných dat. Lidé s připojením k internetu, kteří se zajímají o výši bankovních poplatků, lze považovat za potenciální skupinu v celkové populaci klientů českých bank. Proto lze předpokládat, že i jejich bankovní charakteristiky budou podobné. I tak však z analýzy vyplynuly některé odlišnosti, především ve výši částky vybírané z bankomatu cizí banky a měsíčního zůstatku.

V případě diskriminační analýzy dosažené výsledky bezpečně vyvracejí stanovenou nulovou hypotézu. Klienti s vyšší mírou využití standartních bankovních služeb také častě využívají služby okrajové. Přesnost klasifikace modelů dosáhla hodnot okolo 70%.

Budoucí analýzy mohou zahrnovat použití logistické regrese. Ta se podobá diskriminační analýze a umožňuje také použití nezávislých kategoriálních proměnných. V praxi se často využívá obou metod a výsledky jsou porovnány. V případě shlukové analýzy lze doporučit použití uzlu dvoukrokové analýzy AS,

který umožňuje pokročilejší nastavení. Pro některé specifické úlohy lze také doporučit použití rozhodovacích stromů a neuronových sítí.

9 Seznam použité literatury

[1] An Introduction to Data Mining [online]. 2010 [cit. 2016-04-10]. Dostupné z: http://www.saedsayad.com/data_mining_map.htm

[2] Annotated SPSS Output Discriminant Analysis. Institute for Digital Research and Education [online]. California, 2015 [cit. 2016-04-10]. Dostupné z: http://www.ats.ucla.edu/stat/spss/output/SPSS_discrim.htm

[3] Bankovnipoplatky.com: Internetový ombudsman bankovních klientů [online]. Česká republika: Digitaiment, 2005 [cit. 2016-04-10]. Dostupné z: http://www.bankovnipoplatky.com/

[4] BELAS, Jaroslav, Eva CIPOVOVA a Valer DEMJAN. CURRENT TRENDS IN AREA OF SATISFACTION OF BANK CLIENTS IN THE CZECH REPUBLIC AND SLOVAKIA. Transformation in Business [online]. 2014, 13(3), 219-234 [cit. 2016-04-10]. ISSN 16484460.

[5] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003. ISBN 80-200-1062-9.

[6] BERRY, Michael J a Gordon LINOFF. Data mining techniques for marketing, sales, and customer relationship management. 2nd ed. Indianapolis, Ind.: Wiley, 2004. ISBN 0-471-47064-3.

[7] CHAPMAN, Pete, Julian CLINTON, Randy KERBER, Thomas KHABAZA, Thomas REINARTZ, Colin SHEARER a Rüdiger WIRTH. CRISP-DM 1.0: Step-by-step data mining guide. [online]. 2000 [cit. 2016-03-08]. Dostupné z: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[8] CHITRA, K. a B. SUBASHINI. Data Mining Techniques and its Applications in Banking Sector. International Journal of Emerging Technology and Advanced Engineering [online]. 2013, 3(8), 8 [cit. 2016-04-10]. ISSN 2250–2459. Dostupné z: http://www.ijetae.com/files/Volume3Issue8/IJETAE_0813_35.pdf

[9] ČERNOHORSKÁ, Liběna, Jan ČERNOHORSKÝ a Petr TEPLÝ. THE BANKING STABILITY IN THE CZECH REPUBLIC BASED ON DISCRIMINANT AND CLUSTER ANALYSES. Anadolu University Journal of Social Sciences [online]. 2007, 7(2), 85-96 [cit. 2016-04-10]. ISSN 13030876.

[10] DATA MINING IN BANKING AND FINANCE: A NOTE FOR BANKERS. CiteSeerX [online]. Pennsylvania, 2007 [cit. 2016-04-10]. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.102.7502

[11] HEBÁK, Petr, Jiří HUSTOPECKÝ, Eva JAROŠOVÁ a Iva PECÁKOVÁ. Vícerozměrné statistické metody (1). Praha: Informatorium, 2004. ISBN 80-7333-025-3.

[12] HEBÁK, Petr, Jiří HUSTOPECKÝ, Iva PECÁKOVÁ, Milan PRŮŠA, Hana ŘEZANKOVÁ, Petr VLACH a Alžběta SVOBODOVÁ. Vícerozměrné statistické metody (3). Praha: Informatorium, 2005. ISBN 80-7333-039-3.

[13] IBM Knowledge Center [online]. IBM [cit. 2016-04-10]. Dostupné z: http://www.ibm.com/support/knowledgecenter/

[14] IBM SPSS Modeler 17.0 Documentation: CRISP-DM Guide [online]. U.S.A.: IBM Corporation, 2011 [cit. 2016-04-10]. Dostupné z: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

[15] IBM SPSS Modeler 17.0 Documentation: Modeling Nodes [online]. U.S.A.: IBM Corporation, 2015 [cit. 2016-04-10]. Dostupné z: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.0/en/ModelerModelingNodes.pdf

[16] Kurzycz: Kurzy měn, akcie, komodity, online zpravodajství [online]. AliaWeb, 2000 [cit. 2016-04-16]. Dostupné z: http://www.kurzy.cz/

[17] Laerd statistics: We make statistics easy [online]. Lund Research, 2013 [cit. 2016-04-10]. Dostupné z: https://statistics.laerd.com/

[18] MANN, Bikram Jit Singh a Sunpreet Kaur SAHNI. Profiling Adopter Categories of Internet Banking in India: An Empirical Study. Vision (09722629) [online]. 2012, 16(4), 283-295 [cit. 2016-04-10]. DOI: 10.1177/0972262912460187. ISSN 09722629.

[19] MELOUN, Milan a Jiří MILITKÝ. Přednosti analýzy shluků ve vícerozměrné statistické analýze [online]. Česká republika: 2 Theta, 2004 [cit. 2016-04-10]. ISBN 80-86380-22-X. Dostupné z: http://meloun.upce.cz/docs/publication/152.pdf

[20] MEMIĆ, Deni. ASSESSING CREDIT DEFAULT USING LOGISTIC REGRESSION AND MULTIPLE DISCRIMINANT ANALYSIS: EMPIRICAL EVIDENCE FROM BOSNIA AND HERZEGOVINA. Interdisciplinary Description of Complex Systems [online]. 2015, 13(1), 128-153 [cit. 2016-04-10].DOI: 10.7906/indecs.13.1.13. ISSN 13344684.

[21] MOIN, Kazi Imran a Qazi Baseer AHMED. Use of Data Mining in Banking. International Journal of Engineering Research and Applications [online]. 2012, 2(2), 5 [cit. 2016-04-10]. ISSN 2248-9622. Dostupné z: http://www.ijera.com/papers/Vol2_issue2/DU22738742.pdf

[22] RENCHER, Alvin C. Methods of multivariate analysis. 2nd ed. Hoboken: Wiley-interscience, 2001. Wiley series in probability and statistics. ISBN 0-471-41889-7.

[23] ŞCHIOPU, Daniela. Applying Two-Step Cluster Analysis for Identifying Bank Customers' Profile. Petroleum Gas University of Ploiesti Bulletin, Economic Sciences Series [online]. 2010, 62(3), 66-75 [cit. 2016-04-10]. ISSN 12246832.

[24] SHOBANA, V. K. a G. SHANTHI. Profitability of Foreign Banks Operating in India: A Multi-Discriminant Model. IUP Journal of Bank Management [online]. 2010, 9(1/2), 21-27 [cit. 2016-04-10]. ISSN 09726918.

[25] SWICEGOOD, Philip a Jeffrey A. CLARK. Off-site Monitoring Systems for Predicting Bank Underperformance: A Comparison of Neural Networks, Discriminant Analysis, and Professional Human Judgment. International Journal of Intelligent Systems in Accounting Finance [online]. 2001, 10(3), 169-186 [cit. 2016-04-10]. DOI: 10.1002/isaf.201. ISSN 1055615X.

[26] TRPKOVA, Marija a Dragan TEVDOVSKI. APPLIED DISCRIMINANT ANALYSIS IN ESTIMATION OF POTENTIAL EU MEMBERS. Young Economists Journal / Revista Tinerilor Economisti [online]. 2010, 8(15), 135-147 [cit. 2016-04-10]. ISSN 15839982.

[27] YAZICI, Mehmet. COMBINATION OF DISCRIMINANT ANALYSIS AND ARTIFICIAL NEURAL NETWORK IN THE ANALYSIS OF CREDIT CARD CUSTOMERS. European Journal of Finance [online]. 2011, 4(4), 1-10 [cit. 2016-04-10]. ISSN 19333420.

[28] Y. SAHIN a E. DUMAN. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. Lecture Notes in Engineering and Computer Science [online]. 2011, 2188(1), 442 [cit. 2016-04-10]. ISSN 20780958.

10 Přílohy

1) Tabulka proměnných připraveného datasetu

Proměnná	Použitá zkratka	Typ	Rozmezí hodnot
Data o připojení			
Datum	Datum	Kategoriální	2010 - 2014
Účet			
Min. měsíční obrat	Obrat	Numerická	20 - 1 000 000
Průměr. Měsíční zůstatek	Zustatek	Numerická	0 - 1 000 000
Výpis z účtu			
Forma výpisu	Vypis.Forma	Kategoriální	0 – Elektronicky 1 – Poštou 2 - Telefonicky
Frekvence výpisu	Vypis.Frekv	Kategoriální	1 – Týdně 2 – Měsíčně 3 – Čtvrtletně 4 - Ročně
Karta			
Debet neemboss	Debet.neem	Kategoriální	0 – Ne 1 – Ano
Debet emboss	Debet.em	Kategoriální	0 – Ne 1 – Ano
Kreditní	Kreditni	Kategoriální	0 – Ne 1 – Ano
Výběr z ATM vlastní banky	ATM.vlastni	Numerická	0 - 30
Výběr z ATM cizí banky	ATM.cizi	Numerická	0 - 30
Výběr z ATM cizí banky - částka	ATM.cizi.castka	Numerická	0 - 50000
Výběr z ATM vlastní banky v zahraničí	ATM.vlastni. zahr	Numerická	0 - 30
Výběr z ATM vlastní banky v zahraničí - částka	ATM.vlastni. zahr.castka	Numerická	0 - 50000
Výběr z ATM cizí banky v zahraničí	ATM.cizi.zahr	Numerická	0 - 30

Výběr z ATM cizí banky v zahraničí - částka	ATM.cizi.zahr.castka	Numerická	0 - 50000
Přímé bankovnictví			
Přímé bankovnictví	Prime.bank	Kategoriální	0 – Ne 1 – Ano
Internetbanking	Inter.bank	Kategoriální	0 – Ne 1 – Ano
Telebanking	Tele.bank	Kategoriální	0 – Ne 1 - Ano
Platby - jednorázové			
Příchozí platba z vlastní a cizí banky - počet	Prich.platba.pocet	Numerická	0 - 200
příkaz k úhradě do vlastní a cizí banky - pobočce	Uhrada.pobocka	Numerická	0 - 100
příkaz k úhradě do vlastní a cizí banky - Po telefonu	Uhrada.telefon	Numerická	0 - 30
příkaz k úhradě do vlastní a cizí banky - sběrný box	Uhrada.box	Numerická	0 - 40
příkaz k úhradě do vlastní a cizí banky - internetbanking	Uhrada.interbank	Numerická	0 - 200
Platby - trvalé příkazy (TPÚ)			
do vlastní a cizí banky - pobočce	TPU.pobocka	Numerická	0 - 20
do vlastní a cizí banky - Po telefonu	TPU.telefon	Numerická	0 - 20
do vlastní a cizí banky - internetbanking	TPU.interbank	Numerická	0 - 40
Platby - povolení k inkasu (včetně SIPO)			
do vlastní a cizí banky - Po telefonu2	SIPO.pobocka	Numerická	0 - 20
do vlastní a cizí banky - internetbanking2	SIPO.telefon	Numerická	0 - 10

do vlastní a cizí banky - pobočce	SIPO.interbank	Numerická	0 – 30
Hotovostní operace			
Vklad účet pobočce - počet	Vklad.poboc.pocet	Numerická	0 - 20
Nadměrný vklad - počet	Nadmerny.pocet	Numerická	0 - 10
Nadměrný vklad - částka	Nadmerny.castka	Numerická	0 - 1000 000
Vklad účet přes bankomat - počet	Vklad.bankomat. pocet	Numerická	0 - 10
Výběr pobočce - počet	Vyber.poboc. pocet	Numerická	0 - 20
Služba cashback - počet	Cashback.pocet	Numerická	0 - 20
Aktuální užívaná banka	Banka	Kategoriální	1 - 17
Aktuální užívaný účet	Ucet	Kategoriální	2 - 113
Nové proměnné			
Vklad účet pobočce - flag	Vklad.poboc.flag	Kategoriální	0 – Ne 1 – Ano
Nadměrný vklad - flag	Nadmerny.flag	Kategoriální	0 – Ne 1 – Ano
Vklad účet přes bankomat - flag	Vklad.bankomat. flag	Kategoriální	0 – Ne 1 – Ano
Výběr pobočce - flag	Vyber.poboc. flag	Kategoriální	0 – Ne 1 – Ano
Služba cashback - flag	Cashback.flag	Kategoriální	0 – Ne 1 – Ano
Banky podle aktiv	Banka.aktiva	Kategoriální	Tabulka č. 3
Banky podle bankomatů	Banka.bankomaty	Kategoriální	Tabulka č. 3
Banky podle poboček	Banka.pobocka	Kategoriální	Tabulka č. 3
Celkový počet úhrad	Uhrada.celkem	Kategoriální	0 – 204
Celkový počet trvalých příkazů	TPU.celkem	Kategoriální	0 – 47
Celkový počet SIPO	SIPO.celkem	Kategoriální	0 – 30
Výběr z bankomatu v zahraničí – flag (vlastní nebo cizí banky)	ATM.vlastni.cizi.zahr .flag	Kategoriální	0 – Ne 1 – Ano

Univerzita Hradec Králové
Fakulta informatiky a managementu
Akademický rok: 2015/2016

Studijní program: Systémové inženýrství a informatika
Forma: Prezenční
Obor/komb.: Informační management (im2-p)

Podklad pro zadání DIPLOMOVÉ práce studenta

PŘEDKLÁDÁ:	ADRESA	OSOBNÍ ČÍSLO
Snášel Norbert	Lužec nad Cidlinou 23, Lužec nad Cidlinou	I1447

TÉMA ČESKY:

Data mining: Analýza klientů retailového bankovnictví

TÉMA ANGLICKY:

Data mining: Analysis of retail banking clients

VEDOUCÍ PRÁCE:

Mgr. Jan Draessler, Ph.D. - KIKM

ZÁSADY PRO VYPRACOVÁNÍ:

Cílem je pomocí vybraných nástrojů analyzovat data týkající se využívání služeb běžných účtů bank v ČR.

Osnova:

1, Data mining v business sféře
2, Popis a výběr analýz
3, Popis a příprava dat
4, Vybrané analýzy business dat
5, Interpretace výstupů analýz
6, Shrnutí a dosažené výsledky

SEZNAM DOPORUČENÉ LITERATURY:

BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.

HAN, Jiawei a Micheline KAMBER. Data mining: concepts and techniques. 2nd ed. London: Elsevier, c2006, xviii, 770 p. Morgan Kaufmann series in data management systems. ISBN 978-1-55860-901-3.

Software SPSS: Software a řešení pro prediktivní analýzu [online]. [cit. 2015-10-03]. Dostupné z: http://www-01.ibm.com/software/cz/analytics/spss/

The ultimate IBM? SPSS? Statistics guides. [online]. [cit. 2015-10-03]. Dostupné z: https://statistics.laerd.com/

Podpis studenta: Snášel ..... **Datum:** 7.10.2015

Podpis vedoucího práce: ..... **Datum:** 7.10.2015