Mendelova univerzita v Brně
Provozně ekonomická fakulta

Aplikace dataminingových metod na bankovní data

Diplomová práce

Vedoucí práce:

Ing. Naděžda Chalupová, Ph.D.

Bc. Miloš Melichar

Brno 2015

Poděkování

Zde bych chtěl poděkovat Ing. Naděžde Chalupové, Ph.D. za vedení mojí diplomové práce, připomínky a věnovaný čas. Také bych chtěl poděkovat rodičům za podporu a všem ostatním, kteří mi byli oporou během psaní práce.

Čestné prohlášení

Prohlašuji, že jsem tuto práci: **Aplikace dataminingových metod na bankovní data** vypracoval/a samostatně a veškeré použité prameny a informace jsou uvedeny v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů, a v souladu s platnou *Směrnicí o zveřejňování vysokoškolských závěrečných prací.*

Jsem si vědom/a, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 Autorského zákona.

Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity o tom, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

V Brně dne 30. prosince 2015 _____

Abstract

Melichar, M. Application of data mining techniques on bank data. Diploma thesis. Brno: Mendel University, 2015.
The thesis deals with pre-processing of two data sets with information on clients, loans and debit cards. The data sets were separately pre-processed and modeled by SPSS Modeler using a number of methods and algorithms. For the modeling purposes, three classification data mining tasks were defined: loan approving or rejecting, loan rating and debit card type assignment. By using the selected methods of machine learning techniques the classification models were built for each task. Models accuracy was tested by script written in SPSS language for automation. All tasks were supplemented by clustering technique based on latent factors gained by factor analysis. Factor analysis combined with clustering presents another approach in pattern discovery.

Keywords

SPSS Modeler, bank dataset, peer-to-peer loans, loan classification, machine learning techniques, clustering, factor analysis.

Abstrakt

Melichar, M. Aplikace dataminingových metod na bankovní data. Diplomová práce. Brno: Mendelova univerzita v Brně, 2015.
Práce se zabývá procesem přípravy dvou datových množin, které dohromady obsahovaly informace o klientech, úvěrech a kartách. Datové množiny byly odděleně předzpracovány a modelovány pomocí softwaru SPSS Modeler, který obsahuje množství metod a algoritmů pro podporu obou těchto fází. Pro modelování byly definovány tři klasifikační dataminingové úlohy: schvalování či zamítání úvěrů, jejich rating a přidělování vhodného typu debetní karty. Pro každou úlohu byly pomocí vybraných metod strojového učení vytvořeny klasifikační modely, jejichž přesnosti byly testovány s použitím skriptovacího jazyka v SPSS. Problematika řešených úloh byla doplněna o metodu shlukování pomocí skrytých faktorů vytvořených pomocí faktorové analýzy. Shlukování v kombinaci s faktorovou analýzou představuje další přístup v poznávání vzorů v datech v kontextu řešení zkoumaného jevu.

Klíčová slova

SPSS Modeler, bankovní data, nebankovní půjčky, klasifikace úvěrů, metody strojového učení, shlukování, faktorová analýza.

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod a cíl práce

1.1 Úvod

Data mining se používá ve finančnictví řadu let, a právě banky byly jeho průkopníky v zapojení do praxe. Principy dolování z dat zde hrají roli např. při tvorbě modelů stavějících na množství informací o klientech. Tyto modely pak obvykle slouží k automatizaci rozhodování, aniž by bylo ve větší míře potřeba lidského faktoru. Typickým příkladem zautomatizování takového procesu je žádost klienta o úvěr. Po získání parametrů požadovaného úvěru jej model porovná s historickými úvěry, u kterých je znám stav a pokusí se určit, zda by mohl být splacen. V tomto ohledu jsou dolovací algoritmy pravděpodobně člověkem nepřekonatelné. Jsou totiž schopné najít v datech takové vzory, které člověk nemá šanci postřehnout. Zůstanou tedy opomenuty, čímž se automaticky „dopouští chyby“ při rozhodování o poskytnutí půjčky. Tento způsob využití dolování spadá do oblasti řízení rizik v bankách. Rizikem se zde rozumí potenciální zisk či ztráta, které vznikne v případě splacení či nesplacení úvěru. Cílem je obvykle udržovat hladinu problémových úvěrů co nejnižší. Výsledný model tedy musí být vytvořen tak, aby byl schopen tento trend udržet. S rizikem obvykle souvisí potřeba určovat jeho míru, tedy pravděpodobnost s jakou bude úvěr splacen v případě schválení. Míru rizika lze vyjádřit pomocí úrokové míry.

Pokud je hlavním cílem zautomatizování nějakého rozhodovacího procesu, obvykle víme, co je jeho součástí a jaké proměnné zahrnout. V takovém případě usilujeme o vytvoření co nejpřesnějšího modelu, u kterého nás nemusí příliš zajímat, jak jsou nastavena jeho vnitřní kritéria pro určování. Znamená to tedy, že můžeme upřednostnit model algoritmu, u kterého dosáhneme vyšší přesnost i za cenu toho, že se nedozvíme, jak k výsledku došel. Tento typ úlohy lze použít v případě, kdy rozhodujeme o žádosti na základě iniciativy samotného klienta.

Na druhou stranu však banky často vyvíjí vlastní aktivitu v kontaktu se zákazníky. To se týká úloh, kdy nemáme přesně stanovený cíl či kritérium jeho plnění. Požadavkem může být například nalezení skupiny klientů se stejnou vlastností a jejich vymezení v určitých dimenzích. V takovém případě by nás mělo zajímat, jaká propojení se v datech nachází, abychom nalezli optimální kombinaci dimenzí pro vyobrazení skupin. Jednotlivé skupiny si pak můžeme sami popsat a oslovit je s konkrétní nabídkou produktu. Tento typ úlohy lze také zkombinovat s předchozím problémem. Pokud je žádost klienta o daný produkt zamítnuta, může se banka podívat na příslušnost klienta do jedné ze skupin, podle které může nabídnout vhodnější variantu.

1.2 Cíl práce

Cílem diplomové práce je netriviální zpracování vybraných datových množin a následné modelování, jehož výsledky mají být použitelné v bankovní praxi. V případě analýzy více datových množin, bude formulována společná dolovací úloha řešící podobný problém. Data budou dále zkoumána i nezávisle na sobě pro další možnosti využití. Příprava dat a modelování se provede ve zvoleném dataminingovém nástroji. V závěru budou zhodnoceny dosažené výsledky spočívající v porovnání vybraných metod použitých u jednotlivých úloh a budou shrnuty další poznatky.

1.3 Metodika

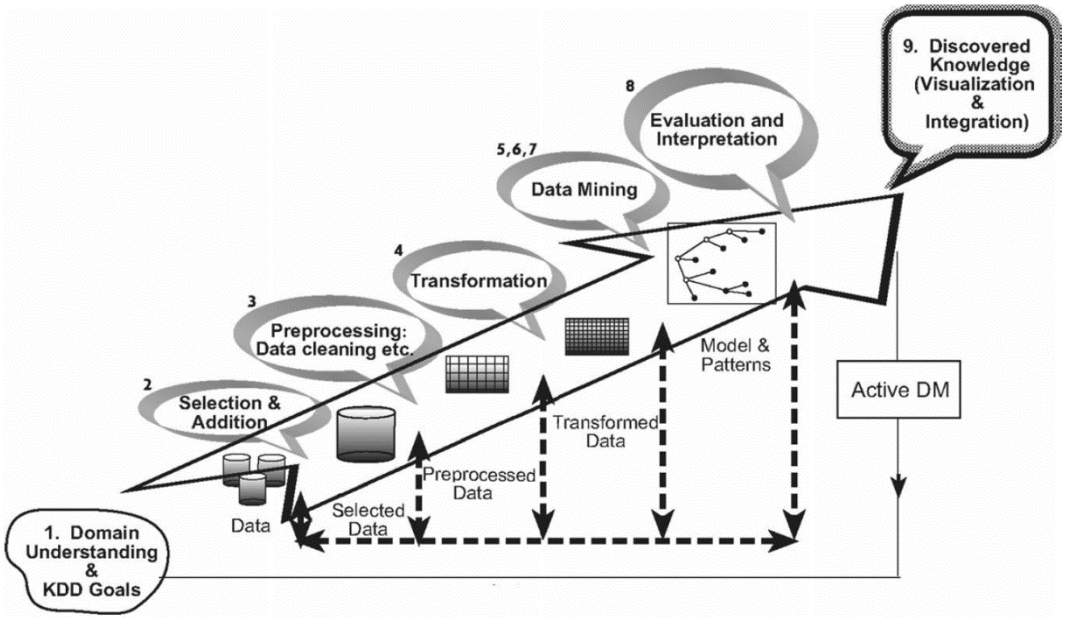
V teoretické části práce je provedena rešerše relevantních zdrojů pro postupy získávání poznatků pomocí data miningu. Do této části jsou zahrnuty principy dolování znalostí, proč se používají a jaké jsou typické dolovací úlohy v běžné i bankovní praxi. Dále teoretická část obsahuje rešerši současně používaných metodologií pro zpracování dat, které zahrnují řadu obecných postupů pro přípravu dat. Další kapitola je věnována vybraným dolovacím či podpůrným metodám pro analýzu, zejména fungování algoritmů a práci s nimi.

Praktická část diplomové práce je vypracována s ohledem na zvolenou metodologii, kterou je CRISP-DM. Jedná se o nejpoužívanější metodologii pro práci s daty. Dvě vybrané datové množiny jsou předzpracovány do podoby použitelné pro modelování. Bankovní data nacházející se v relačních tabulkách je zapotřebí vhodně zkombinovat a odvodit nové atributy, zejména z transakčních údajů, které jsou poté sloučeny do jedné matice. Vzhledem k zaměření obou datových množin, jejichž obsahem jsou informace o klientech a poskytnutých úvěrech, je definována hlavní společná dolovací úloha, tedy klasifikace úvěrů, popř. shlukování podle stanoveného kritéria. Klasifikace bude spočívat v predikci, zda nový klient bude či nebude schopen úvěr splácet. Protože datové množiny pochází od jiných subjektů a obsahují odlišné atributy, jsou dále zkoumány odděleně nezávisle na sobě. Z tohoto důvodu jsou definovány další dolovací úlohy, přizpůsobené pro vybraná data. Pro naplnění cílů těchto úloh jsou zvoleny algoritmy vhodné pro klasifikaci, mezi které byly zařazeny rozhodovací stromy, neuronové sítě a SVM. Pro shlukovací úlohy je použit algoritmus k – means v kombinaci s faktorovou analýzou. Příprava dat a modelování byly provedeny v dataminingovém nástroji IBM SPSS Modeler. Důvodem pro výběr tohoto nástroje je předchozí zkušenost získaná v rámci studia, a také široký rozsah dostupných metod, které je možné vyzkoušet. Uživatelské rozhraní metod je přizpůsobeno pro jednotné používání, což urychluje seznamování se s vybranými algoritmy či postupy pro práci s daty, díky čemuž lze lépe pochopit podstatu procesu dolování znalostí. Vygenerované modely jsou testovány pomocí skriptovacího jazyka v SPSS.

2 Metodická východiska

2.1 Proces dobývání znalostí z databází

Objevování znalostí z databází (KDD) je automatická, průzkumná analýza dat a jejich modelování. Jedná se o proces identifikace validních, nových, užitečných a srozumitelných vazeb v rozsáhlých datových množinách. Data mining je jádrem KDD procesu, zahrnující použití příslušných algoritmů pro průzkum dat, vytvoření modelu pro objevování původně neznámých vzorů. Získaný model se používá pro pochopení jevů v datech, analýzu a predikci. Proces dobývání znalostí z databáze je iterativní a obecně sestává z devíti kroků. V každém kroku je možné vrátit se zpět, přičemž někdy je to přímo vyžadováno. Je potřeba porozumět celému procesu jako celku, ale také různým požadavkům a možnostem v jednotlivých krocích. Celý proces KDD je zachycen na obrázku níže. (Maimon, Rokach, c2005)



Obr. 1 KDD proces
Zdroj: Maimon, Rokach, c2005

Proces dolování začíná určením cílů projektu a končí aplikací objevené znalosti. V jednotlivých KDD procesech se při dosahování cílů prochází následujícími devíti kroky:

Porozumění problémové doméně. Jedná se o přípravný krok sloužící k pochopení celého problému souvisejícího s transformací dat, výběrem algoritmů, reprezentací výsledků, atd. Je třeba se důkladně obeznámit s definovaným cílem pro koncového uživatele a prostředím, ve kterém se bude celý proces odehrávat.

Výběr a tvorba datové množiny pro objevování znalosti. Po definování cílů by mělo být rozhodnuto, která data budou pro průzkum použita. Tento krok v sobě za-

hrnuje zajištění dostupných dat včetně dalších nezbytných kroků pro provedení daného průzkumu. Vybraná data jsou integrována do jedné množiny obsahující veškeré potřebné atributy, které by mohly hrát roli v procesu objevování znalosti. Na důkladnosti provedení toho kroku velice záleží, protože algoritmy se z vybraných dat učí a objevují v nich znalosti. Pokud některý z důležitých atributů chybí, může sestavený model dosahovat nesprávných výsledků. Obecně platí, že čím více atributů data obsahují, tím lépe.

Předzpracování a čištění dat. V této fázi se zvyšuje spolehlivost dat pomocí čištění, manipulace s chybějícími hodnotami a odstraňováním šumu a extrémů. Krok může zahrnovat použití různých statistických metod a dataminingových algoritmů. Pokud například bude zjištěno, že některý z atributů má nedostatečnou kvalitu nebo obsahuje mnoho chybějících hodnot, může být označen za výstupní proměnnou, jejíž hodnoty mohou být predikovány a doplněny.

Transformace dat. Tento krok v sobě zahrnuje metody pro dimenzionální redukci, jako je vyhledání relevantních atributů a jejich transformace (např. diskretizace numerických atributů). Tento krok bývá nejvíce kritický a zároveň je pro každý projekt specifický.

Výběr vhodných dataminingových úloh. V této fázi rozhodujeme, jakou dataminingovou úlohu nad daty provedeme, zda se bude jednat o klasifikaci, regresi nebo shlukování. Tato část nejvíce souvisí s definovanými cíli KDD procesu a samozřejmě na předchozích krocích. V podstatě rozlišujeme dva druhy úloh, tj. predikci a deskripci. Predikce bývá často označována jako metoda učení s učitelem, zatímco deskripce je učení bez učitele. Většina dataminingových technik je založena na induktivním učení, kde je explicitně nebo implicitně vytvářen model zobecněním na základě dostatečného množství trénovacích příkladů. Základním požadavkem je vytvoření natrénovaného modelu použitelného pro nově příchozí případy.

Výběr vhodných algoritmů. Tento krok v sobě zahrnuje výběr specifických metod pro nalezení závislostí. Je důležité zvážit, zda upřednostnit přesnost nad srozumitelností algoritmů, tedy zda například zvolit neuronovou síť místo rozhodovacího stromu. V této fázi je nutné seznámení se s podmínkami, pod kterými daný algoritmus pracuje nejlépe. Každý algoritmus má určité parametry a svůj způsob učení, jako např. počet kroků křížové validace nebo dělení dat na množinu trénovacích a testovacích instancí.

Nasazení algoritmů. V této fázi aplikujeme dataminingové algoritmy. Je možné, že pro dosažení uspokojivých výsledků bude nutné algoritmus spustit několikrát, pokaždé s jiným nastavením vstupních parametrů.

Zhodnocení modelu. Zhodnocujeme vydolované vazby v datech s ohledem na vytyčené cíle v prvním kroku. Ověřujeme, zda kroky provedené ve fázi předzpracování jsou dostatečné a vedly k dosažení cíle nebo je nutné se do fáze modelování vrátit a data upravit. Tento krok se zaměřuje na srozumitelnost a použitelnost vytvořeného modelu.

Využití objevené znalosti. Nyní je možné použít vydolované znalosti nebo je zakomponovat do podnikového systému pro další využití. (Maimon, Rokach, c2005)

2.1.1 Druhy dataminingových metod

Metodami se v tomto kontextu rozumí skupiny algoritmů sloužících k naplnění dataminingových úloh definovaných nad vybranými daty. U konkrétní úlohy se vybírá vhodný algoritmus nebo se vezmou z dané skupiny algoritmy všechny a jejich výsledky se porovnají.

Cílem dolovacích úloh bývá *predikce* nebo *deskripce*. Členění typů úloh není vždy jednoznačné, nicméně používá se základní rozdělení dle dvou zmíněných cílů. Dělení vychází z Maimona a Rokacha (c2005), kde je *predikce* dále rozdělena na *klasifikační* a *regresní* úlohy.

Prediktivní metody popisují jednu nebo více proměnných ve vztahu k ostatním. Občas se také nazývají jako asymetrické, učící se s učitelem nebo přímé metody. Na základě pravidel lze predikovat budoucí výsledky jedné nebo více odezev či cílových proměnných v kontextu s tím, co se stane s vysvětlující či vstupní proměnnou. Mezi tyto metody patří algoritmy strojového učení, jako např. neuronové sítě, rozhodovací stromy, ale také statistické metody, kam řadíme lineární a logistické regrese. (Giudici, 2003)

Během *predikce* dochází k určování dodatečných nebo chybějících hodnot daného atributu konkrétní instance. V *klasifikačních* úlohách se atributu přiřazuje příslušnost do určité třídy objektu. Tento atribut má diskrétní hodnoty. Naproti tomu v *regresních* úlohách se atributům na základě vytvořeného klasifikačního modelu dopočítají hodnoty spojité. *Klasifikace* se skládá ze dvou fází (Rychlý, 2003):

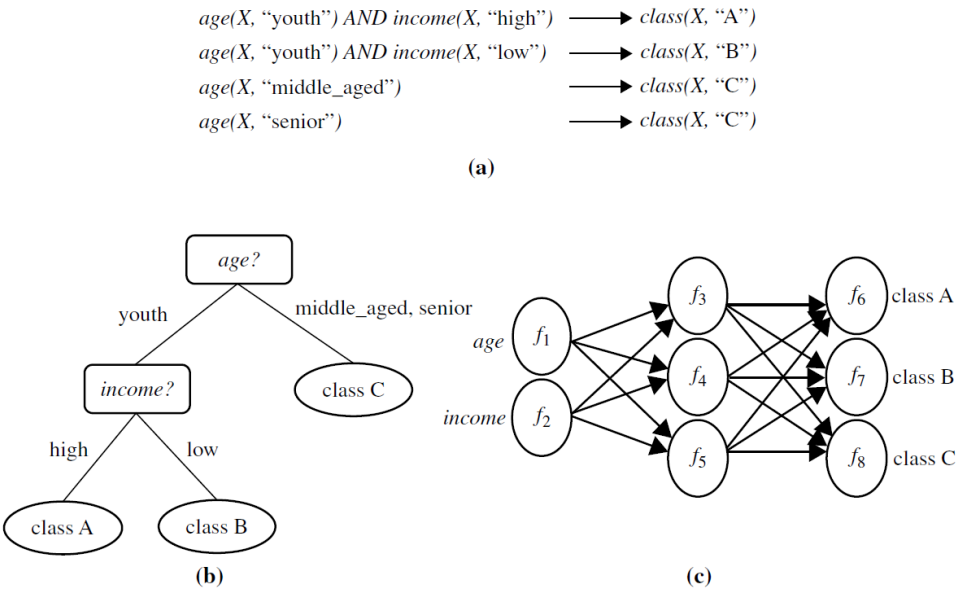
- *Fáze učení* – na základě trénovacích dat, u nichž je známa hodnota cílové proměnné (její třída), je vytvořen klasifikační model.
- *Vlastní klasifikace* – vytvořený model se použije ke klasifikaci nových instancí.

Zatímco predikce spíše předpovídá hodnoty atributů, *deskripce* slouží ke stručnému popisu skupin instancí nacházejících se v datech. Nazývají se také jako metody symetrické, učící se bez učitele nebo nepřímé metody. Zkoumané objekty bývají seskupeny bez předchozí znalosti o spojení, k čemuž se používají algoritmy pro shlukování nebo Kohonenovy mapy. Jednotlivé proměnné mohou být spojovány s druhými, aniž by tato spojení byla předem známa. Pro tento účel se také používají asociační pravidla, logistické lineární modely a modely grafické. V tomto případě jsou veškeré proměnné zkoumány rovnocenně a nejsou zde známé žádné příčinné souvislosti. (Giudici, 2003)

Ze základního dělení metod na prediktivní a deskriptivní vycházejí i Tan, Steinbach a Kumar (c2006) s tím rozdílem, že popisné modely rozdělují na tři další domény využívající různé algoritmy. Dohromady tedy mluvíme o čtyřech základních úlohách modelování, kterými jsou:

- prediktivní modelování (klasifikace a regrese)
- asociační pravidla
- shluková analýza
- detekce anomálií

Klasifikační modely mohou mít více podob, kdy jedna z možných reprezentací je ve formě *klasifikačních pravidel* (např. IF-THEN). Odvozovací pravidlo je tvořeno předpoklady, které se nacházejí na levé straně pravidla, kdy po jejich splnění je objektu přiřazena třída z pravé strany pravidla. Dále se pro predikci používají *rozhodovací stromy*. Rozhodovací strom je v podstatě vývojový diagram tvořený uzly, ve kterých se provádí test hodnoty atributu, kde větve představují výstupy těchto testů a listy nesou jednu nebo více tříd. Rozhodovací stromy mohou být jednoduše převedeny na klasifikační pravidla. *Neuronové sítě* se rovněž používají pro klasifikaci. Skládají se z umělých neuronů tzv. perceptronů, které jsou propojené s dalšími pomocí ohodnocených spojení. Pro klasifikaci se dále používají naivní Bayesovský klasifikátor, podpůrné vektory (SVM) nebo k - nejbližších sousedů. (Han, Kamber a Pei,c2012)



Obr. 2 Klasifikační modely: (a) klasifikační pravidla, (b) rozhodovací stromy, (c) neuronové sítě
Zdroj: Han, Kamber a Pei, c2012

Asociační analýza se používá pro odkrývání silně souvisejících vlastností v datech. Obvyklá forma objevených vzorů je implikace nebo podmnožina atributů. Vzhledem k exponenciálně narůstajícímu prohledávanému prostoru je úkolem těchto metod extrahovat co nejvíce významných vzorů. Asociační analýza se používá například pro nacházení skupiny genů s podobnými vlastnostmi, identifikace webových stránek často prohlížených společně nebo pro analýzu efektů ovlivňujících zemské klima. Velmi častou aplikací asociačních pravidel je analýza nákupního koše. (Tan, Steinbach a Kumar, c2006)

Shluková analýza pracuje na rozdíl od predikce s tzv. neoznačenými či nezatříděnými daty. Používá se pro vytvoření tříd skupin případů. Objekty bývají shlukovány podle principu maximální podobnosti v rámci jedné třídy a minimální podob-

nosti mezi různými třídami. Objekt z jedné třídy je svými vlastnostmi podobný objektu z té samé třídy, avšak má zcela jiné vlastnosti ve srovnání s objektem z jiné třídy. Následně je možné z takto vytvořených skupin objektů odvozovat pravidla. (Han, Kamber a Pei, c2012)

Detekce anomálií je založena na vyhledávání případů, které nejsou v souladu se zvyklostmi v daném souboru dat. Tyto objekty nazýváme jako odlehlé hodnoty neboli anomálie. Mnoho dolovacích metod považuje anomálie za šum nebo výjimku, avšak ve skutečnosti se může jednat například o klienta s podvodným jednáním. Detekce odlehlých hodnot se provádí pomocí statistických testů předpokládajících určité rozdělení pravděpodobnosti, případně měřením vzdálenosti, odkud objekty nacházející se mimo kterýkoli shluk jsou považovány za odlehlé. (Han, Kamber a Pei, c2012)

2.1.2 Využití dolovacích metod v souvisejících oblastech

Rozhodování vyžadující lidský faktor – výběrem správné dataminingové techniky lze eliminovat nutnost zásahu pracovníka do procesu klasifikace žadatele o půjčku. *Statistické* metody lze použít pro určení klienta jasně náležícího do tříd „přijat“ či „zamítnut“. Existuje však množina hraničních případů, kdy nelze s pomocí těchto metod automaticky rozhodnout, zda žadatele akceptovat či ne. Například firma, která používá pro statistické rozhodování proceduru spočívající ve výpočtu parametru z atributů získaných od klienta, přijímá žadatele na základě překročení hodnoty daného parametru a zamítá pod touto hodnotou. Problém nastává u případů nacházejících se blízko určené prahové hodnoty. V takových situacích přichází na řadu lidský faktor, který rozhodne o přijetí klienta. Z podnikatelského hlediska by byla totiž chyba danou skupinu klientů automaticky zamítat. Metody *strojového učení* však tento nedostatek odstraňují, protože fungují na jiném principu. Tyto metody vytváří z množiny trénovacích dat řadu klasifikačních pravidel, které jsou následně využity rovněž pro správné zatřídění horních 10 % případů. Získaná pravidla také mohou posloužit k vysvětlení důvodu pro rozhodnutí. (Witten, Frank a Hall, c2011)

Marketing a prodej – v této oblasti je běžný výskyt obrovského rozsahu dat. Predikce zde představuje hlavní předmět zájmu dolování znalosti z dat, ovšem způsob tvorby rozhodovacího procesu většinou není podstatný. Do této oblasti patří problém pomíjivé loajality zákazníků, u nichž je velká pravděpodobnosti *odchodu ke konkurenci*. Tento problém se týká i bank, které byly průkopníky v nasazování data miningu do praxe, hlavně díky úspěšnosti v nasazování metod strojového učení pro hodnocení úvěrového rizika. V současné době banky zjišťují změny v zákazníkově chování, což může předcházet možné změně banky. Velmi častým příkladem užití data miningu v marketingu je predikce přechodu zákazníka ke konkurenčnímu operátorovi. Obvykle bývá snahou odhalit vzorce chování rentabilního zákazníka, poznat jeho potřeby a nabídnout mu vhodnou službu. (Witten, Frank a Hall, c2011)

Analýza spotřebního koše – za pomoci asociačních pravidel zjišťujeme, které položky se společně vyskytují v nákupním koši zákazníků. Tato metoda je hojně využívána v supermarketech a bývá často jediným zdrojem informací o chování zákazníků. Poznané skutečnosti mohou být využity při plánování prodejního prostoru,

snížení počtu slev výrobků na jeden, který je součástí skupiny výrobků často kupovaných společně, nebo k nabízení slevových kuponů na daný produkt v rámci této skupiny. (Witten, Frank a Hall, c2011)

Detekce podvodů – odhalování podvodného jednání klienta je založeno na detekci anomálií v datech, která spočívá v identifikaci případů, jejichž vlastnosti se významně liší od ostatních. Algoritmus provádí průzkum dat za účelem objevení skutečných anomálií ve snaze předejít označení legitimních případů za falešné. Dobrý detekční algoritmus dosahuje vysoké míry odhalených klientů, z nichž počet nesprávně označených je minimální. Vzhledem k obvykle menšímu počtu podvodných klientů oproti zbytku případů v databázi, je detekce anomálií použita na vytvoření profilu klienta legitimních transakcí. Profil je následně použit pro porovnávání charakteristik s nově příchozím případem, a v případě vzájemně významně odlišných vlastností je případ označen za podvodný. (Tan, Steinbach a Kumar, c2006)

2.1.3 Metodologie pro práci s daty

Metodologie vychází z obecných postupů pro práci s daty a často bývají přizpůsobeny pro modelování v konkrétním softwaru. Za samostatnou metodologii lze považovat KDD proces, který obsahuje ze zmíněných metodologií nejvíce průvodních kroků. Za vznikem a rozvojem těchto metodologií stojí velké společnosti zabývající se danou problematikou, jako je SAS nebo SPSS. Softwarově závislými metodologiemi jsou SEMMA od firmy SAS používaná v nástroji Enterprise Miner a 5A od společnosti SPSS. Univerzální metodologií je CRISP-DM, na jejímž rozvoji se rovněž podílela společnost SPSS. (Berka, 2003)

Metodologie 5A

Jednotlivé fáze metodologie 5A jsou následující (Berka, 2003):

- *Assess* – volba vhodných dat k provedení analýzy. Zajišťují se analytické nástroje a vyškolení odborníci pro práci s těmito nástroji.
- *Access* – sběr, příprava a skladování dat. Data jsou získávána z datových skladů, databází a dalších interních zdrojů. Je možné použít data z veřejných databází nebo z průzkumů, a to vlastních či provedených externí firmou.
- *Analyze* – samotné použití analytických nástrojů. V této fázi jsou zodpovídány definované otázky a získávány znalosti. SPSS doporučuje k analýze deskriptivní statistiku, OLAP technologii a metody strojového učení. Je vhodné použít najednou více metod a jejich výsledky poté porovnat.
- *Act* – znalosti získané v předchozím kroku se mění na tzv. znalosti akční, ze kterých se formulují doporučení.
- *Automate* – slouží k usnadňování provádění opakujících se úloh v podobě automatizace analýzy.

Metodologie SEMMA

Zkratka metodologie SEMMA představuje názvy fází Sample, Explore, Modify, Model a Assess, jejichž popis včetně překladu názvů je následující (Olson, Delen, c2008):

- *Vzorek (Sample)* – jedná se o takový vzorek dat, jehož velikost je dostatečně velká na to, aby byla zachována jeho informační hodnota a zároveň byl tak malý, aby se s ním rychle pracovalo.
- *Průzkum (Explore)* – fáze je zaměřena na porozumění datům, čemuž je napomáháno vyhledáváním neočekávaných trendů a odchylek v datech.
- *Modifikace (Modify)* – vychází z objevených skutečností v průzkumné fázi, kdy dochází k úpravám dat, jako je vytváření, výběr nebo transformace proměnných.
- *Modelování (Model)* – jakmile jsou data připravena, je možné přejít do fáze modelování, ve které se vysvětlují závislosti v datech. Vyhledáváme kombinaci proměnných, která spolehlivě předpoví požadovaný výstup.
- *Vyhodnocení (Assess)* – na konci se zhodnotí užitečnost a spolehlivost výsledků dataminingového procesu. V tomto kroku se modely vyhodnocují za účelem odhadu správnosti výsledků.

Metodologie CRISP-DM

CRISP-DM vznikla jako výsledek spolupráce firem DaimlerChrysler, SPSS zabývající se data miningem a tvorbou nástroje Clementine, NCR poskytující služby v oblasti data warehousingu a pojišťovnou OHRA. Metodologie CRoss-Industry Standard Process for Data Mining představuje projekt pod záštitou Evropské komise, který si kladl za cíl, vytvořit univerzální metodiku bez vázanosti na konkrétní nástroj. Účelem této metodiky je zrychlení, zefektivnění, zvýšení spolehlivosti a snížení nákladů procesu dolování znalosti. Metodologie také slouží jako průvodce možnými problémy a jejich řešeními, které mohou nastat během dolování. (Petr, 2006)

V současnosti nejpoužívanější metodologií je právě CRISP-DM. Metodologie skládající se ze šesti fází je vhodná pro popis postupu analýzy, avšak detaily a různá specifika by měla být aktualizována vzhledem k aktuálním trendům v oblasti Big Data projektů. O podporu této metodologie v dnešní době usiluje pouze nástroj IBM SPSS Modeler, což je nástupce Clementine. (Piatetsky, 2015)

2.2 CRISP-DM

Vzhledem k volbě nástroje, kterým je IBM SPSS Modeler 14.1, bude použita tato metodologie přizpůsobená pro práci v daném softwaru. Provedená rešerše je vypracována pro potřeby praktické části a zahrnuje především postupy a doporučení, které mají význam pro řešení.

Životní cyklus metodologie se skládá ze šesti fází, jejichž posloupnost není pevně dána. Ve většině případů se přechází z jedné fáze do druhé směrem dopředu i zpět. (IBM CORPORATION, 2011)

Metodologie se skládá z těchto šesti fází:

1. *Porozumění problému (Business Understanding)*
2. *Porozumění datům (Data Understanding)*
3. *Příprava dat (Data Preparation)*
4. *Modelování (Modeling)*
5. *Hodnocení (Evaluation)*
6. *Nasazení (Deployment)*

Příručka k metodologii CRISP-DM přehledně popisuje jednotlivé fáze data miningu z pohledu přínosu pro podnik, které byly shrnuty do následujících etap (IBM CORPORATION, 2011):

Porozumění problému

Nejdříve je potřeba *definovat podnikové cíle*, tj. čeho má být dosaženo z obchodního hlediska. Pokud má podnik problém s odchodem zákazníků, pak lze definovat cíl data miningu jako snahu o zamezení odlivu zákazníků ke konkurenci. Dalším cílem může být například zvýšení loajality zákazníků prostřednictvím nabízení přizpůsobených služeb. Poté se definují *kritéria plnění cíle*. Za *objektivní kritérium* lze považovat např. plánované snížení počtu odchozích zákazníků. Jako *subjektivní kritérium* je možné vybrat shluky, na nichž se pozitivně podepsalo zavedené opatření, jejichž postřehnutí však bývá složitější. V dalším kroku je nutné projít *inventář dostupných zdrojů*. Důkladnou analýzou hardwaru, datových zdrojů popřípadě lidských lze uspořit mnoho času. V případě hardwaru nás zajímá, zda je podporován. U dat je potřeba mít přehled o datových typech, formátu a jakým způsobem jsou skladována.

Porozumění datům

Počátek fáze je zaměřen na *sběr dat*. Data mohou pocházet přímo z podniku, kdy se může jednat o transakční data, data z průzkumu či z webového logu. Data je možno rovněž zakoupit a připojit k základním. K tomuto účelu jsou vhodná tzv. data demografická. Je třeba klást důraz na větší množství dat, což může vést k přesnějším modelům, nevýhodou však může být prodloužení doby zpracování. Shromážděná data musí být následně *popsána*, zejména datové typy (např. numerický neboli kategoriální, string, či boolean). V případě dat z různých zdrojů může dojít k nekonzistenci hodnot atributů. Např. označení pohlaví může být v jednom souboru reprezentováno pomocí hodnot „muž“ či „žena“ a ve druhém „m“ a „ž“. V takovém případě je nutné zavést jednotný způsob kódování hodnot, aby bylo možno datové množiny sloučit. U atributů nabývajících více než dvou hodnot lze použít číselné identifikátory, ke kterým musí být přiřazen popis. Následně provedeme průzkum dat s možnou formulací hypotéz. *Explorace* pomáhá k hlubšímu pochopení významu dat a případné úpravě jednotlivých hypotéz. Nakonec je prováděna *kontrola kvality* dat,

která spočívá v odhalování problémů, jako jsou chybějící hodnoty, chyby v datech při zadávání, chybná měření nebo nekonzistence v kódování.

Příprava dat

Příprava dat je jedna s nejpracnějších částí celého procesu dolování. Velmi často zabírá 50 až 70 % celkové doby projektu. Úspory času lze do jisté míry dosáhnout důkladnou přípravou přechozích dvou fází, přesto je však nutné věnovat této fázi značné množství času. *Selekce dat* probíhá buď horizontálně po řádcích, kdy vybíráme instance, jako jsou účty, produkty nebo zákazníky, které hodláme zahrnout do analýzy. V případě výběru po sloupcích vybíráme vlastnosti, které nás u položek zajímají. Za vlastnost lze považovat například objem transakcí nebo příjem domácnosti. Selekce atributů se provádí i z důvodu vypuštění citlivých údajů, jako jsou jména klientů, adresy, telefonní čísla nebo čísla kreditních karet. V případě *chybějících dat* lze jednotlivé záznamy odstranit. Pokud je chybějících hodnot u konkrétního atributu příliš mnoho, je možné celý sloupec z dat odstranit. V některých případech je možné chybějící údaj doplnit pomocí technik odhadu. Je potřeba rozlišovat mezi chybějícím údajem a prázdnou hodnotou. Ve druhém případě se v podstatě jedná o údaj sdělující absenci sledované vlastnosti. *Chyby v datech* je možné opravovat manuálně, odhadovat, případně odstraňovat záznamy či celé sloupce. Dalším problémem z hlediska kvality dat je zmíněná *nekonzistence v kódování*. Nekonzistence je potřeba napravit, aby bylo možné datové množiny slučovat. V případě potřeby tvorby nových dat se *odvozují atributy*. Nové atributy mohou být vytvořeny ze stávajících proměnných, kde mohou představovat rozdíl mezi dvěma časovými událostmi. Dále lze atributy vytvořit na základě výsledků slučování dat nebo jiné restrukturalizace. Nový atribut tak může obsahovat např. počty transakcí, nákupů, které byly vypočteny na základě průchodu daty. Dalším krokem je *integrace* dat, která má za účel spojení do jedné tabulky za účelem analýzy. Pro integraci existují dvě základní metody. První způsob je *slučování*, kdy bývá více datových množin sloučeno dohromady pomocí unikátního identifikátoru nacházejícího v daných množinách. Typicky může jít například o ID zákazníka. Druhý způsob, tzv. *připojení*, je založeno na integraci dvou nebo více množin s podobnými vlastnostmi, avšak s různými záznamy. Integrace dat je založena na podobných hodnotách atributů (např. název produktu nebo délka smlouvy). Posledním krokem před zahájením modelování je *formátování* dat. Každý algoritmus potřebuje určitý formát dat nebo řazení, aby mohl pracovat správně nebo rychleji. Formátování nemá za následek změnu významu dat, provádí se pouze v případě potřeb vyžadovaných konkrétním algoritmem.

Modelování

Modelování je obvykle založeno na několika iteracích, kdy se spouští řada modelů s výchozím nastavením parametrů a až poté se nastavení ladí, či dokonce vrací do fáze přípravy dat. Většinou nelze dosáhnout uspokojivých výsledků sestavením a spuštěním jediného modelu, což činí data mining zajímavým, protože existuje

mnoho možností nastavení. Před *výběrem modelu* bychom měli zvážit, o jaký typ cílové proměnné se jedná (např. symbolické neboli kategoriální). Dále uvažujeme definované dataminingové cíle a vybíráme podle toho, zda chceme spíše vyhledat zajímavé vzory chování zákazníka v datech týkajících nákupů nebo identifikovat studenty se sklonem k problémovému splácení půjčky. Algoritmy mohou mít speciální požadavky např. na velikost dat, typ, nebo se mohou navzájem lišit způsobem prezentování výsledků. Posledním krokem před zahájením samotného modelování je vytvoření tzv. *testovacího designu*. Klíčovým elementem v této fázi je stanovení tzv. *kritéria správnosti*. U metod učení s učitelem, jako je např. strom C5.0, lze správnost určit obvykle pomocí odhadu chybovosti vytvořeného modelu. Pro algoritmy učící se bez učitele, jako jsou Kohonenovy mapy, může být kritériem snadnost interpretace výsledku, nasazení nebo potřebný čas pro zpracování. Testovací design je prováděn za účelem popisu kroků potřebných pro testování vytvořeného modelu. Vzhledem k tomu, že samotné modelování je rovněž iterativní proces, je dobré vědět, kdy je vhodné přestat se změnou parametrů daného modelu a vyzkoušet jinou metodu, k čemuž nám poslouží vytvořený popis. Před koncem procesu tvorby modelů bývají obvykle zjišťovány nastavené parametry včetně poznámek, které vedou k nejlepším výsledkům. Dále jsou zjišťovány informace o *aktuálně vytvořeném modelu* a *popsány jeho výsledky* zahrnující výkonnost a objevené problémy s daty. Nyní již máme k dispozici *množinu modelů*, o kterých lze prohlásit, zda dosahují požadovanou míru přesnosti či efektivnosti, a mohou být tak označeny za konečné. Konečnost modelu může spočívat ve vhodnosti jeho nasazení nebo například v ilustraci zajímavých vzorů v datech. Vyhodnocování modelů by mělo probíhat s ohledem na stanovená kritéria v testovacím designu. Pro hodnocení modelů lze použít například evaluační grafy.

Hodnocení a nasazení

Fáze spočívá ve *zhodnocení* celého dataminingového procesu v kontextu plnění kritérií definovaných v počáteční fázi. V průběhu evaluace by mělo být zhodnoceno, zda jsou výsledky jasné, unikátní, umožňují jednoznačnou interpretaci a zda lze poznatky použít pro naplnění podnikových cílů. Po zhodnocení se dolovací proces ukončí a přejde se do fáze nasazení nebo se vrací zpět například z důvodu nedostatečné přesnosti modelů. *Nasazením* se rozumí proces využití nově získaných znalostí ke zdokonalování fungování konkrétní organizace. Toho může být dosaženo implementací modelů získaných v daném nástroji za účelem generování „churn skóre“ do podnikového datového skladu. Nasazení však může proběhnout v podobě využití získané znalosti za účelem vyvolání změn v organizaci. Například, pokud objevíme alarmující vzory v chování zákazníků, jejichž věk dosáhl určité hranice, není nutné tyto výsledky formálně integrovat do informačního systému, ale stačí je zahrnout do plánování a marketingového rozhodování.

2.3 Použité metody

2.3.1 Rozhodovací stromy

Jednou z nejběžnějších dataminingových metod jsou rozhodovací stromy. Jedná se o strukturovaný predikční model, kde jednotlivé uzly provádí test na atribut, z uzlů vycházejí větve reprezentující výsledek testu, a kde každý list je značen samostatnou třídou či jejich rozložením. Algoritmus je založen na strategii „rozděl a panuj", kdy se rekurzivně shora dolů postupně konstruuje rozhodovací strom. Každý objekt je klasifikován podle cesty z vrcholového či kořenového uzlu směrem k listům, kde hrany odpovídají hodnotám atributů daného objektu. V kořenovém uzlu jsou trénovací data rozdělena na dvě nebo více množin podle tříd atributu, které se směrem dolů dělí dle rozdělovacího kritéria v jednotlivých uzlech. (Wang, c2009)

CART stromy

Výhoda toho typu stromu spočívá v možnosti použití pro cílové proměnné konečného i neurčitého počtu hodnot, což strom předurčuje k použití nejen pro klasifikaci, ale i regresi. Těžištěm výkonnosti tohoto typu stromu je vyčerpávající způsob hledání možného rozdělení trénovací množiny, které zajišťuje výběr optimálního rozdělení. Kritériem pro rozdělení je výpočet Giniho indexu. Doba hledání se však může značně prodloužit v případě práce s kvalitativními proměnnými s velkým množstvím k kategorií, kde počet testovaných rozdělení bude 2^{k-1}. Ve své základní variantě je CART strom binární. Nevýhodou binární struktury je tvorba stromů o mnoha hladinách, což je často dělá příliš komplexní a nesrozumitelné. (Tufféry, 2011)

C4.5 a C5.0

Algoritmus C4.5 vznikl jako nástupce původního Quinlanova ID3 algoritmu. Tento algoritmus se liší od CART rozdělovacím kritériem, kde je uvažován informační zisk místo Giniho indexu. V každém uzlu stromu je vybrán jeden atribut, který nejefektivněji rozděluje množinu trénovacích případů na podmnožiny zvýšeného počtu případu jedné třídy či více. Kritériem je normalizovaný informační zisk neboli změna v neurčitosti, což je výsledek výběru určitého atributu pro rozdělení dat. Jako rozhodovací kritérium se volí atribut s nejvyšším ziskem. Díky sledování míry snížené entropie je možné pracovat i s atributy, ve kterých chybí hodnoty. (Hssina et al., 2014)

Pokud je strom příliš přizpůsoben konkrétnímu datovému vzorku, dochází často k přetrénování. Vytvořený strom je pak citlivý na šum a na dosud nepoznaných vzorcích vykazuje horších výsledků. Aby byl tento efekt eliminován, využívá se tzv. prořezávání, které snižuje klasifikační chybu. Prořezávání zmenšuje velikost stromu odstraňováním sekcí, které mají slabou schopnost klasifikovat. Ve výsledku má prořezávání dvojí efekt, kdy za prvé, snižuje velkou složitost vytvořeného klasifikátoru a za druhé, zvyšuje přesnost odstraněním šumem či chybou ovlivněných větví. (Hssina et al., 2014)

C5.0 je jeho nástupce, který bývá součástí komerčních řešení. Oproti starší verzi nabízí několik vylepšení. C5.0 je efektivnější z hlediska potřeb paměti a výpočetního času. V mnoha případech vedlo toto významné zlepšení ke zkrácení doby generování modelu z možné hodiny a půl na 3,5 sekundy. Dále obsahuje oproti C4.5 metodu boosting pro zvýšení přesnosti klasifikace. (Maimon, Rokach, c2015)

2.3.2 Neuronové sítě

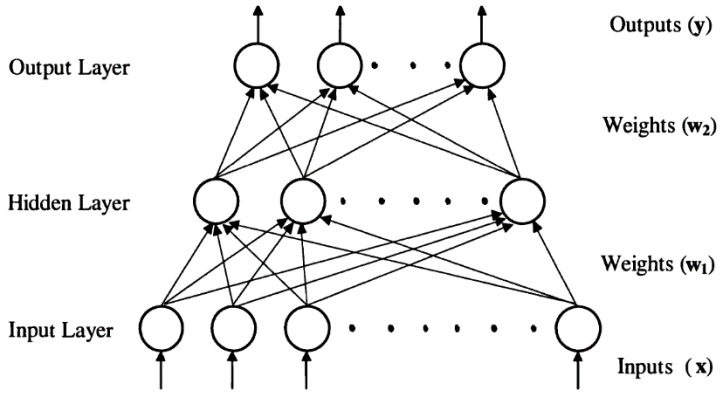
Umělá neuronová síť je tvořena jednotkami zvanými perceptrony. Každý perceptron získává na vstupu signál, který je tvořen kombinací výstupů předchozích perceptronů. Signál je následně předán obvykle nelineární aktivační funkci, která vypočtenou hodnotu směruje na perceptrony v nižších vrstvách. (Larose, c2005)

Jednou z výhod neuronových sítí je vzhledem k jejich robustnosti odolnost vůči šumu v datech. Neuronové sítě jsou tvořeny mnoha jednotkami propojených vahami, čímž jsou předurčeny k použití i na velmi šumem poznamenaných či až chybných datech. Na druhou stranu neposkytují ve srovnání, především s rozhodovacími stromy, takovou formu informace, která je podobná lidskému způsobu interpretace. Neméně významným rozdílem je vyšší potřebný čas k natrénování. Dále je potřeba, aby hodnoty veškerých atributů byly převedeny na interval 0 až 1. To je v případě numerických atributů snadno řešitelný problém, nicméně u atributů kategoriálních je někdy problém číselnou hodnotu přiřadit. Jde především o případy, kdy atribut nominálního charakteru dosahuje více hodnot, které po vyjádření v daném číselném intervalu získají určitou pozici, z čehož může neuronová síť vyvodit nesmyslný závěr. Vzhledem k produkování spojitého výstupu je možné neuronové sítě použít nejen pro klasifikaci, ale i regresi. (Larose, c2005)

Vícevrstevná neuronová síť

Vícevrstevná neuronová síť (MLP) nebo také vícevrstevná dopředná neuronová síť je v současnosti nejpoužívanější typ neuronové sítě v praxi. Tento typ sítí je určen k hledání vztahů mezi množinou vstupních proměnných neboli prediktorů a výstupních proměnných. Obvyklou strukturu znázorňuje Obr. 3. Neurony ve vstupní vrstvě odpovídají nezávislým proměnným neboli prediktorům, které jsou důležité pro predikci hodnot závislé proměnné ve výstupní vrstvě. Kromě neuronů ve vstupní vrstvě každý přijímá sumu vážených vstupů následně předanou aktivační funkci, která vypočtený výsledek předá do vedlejší vrstvy. V praxi se používá několik typů aktivačních funkcí, z nichž některé jsou následující:

- Sigmoidální (logistická) funkce, $f(x) = (1 + \exp(-x))^{-1}$
- Hyperbolická tangentní funkce, $f(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$
- Sinová a kosinová funkce, $f(x) = \sin(x)$, $f(x) = \cos(x)$
- Lineární funkce nebo funkce identity, $f(x) = x$



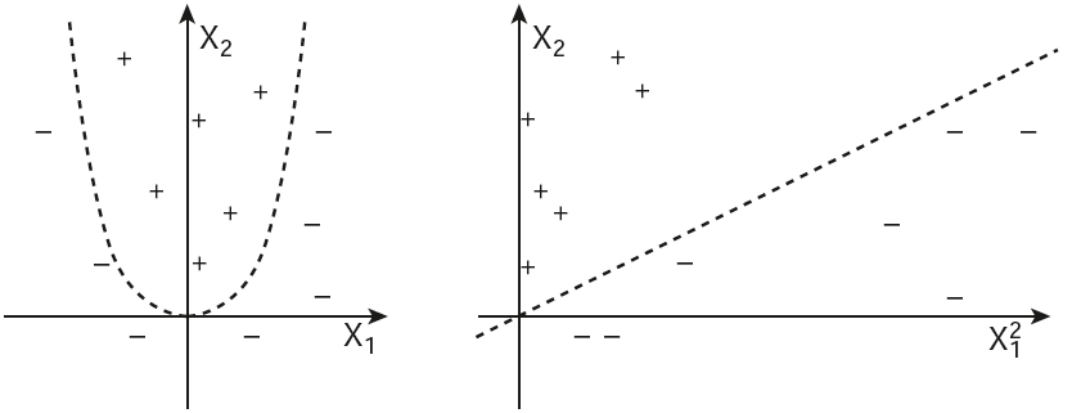
Obr. 3 Struktura vícevrstevné neuronové sítě
Zdroj: Maimon, Rokach, c2005

V současnosti nejpoužívanějším typem transformační funkce je logistická funkce. U klasifikačních problémů, kde výstupní proměnná je binární či kategoriální, lze tuto funkci použít ve výstupní vrstvě za účelem snížení rozsahu hodnot výstupů. Váhy jednotlivých parametrů musí být stanoveny ještě před zahájením fáze klasifikace. Nejprve je síť naplněna trénovacími případy, které sestávají z množiny vstupních vzorků s příslušným výstupem. Pro každý trénovací vzorek jsou vstupní hodnoty zváženy a sečteny v každém neuronu ve skryté vrstvě. Vážený součet je poté předán aktivační funkci, která přenese výsledek k neuronům ve výstupní vrstvě. Pomocí jejich aktivační funkcí jsou ze sítě získány výstupy, které jsou porovnány s hodnotou cílové proměnné. Váhy spojení jsou upravovány tak, aby síť poskytovala lepší aproximaci vzhledem k požadovanému výstupu. Tento proces se obvykle opakuje tolikrát, dokud rozdíl mezi výstupem ze sítě a cílovou hodnotou není co nejmenší, resp. dokud není dosaženo požadované úrovně chyby. (Maimon, Rokach, c2005)

2.3.3 Support vector machines

SVM je metoda učení s učitelem používaná ke klasifikaci nebo regresi. U klasifikace se obvykle používá nelineární funkce, která transformuje vstupní data do vysoko dimenzionálního prostoru, ve kterém je možné případy jednodušeji rozdělit (viz Obr. 4). Následně je zkonstruována co nejširší nadrovina za účelem optimálního rozdělení dat do tříd. Dvě nadroviny jsou zkonstruovány na obou stranách dělící nadroviny, aby byla maximalizována vzdálenost mezi těmito dvěma paralelními nadrovinami. Výsledná generalizační chyba se odvíjí od velikosti rozdělovací hranice, která klesá s rostoucí vzdáleností. Pro transformaci dat do oddělitelné podoby se používá několik typů nelineárních funkcí. Obvykle se jedná o následující funkce (Olson, Delen, c2008):

- polynomiální
- radiální bázová
- sigmoidální



Obr. 4 Příklad transformace dat pomocí SVM do lineárně oddělitelné podoby
Zdroj: Tufféry, 2011

2.3.4 Analýza hlavních komponent a faktorová analýza

Analýza hlavních komponent (PCA)

Tato metoda patří mezi techniky sloužící k redukci počtu dimenzí, tj. počtu atributů. Cílem je nalezení takového počtu nových dimenzí či komponent, které maximálně zachovávají variabilitu původních dat. V typickém případě, první komponenta zachycuje co největší variabilitu. Následně druhá dimenze bývá ortogonální k první a zachycuje co nejvíce zbývající variability. PCA má několik zajímavých výhod. První je, že identifikuje nejsilnější vazby v datech. Toto ji předurčuje k použití hledání vzorů v datech. Druhá a hlavní výhoda spočívá ve schopnosti shrnutí převážného množství variability do několika málo dimenzí. Metoda je vhodná v případě, kdy máme k dispozici příliš mnoho proměnných pro použití určité techniky, která by jinak vedla ke špatným výsledkům. Třetí výhodou je, že za předpokladu nižšího šumu a většího množství vzorů v datech, lze tento šum značně eliminovat. Toto je rovněž vhodné pro řadu dataminingových technik či jiných algoritmů používaných k analýze dat. (Tan, Steinbach a Kumar, c2006)

Datovou množinu s k – numerickými atributy si lze představit jako oblak bodů k - rozměrného prostoru. Atributy představují souřadnice v prostoru. Volba os, podle kterých chceme systém koordinovat je libovolná. Pomocí vybraných horizontálních a vertikálních os lze jednotlivé případy vynést do bodového grafu. Ať už použijeme jakýkoliv koordinační systém, mrak bodů má určitý rozptyl ve všech směrech indikující stupeň šíření kolem střední hodnoty v daném směru. (Witten, Frank a Hall, c2011)

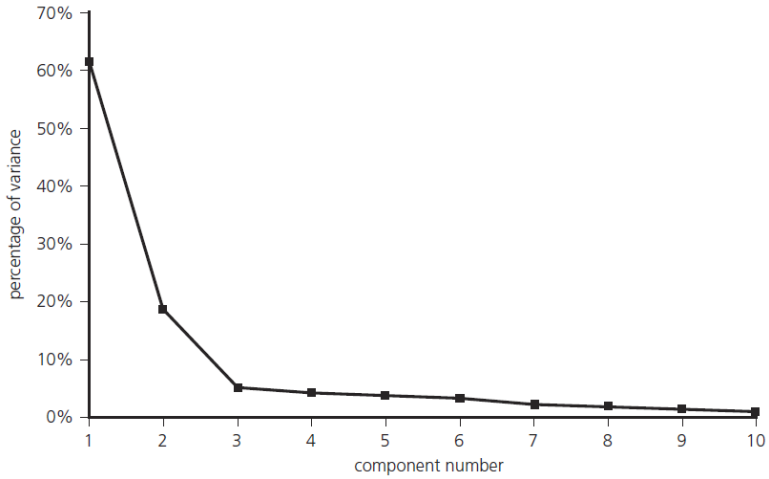
Příklad užití PCA je následující: umístíme první osu, která je ve směru největšího rozptylu bodů, abychom jej maximalizovali podél této osy. Druhá osa bude kolmá k první. V případě dvoudimenzionálního grafu je směr šíření určen první osou. U grafu s více dimenzemi máme více možností jak postupovat, nicméně je potřeba opět osy umístit tak, aby byl maximalizován rozptyl podél jednotlivých os. Celý tento problém lze většinou řešit pomocí specializovaného programu a není složité

mu porozumět za předpokladu, že uživatel zná následující pojmy (Witten, Frank a Hall, c2011):

- *kovarianční matice*
- *diagonalizace matice*
- *vektor vlastních čísel* (angl. Eigenvector)

Kovariance dvou atributů měří, jak silně jsou proměnlivé mezi sebou. *Kovarianční matice* tedy vyjadřuje jednotlivé proměnlivosti závislých atributů, které jsou v jedné komponentě. Následně je matice *transformována (diagonalizována)* za účelem nalezení vektorů vlastních čísel, tzv. eigenvektorů. *Eigenvektory* představují množinu nových proměnných či os, tedy vzniklé komponenty. V podstatě lze metodu PCA považovat za příklad rotace původních souřadných os do nové množiny samozřejmě s nižším počtem tak, aby byla co nejvíce zachována variabilita původních dat. Nově vzniklé osy spolu nekorelují. (Tan, Steinbach a Kumar, c2006)

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100%



Obr. 5 Příklad komponent a distribuce rozptylu (angl. variance)
Zdroj: Witten, Frank a Hall, c2011

Faktorová analýza

Faktorová analýza navazuje na principy metody PCA, liší se však například použitím matice korelační místo kovarianční. Faktorová analýza se provádí za účelem vyjádření původních proměnných jako lineárních kombinací malého počtu skrytých či latentních faktorů. Předpokládá se, že jedna skupina atributů není v silné korelaci s jinou, ale atributy v této skupině jsou navzájem silně korelovány, pravděpodobně v důsledku existence nějakého vysvětlujícího atributu. Tímto vysvětlujícím atributem je vypočtený faktor, který reprezentuje skupinu původních proměnných. Pomocí transformací je vytvořena matice faktorových zátěží, která zobrazuje závislosti hodnot původních proměnných na těchto skrytých faktorech. Příkladem mohou být například data o výkonech desetibojařů. Můžeme zjistit, že stejní atleti dosahují po-

dobných výsledků v disciplínách, kde rozhoduje např. rychlost. Obdobně atlet vyhrávající v disciplíně vyžadující vytrvalost, bude vítězit v jiných disciplínách, kde hraje hlavní roli tento faktor. Z toho můžeme vyvodit hypotézu, že výkon atleta v disciplíně je dán jejím druhem a dvou skrytých faktorech: rychlosti a vytrvalosti. (Tan, Steinbach a Kumar, c2006)

Rotace

V souvislosti s PCA a faktorovou analýzou je dále potřeba nastínit význam pojmu rotace. V obou případech může být převážná většina zátěže umístěna na první osu, to však v případě více skupin atributů může být nežádoucí. Ačkoliv atributy uvnitř faktoru jsou silně korelačně provázány, všechny proměnné mohou být orientovány ve směru první osy v důsledku převahy vysokých hodnot nad nízkými a skupiny tak nevyniknou. Rotace os může být řešením k vhodnější distribuci variability či korelace, jako náhrada za kritérium nejvyšší zátěže na první osu. Nutno podotknout, že po aplikování rotace dojde ke změně distribuce zatížení, nicméně faktory nadále zachovají celkové zatížení, tj. informační hodnotu. (Tufféry, 2011)

Rozlišujeme dva základní typy rotací: ortogonální a neortogonální (angl. obligue). V prvním případě nejsou faktory korelovány a umožňují snadnější interpretaci. Druhý typ rotace předpokládá, že faktory již na sebe nejsou kolmé a není je tak snadné interpretovat. Na druhou stranu mají však výhodu v silnější korelaci faktorů s proměnnými. Vzhledem k tomu, že v praktické části budeme vyžadovat metodu pro snadnější interpretaci faktorů, budou nás zajímat rozdíly mezi hlavními ortogonálními rotacemi os, které jsou např. (Tufféry, 2011):

- *Quartimax* – kde veškeré proměnné mají vysoký vliv na daný faktor, nenulový vliv na faktor jiný a prakticky nulový na všechny faktory.
- *Varimax* – nejčastěji používaný typ rotace. Každý faktor silně koreluje s některými proměnnými a slabě s ostatními. Jinak řečeno, některé z proměnných mají velký vliv na danou osu, zatímco jiné velmi malý. Osy lze snadněji interpretovat.

2.3.5 K – means

Jedná se o algoritmus používaný k nehierarchickému shlukování. Na rozdíl od jiných clusterizačních metod sami určujeme počet clusterů, do kterých mají být jednotlivé případy zařazeny. Na počátku určíme inicializační množinu přiřazením objektů do jednotlivých shluků. Vhodné rozmístění této množiny nám může urychlit celý proces. Poté začneme počítat středy jednotlivých shluků. Tyto středy neboli centroidy se počítají pro všechny případy v rámci jednotlivých shluků. Následně musíme zkontrolovat, zda každý případ je co nejblíže středu svého shluku a co nejdále od středu sousedního shluku. Pokud narazíme na případ, který je zařazen do nesprávného shluku, tj. ve skutečnosti je blíže středu jiného shluku, pak jej do bližšího shluku přeřadíme. Poté musíme znovu přepočítat středy těchto dvou shluků. Pokud po další kontrole vzdáleností nezaznamenáme žádný nesprávně zařazený objekt, pak je proces shlukování u konce a bylo dosaženo optimálního výsledku. V opačném případě je nutné záznamy přetřídit a vzdálenosti znovu přepočítat. Kvalitu shlukování je

možné zvýšit několika způsoby, např. výběrem inicializační množiny s nejmenším součtem kvadratických chyb. (Cleff, 2014)

V SPSS Modeleru je kvalita shlukování určována metodou siluety míry koheze a separace shluku. Výsledek shlukování může nabít hodnot „špatný“, „přijatelný“ či „dobrý“. Silueta měří průměry nad všemi záznamy dle následujícího vzorce:

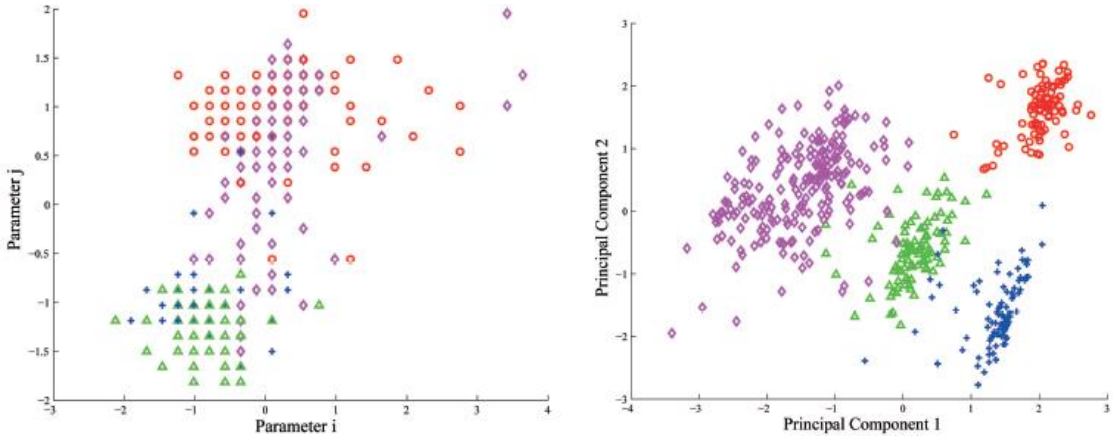
$$\frac{B-A}{maximum\ (A,B)}$$

Kde A je vzdálenost záznamu od středu svého shluku a B vzdálenost záznamu od nejbližšího středu shluku, kterého záznam není součástí. Výsledek „1“ znamená, že všechny případy jsou v blízkosti středu svého shluku. Jestliže vyjde hodnota „-1“, pak veškeré případy jsou umístěny mimo své shluky. „0“ znamená, že záznamy se nachází v polovině vzdálenosti mezi středem svého shluku a středem nejbližšího shluku. (IBM CORPORATION, 2012)

2.3.6 Kombinace PCA a K – means

Metoda PCA je široce používaná statistická metoda pro dimenzionální redukci na bázi učení bez učitele. K – means se používá pro deskriptivní úlohy rovněž bez učitele. Bylo dokázáno, že použití analýzy hlavních komponent má za následek zlepšení výsledků shlukování využívajícího spojitého měřítka. Způsob rozmístění centroidů algoritmu k – means je dán použitím kovarianční matice metody PCA. Dimenzionální redukce PCA automaticky provádí shlukování ještě před použitím K – means. (Ding, He, 2006)

Výhody výsledného shlukování jsou evidentní na Obr. 6. V případě zobrazování případů do grafu za účelem vizualizace shluků je po použití metody PCA potřeba na horizontální a vertikální osu vynést faktory (či komponenty) místo náhodných proměnných. Shluky s použitím PCA lze lépe vizuálně oddělit.



Obr. 6 Srovnání shlukování pomocí K-means bez použití PCA (vlevo) a s ní (vpravo)
Zdroj: Jollifee, c2002

3 Současný stav

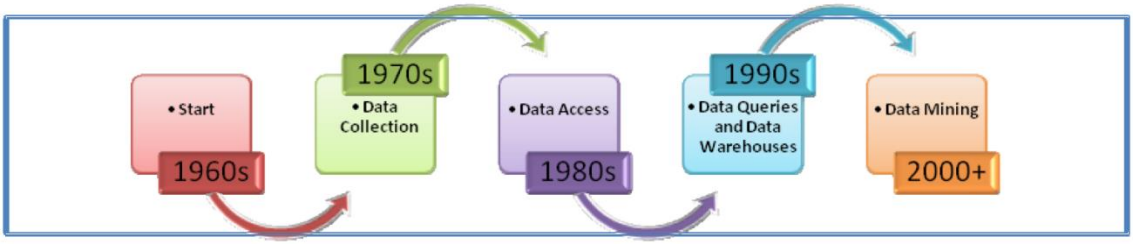
3.1 Uplatnění data miningu v komerční praxi

Ačkoliv jsou dataminingové techniky v praxi poměrně nové, touha po získávání profitabilních informací z dat má již dlouhou historii. Nesahá pouze do dob, kdy byly vynalezeny první počítače, ale ještě daleko předtím. V průběhu času byl data mining často spojován s pojmy, jako např. objevování znalostí, business intelligence, prediktivní modelování, analýza apod. Cílem data miningu v praxi je nalezení takové znalosti, která bude pro podnik představovat přidanou hodnotu. Nalezení informačně přínosného vzoru lze použít při různých rozhodovacích procesech. Zákazník může být aplikací v call centru automaticky označen barvou, určující jeho atraktivitu pro podnik. „Zelená" barva značí spokojenost zákazníka, „žlutá" znamená určité problémy, kdy zákazník je pro podnik cenný, ale jeví známky rizika spojeného s odchodem. „Červený" zákazník je nespokojený a je u něj pravděpodobný přechod ke konkurenci. Na základě tohoto určení může podnik na daný typ zákazníka zacílit za účelem přesvědčení, aby zůstal, nebo se naopak zaměřit na loajální klienty. (Berry, Linoff, c2004)

Tento způsob využití data miningu patří do oboru tzv. řízení vztahu se zákazníkem - Customer relationship managementu (CRM). Jedná se o soubor postupů a nástrojů, které napomáhají organizovaným způsobem budovat vztahy se zákazníky. CRM pracuje s informacemi získanými prostřednictvím kontaktu s klienty a využívá je pro efektivní interakci, a to ve všech fázích vztahu. Díky možnostem informačních technologií a webu je nyní možné budovat vztahy se zákazníky lépe, než tomu bylo v minulosti. Software pro podporu CRM byl původně zaměřen spíše na zjednodušení organizace a správy informací o zákaznících. Možnost hlubší analýzy dat s cílem získat profitabilní znalost se naskytla až s příchodem data miningu. (Gupta, Aggarwal, 2012)

Myšlenka řízení vztahu se zákazníkem je založena na principu napodobování chování malých podniků, které vynikají právě v osobním přístupu k zákazníkovi. Malý podnikatel má větší přehled o svém zákazníkovi a může ho tak lépe oslovit. Firmy se na základě těchto informací mohou rozhodovat, zda zákazník stojí za oslovení, či je vhodné ho nechat odejít ke konkurenci. Zavedení CRM do podnikání je zkrátka jeden z významných kroků v podnikání. Například banky bez podpory CRM si zajišťují svůj zisk pouze na základě udržování co největšího rozdílu mezi úrokovou mírou vkladů a půjčených peněz a nepřizpůsobují své bankovní produkty na míru potřeb klienta. K poznání společných potřeb určité skupiny klientů nám může pomoci dataminingová technika shlukování. V tomto kontextu je dolování dat chápáno jako kolekce nástrojů a technik pro podporu pro-zákaznicky orientovaného podniku. V širším slova smyslu se jedná o rozhodování podložené informacemi získanými ze vztahu se zákazníky, kdy výsledky těchto rozhodnutí přináší prospěch. (Berry, Linoff, c2004)

Počátek data miningu byl zaznamenán přibližně na konci 60. let 20. století. Prvotní etapa s názvem *sběr dat*, byla zaměřena na tvorbu jednoduchých reportů předformátovaných informací z dat uložených v databázích. V 80. letech byla vyžadována vyšší frekvence individuálního *přístupu k datům*, k čemuž sloužilo jednoduché dotazování nad databázemi. Po roce 1990 uživatelé začali vyžadovat okamžitý přístup k detailním informacím prostřednictvím komplexních *dotazů nad databází*, které poskytovaly odpovědi na otázky tzv. za běhu. Informace byly vyžadovány „just-in-time", aby korespondovaly s výrobou a rozhodovacími procesy. Uživatelé formulovali své vlastní dotazy za účelem extrakce požadované informace. Samotný *data mining* v podobě, jakou známe nyní, je používán přibližně od roku 2000, kdy hovoříme o různých specializovaných nástrojích a technikách pro nalezení významných vazeb v datech. (Rohanizadeh, 2009)



Obr. 7 Historický vývoj data miningu.

Zdroj: Rohanizadeh, 2009

Většina dataminingových technik existovala, alespoň v podobě akademických algoritmů, již před několika desítkami let. Pro jejich uplatnění v komerční praxi na konci 90. let měly podíl následující faktory (Berry, Linoff, c2004):

- data jsou vytvářena
- data jsou skladována
- výpočetní síla je již dostupná
- zájem o podporu řízení přístupu k zákazníku je silný
- komerční dataminingové nástroje jsou dostupné

Kombinace těchto faktorů způsobila, že dolování dat se stále častěji objevuje jako základ obchodních strategií mnoha společností. Příkladem úspěšné strategie založené na dolování znalostí z uživatelských dat je vyhledávač společnosti Google, který jako první zkombinoval sofistikované algoritmy pro vyhledávání s obchodním modelem založeným na maximalizaci příjmů z tzv. klikání na reklamy. V podstatě v každém oboru podnikání firmy zjišťují, že mají k dispozici informace o svých klientech, které mohou v budoucnu využít.

Data mining má největší význam, pokud pracujeme s *obrovským množstvím dat*. Ve své podstatě většina algoritmů vyžaduje toto množství pro výstavbu a trénování modelů. Data mining lze tedy uplatnit v podstatě všude, kde jsou dostupná data, což

u firem vlastnících informace o zákaznících je zcela běžné. Pouze jedna osoba může při prohlížení webu e-shopu vytvářet množství dat o velikosti několika desítek kilobajtů za den.

Získaná data se sdružují v *datových skladech*, které mohou obsahovat data z několika zdrojů v různých formátech. Datový sklad poskytuje jednotný pohled na firemní data napříč všemi divizemi. Je to místo, kde se shromažďují veškerá historická data, která mohou být následně použita pro rozhodování. Výhoda datového skladu spočívá v možnosti využít informací nejen pro operativní rozhodování, ale také jako podklad pro vrcholový management. Přítomnost datového skladu tedy evidentně data mining usnadňuje, a v případě jeho absence je pro dolování nutné provést další nezbytné kroky. (Witten, Frank a Hall, c2011)

Dataminingové nástroje obvykle vyžadují několikanásobný průchod obrovským množstvím dat. Z tohoto důvodu jsou také náročné na *výpočetní výkon*. Stále se snižující ceny disků, pamětí, procesorů a síťových prostředků však přispívají k jedinému – drahé technologie, který byly původně používány v laboratořích nebo pro akademické účely se přenáší do prostředí běžných podniků.

Napříč všemi odvětvími si firmy uvědomují, že informace získané z *kontaktu se zákazníky* jsou pro ně klíčovým aktivem. Tyto informace nejsou užitečné jen pro samotnou firmu, ale i pro ostatní. Mohou tedy představovat produkt, se kterým lze obchodovat. Samozřejmě musí firma zvážit, zda se jí vyplatí data o klientech prodat ostatním, čímž může ztratit konkurenční výhodu. Prodat ovšem může data, která sama nehodlá využít. Například data ze společnosti spravující kreditní karty obsahují informace, jak často zákazník nakupuje letenky. Tato data tak mohou být využita leteckou společností, která nabídne levnější kilometry pasažérům, kteří často létají s různými leteckými společnostmi.

Z hlediska správně vybudované datové základny pro podporu CRM řešení je zapotřebí, aby data o zákaznících obsahovala následující informace (Gupta, Aggarwal, 2012):

- *transakční* – měla by být dostupná kompletní historie transakcí (prodejů, u banky např. vklady a výběry)
- *kontaktní údaje* – v dnešní době již není problém kontaktovat zákazníka a máme na výběr z několika možností.
- *deskriptivní informace* – slouží pro segmentaci či jiné analytické úlohy.
- *reakce klientů na marketingové kampaně* - tato část obvykle obsahuje informace o tom, zda byla marketingová iniciativa úspěšná, tedy zda zákazník produkt přijal, či odmítl.

Podpora dataminingových technik *specializovaným softwarem* je vždy otázkou času, tedy než se nově vyvinuté algoritmy zavedou do praxe. Tyto techniky vycházejí z oborů jako je statistika, umělá inteligence, konkrétně strojové učení. Současná doba se kromě vývoje nových algoritmů soustřeďuje na inovaci a zlepšování již používaných, které bývají obsaženy v komerčních nebo open-source softwarových řešení.

3.2 Využití data miningu v bankovní praxi

Bankovní průmysl si nyní uvědomuje, jak dolovací techniky využít za účelem získání výhodnější pozice na trhu. Aplikací těchto metod mohou banky například lépe predikovat, jak se klienti budou chovat po zvýšení úrokových sazeb, či kteří klienti budou pravděpodobně poptávat nově zaváděný produkt. Dále mohou identifikovat klienty s vyšším rizikem prodlení (defaultu) se splácením dluhu nebo získávat ze vztahu se zákazníkem lepší profit. Disponují obrovským množstvím dat zahrnující informace o transakcích, demografických údajích, údajích o využívání kreditních karet atd. Protože bankovnictví spadá do sektoru služeb, je požadováno efektivní a silné CRM. Případy, ve kterých banky data mining využívají, lze shrnout do následujících podoblastí (Moin, Ahmed, 2012):

Marketing

Jde se o jednu z nejčastěji používaných podoblastí dolování dat. Na základě analýzy klientské databáze je možné spojovat klientovo chování do souvislosti s určitým produktem, cenou či distribučním kanálem. Banky tak mohou očekávat reakci od konkrétního typu zákazníka v případě, že nabízí určitý produkt. Dále je možné oddělit profitabilní zákazníky od méně zajímavých klientů z pohledu banky.

Risk management

Banky potřebují odhadovat spolehlivost klientů, zejména při poskytování různých úvěrů, nabízení kreditních karet novým zákazníkům nebo prodlužování již existujících kreditních linií. Úvěr lze poskytnout klientovi po posouzení několika aspektů, jako např. velikost půjčky, úroková míra, typ nemovitosti zatížené hypotékou, demografické údaje, příjem a úvěrová historie. Jedním ze základních nástrojů risk managementu využívajícího dataminingové metody je kreditní skórování používané k ohodnocení klientů.

Detekce podvodů

Díky data miningu bývá mnoho podvodných jednání včas odhaleno a hlášeno. Být v pozici toho, kdo je schopen podvody detekovat, je jistě pro řadu podniků trnem v oku. V praxi existují dva různé přístupy, jak může banka podvod odhalit. V prvním případě banka „proklepne“ datové sklady zájemce o úvěr, aby získala určité vzorce chování. Tyto vzorce následně aplikuje na svoji databázi a výsledky porovná. Druhý způsob je založen na detekci pomocí vlastních zdrojů banky.

Řízení vztahu se zákazníkem

V dnešní době mají zákazníci v nabídce spoustu produktů od různých bank. Banky by měly používat data mining k lepšímu poznání potřeb klienta a přizpůsobovat tak produkty na míru. Výsledkem úspěšného opatření tak může být udržení stávajícího zákazníka. Poznání důvodu odchodu zákazníka ke konkurenci může v budoucnu pomoci.

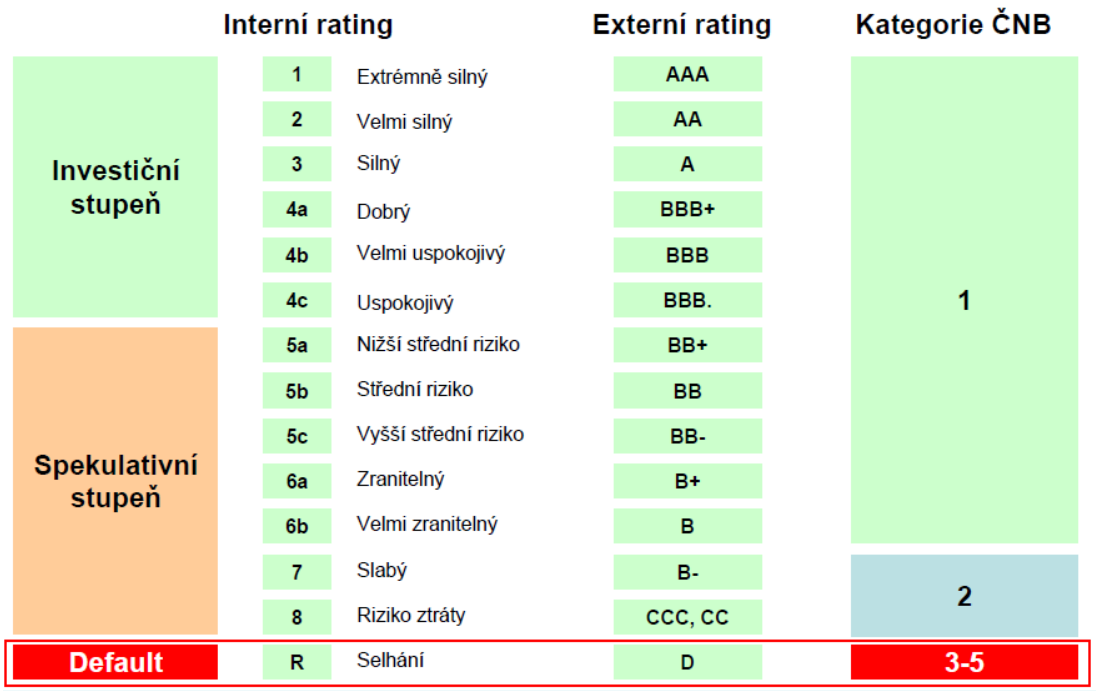
3.2.1 Vybrané metody používané k úvěrovému skórování

Nejvíce výzkumů v odborných článcích bylo věnováno použití *neuronových sítí* pro klasifikaci úvěrů. Jedná se tedy o nejčastěji porovnávanou metodu v odborných článcích orientovaných na kreditní skórování. Umělé neuronové sítě jsou schopné rozpoznat komplexní a nelineární vzorce závislostí mezi cílovým atributem a vstupními proměnnými, což umožňuje predikovat kredibilitu nového uchazeče o úvěr. Dalším nejvíce diskutovaným algoritmem jsou *podpůrné vektory (SVM*) založené na statistickém učení, které se používá pro oddělení běžně nelineárně separabilních dat. Častým postupem bývá vytvoření tzv. *ensemble modelů*, které kombinují predikci několika klasifikátorů. Kombinovat lze modely jednoho typu s různým nastavením modelů, ale i modely vytvořené odlišnými klasifikačními algoritmy. Při tvorbě ensemble modelů se používají různé typy, jako je např. *boosting* či *bagging*. Jednou z nejvíce diskutovaných metod pro použití kreditního skórování, která nebyla v analyzovaných článcích[1] často pozorována, jsou *rozhodovací stromy*. U nich se předpokládá přínos ve vysvětlení příčin klasifikace do dané třídy ve snadněji srozumitelné podobě oproti jiným přístupům. Dále je diskutován tzv. *hybridní přístup* spočívající v kombinaci více metod najednou, např. současným použitím *klasifikace* a *shlukování*. (Sadatrasoul et al., 2013)

Podle společnosti Kreditech (2015) je předností skóringových modelů stavěných na obrovském množství historických dat především to, že klient nemá možnost odhadnout výši svého skóre a přizpůsobit svoje chování tak, aby si svoji bonitu zvýšil. Důvodem je složitost propojení různých sledovaných charakteristik, které klient nemá šanci odhadnout.

Úvěry nemusí být klasifikovány pouze do kategorií problémový či bezproblémový. V praxi se používá rating, který slouží k ohodnocení stupně investice do poskytnutého úvěru. Jinak řečeno, úvěry, které mají větší sklon k nesplácení, mají vyšší výnos. Příklad ratingu je uveden na Obr. 8, kdy banka používá svůj vlastní postup pro určení stupně výnosnosti investice. Kromě vlastního a externího ratingu je uvedeno srovnání s kategoriemi podle ČNB, kde kategorie 1 značí úvěry splacené do 30 dnů od data splatnosti a 2 úvěry splacené do 90 dnů. Ostatní kategorie jsou souhrnně značeny jako selhání.

[1] Vybrané články pochází z období 2000 až 2012.



	Interní rating		Externí rating	Kategorie ČNB
Investiční stupeň	1	Extrémně silný	AAA	1
	2	Velmi silný	AA	
	3	Silný	A	
	4a	Dobrý	BBB+	
	4b	Velmi uspokojivý	BBB	
	4c	Uspokojivý	BBB.	
Spekulativní stupeň	5a	Nižší střední riziko	BB+	
	5b	Střední riziko	BB	
	5c	Vyšší střední riziko	BB-	
	6a	Zranitelný	B+	
	6b	Velmi zranitelný	B	
	7	Slabý	B-	2
	8	Riziko ztráty	CCC, CC	
Default	R	Selhání	D	3-5

Obr. 8 Příklad interního ratingu úvěrů
Zdroj: Česká spořitelna, 2010

4 Porozumění problematice a příprava dat

4.1 Problémová doména

K dispozici máme dvě sady dat, které mají do jisté míry podobné rysy a smysl použití. V případě obou sad nalezneme informace o poskytnutých úvěrech, z nichž některé jsou ukončené s výsledkem, tj. úvěr byl splacen či nesplacen. Liší se od sebe kromě atributů druhem instituce, od které data pochází.

V prvním případě se jedná o data od anonymní banky, která obsahují informace o klientech, z nichž někteří mají úvěr, kartu či oba produkty. Data obsahují transakční záznamy představující obraty na účtech, díky nimž je možné vypočítat charakteristiky klienta. Jedná se o veřejně dostupná data poskytnutá v rámci dolovací soutěže (PKDD'99 Discovery Challenge, 1999). Atributy z transakčních údajů hrají hlavní roli při porovnávání klientů, což obvykle nelze očekávat od tzv. demografických údajů. Pokud je zvoleno více dolovacích úloh, lze z těchto transakčních údajů vybrat takové atributy, které budou relevantní ve více úlohách. Dopředu odhadnout, které atributy budou souviset s cílovou proměnou, však nebývá jednoduché. Pokud klient žádá o určitý typ debetní karty, může být brán v potaz i atribut, který více souvisí s poskytnutím úvěru. Z toho důvodu existují techniky, jako např. Feature selection, umožňující vynechat atributy statisticky bezvýznamné v kontextu zvolené cílové proměnné.

Druhá sada dat pochází od americké společnosti zprostředkovávající nebankovní úvěry, konkrétně tzv. P2P půjčky. Peer-to-peer půjčky se liší od běžných bankovních půjček tím, že dlužník žádá o poskytnutí úvěru od několika věřitelů najednou. Věřitel má dále možnost si rozložit riziko na více dlužníků, tedy poskytnou určitý obnos více žadatelům. Nevýhodou těchto půjček je vyšší úrok oproti klasickým bankovním úvěrům. Dalším mínusem je menší počet atributů pro posouzení spolehlivosti klienta, zejména atributů o transakcích. U těchto půjček lze předpokládat, že o ně žádají klienti, kteří již dříve měli úvěr u banky, ale došlo k problémům se splácením, či měli více otevřených úvěrů. Jde rovněž o veřejně dostupná data. Tyto úvěry spíše reflektují chování amerického trhu, ve kterém se oproti České republice ve větší míře využívají kreditní karty pro financování z cizích zdrojů. (Lending Club, 2015)

4.1.1 Definování cílů a kritérií plnění

S ohledem na dostupné informace v datech byly zvoleny následující obchodní cíle, které lze aplikováním vhodných metod naplnit. Definujme tedy cíle:

- Poskytování úvěru spolehlivým klientům.
- Odhadnutí míry rizika dle spolehlivosti klienta žádajícího o úvěr.
- Udělení správného typu karty klientovi podle kritéria, které má být odvozeno z dat.

Kritéria pro plnění cílů:

- Snížení počtu problémových úvěrů na minimum.
- Správné stanovení úrokové míry - určí rizikovost konkrétního úvěru a zároveň vyčíslí profit, který banka získá v případě bezproblémového splacení úvěru.
- Zvýšení počtu správně přidělených typů karet dle chování klienta.

4.1.2 Inventář dostupných zdrojů

K analýze nám budou sloužit dva počítače. Jeden stroj bude určen k předzpracování dat a shrnutí výsledků do diplomové práce. Druhý stroj s výkonnějším procesorem bude využit k výstavbě modelů. Kontrola stavu generování modelů a případných změn parametrů bude prováděna vzdáleně s použitím nástroje TeamViewer, který funguje na principu vzdálené plochy.

Tab. 1 Inventář použitého HW a SW

Inventář dostupných zdrojů		
Typ stroje	Notebook	Desktop PC
Procesor	Intel Core i3 M330 2,13 GHz	Intel Core i7 4771 3,50 GHz
Počet jader CPU	2	4
RAM	8 GB DDR3 1600 MHz	8 GB DDR3 1600 MHz
Disk	SSD 256 GB	SSD 128 GB
Grafická karta	AMD Mobility HD 5145	GeForce GTX 660
Operační systém	Windows 7 Professional 64-bit	Windows 7 Professional 64-bit
DM nástroj	IBM SPSS Modeler 14.1	
Vzdálené ovládání	TeamViewer	

Data jsou dostupná na lokálním počítači v textové podobě. V případě *bankovních* dat se jedná o více samostatných tabulek (tj. klienti, účty, úvěry, karty atd.) tvořených záznamy (records), které jsou odděleny konci řádků. Jednotlivé atributy záznamů jsou odděleny středníkem. Tabulky se mezi sebou dají propojovat pomocí identifikátorů. Způsob, kterým se atributy (či pole – angl. fields) v SPSS propojují napříč tabulkami, je obdobný principu v databázových systémech. Místo SQL dotazů se v tomto prostředí nejčastěji používá uzel „Merge“, který imituje funkce příkazu JOIN se všemi jeho variantami. Pro kontrolu správného napojení dvou tabulek lze použít samotného pohledu pomocí uzlu „Table“. Dalším způsobem je např. aplikování uzlu „Data audit“, kterým lze zjišťovat, zda se po propojení nezměnil požadovaný celkový počet záznamů, který přebíráme z jedné tabulky (např. tabulka úvěrů a klientů), či zda nevznikly chybějící hodnoty.

Druhá množina *nebankovních půjček* se skládá rovněž z více tabulek, ovšem každá tabulka obsahuje vždy stejné atributy. Záznamy o půjčkách jsou rozděleny chronologicky podle data uzavření úvěru. Pro analýzu je v tomto případě potřeba

záznamy opět sloučit do jediné matice, což se v konkrétním případě provádí horizontálně. K tomuto účelu se používá uzlu „Append“, který slučuje tabulky podle záznamů, které musí mít stejné atributy. Pro slučování dat se tedy nepoužívá žádného klíče, na rozdíl od uzlu „Merge“.

Fáze přípravy dat je obvykle nejdelší a končí v podstatě až ve chvíli, kdy je rozhodnuto o konečné podobě modelů určených k testování. Je typické, že ani fáze výstavby modelů neznamená ukončení etapy předzpracování dat. Z výsledků modelů totiž často vychází najevo, že s daty ještě pořád není něco v pořádku. Nesprávné výsledky nás donutí se vrátit o krok zpět a provádět další úpravy v datech.

4.2 Porozumění datům

4.2.1 Sběr dat

V případě bankovních dat byly kromě transakčních údajů k dispozici data demografická. K těmto atributům patří například míra nezaměstnanosti v daném regionu klienta nebo např. průměrná mzda v regionu. Tyto vlastnosti jsou zajímavé, ale většinou mají velmi malý vliv na výsledek klasifikace. Demografická data je však možné použít opačným způsobem, tedy například uvažovat jako cílovou proměnnou demografický atribut průměrná mzda v regionu, a poté z ostatních atributů zjistit, které mají zásadní vliv na tuto proměnnou. Toto zjištění však nemá pro naše úlohy velký význam.

U nebankovních půjček jsou k dispozici zejména atributy týkající se chování klienta a průběhu splácení úvěru. Demografické atributy nejsou v této sadě vůbec k dispozici, kromě regionu, ze kterého klient pochází. Data by šlo doplnit např. o míru nezaměstnanosti, avšak se nepřepokládá jejich významnost.

4.2.2 Popis dat

Popis atributů původních *bankovních* dat se nachází níže v tabulkách s vypuštěním primárních klíčů a charakteristik, které pro nás nemají význam (např. číslo účtu a označení banky příjemce u transakcí). U atributů jsou kromě názvů zmíněny tzv. rozsahy hodnot, které vyjadřují, jak se mají jednotlivé proměnné chápat. Pokud máme číselný atribut o velkém množství hodnot, je obvyklé jej označit za intervalový (continuous). V případě, že hodnoty atributu jsou spočitatelné např. na prstech rukou, je dobré atribut nastavit jako výčtový (nominal). Rozsah hodnot musí být nastaven před použitím některých úprav, jako je například kategorizace. Kategorizovat lze právě intervalové hodnoty, zatímco nominální atributy lze např. přeuspořádat vytvořením jiných kategorií. Nastavení rozsahu hodnot je v případě modelování obvykle důležitější, než znalost o způsobu uložení hodnot, tedy o datovém typu.

Tab. 2 Atributy bankovních dat

Tabulka	Atributy	Rozsah hodnot	Datový typ	Počet záznamů
Účty	ID regionu pobočky	Interval	Integer	4 500
	Datum zřízení účtu	Interval	Date	
	Frekvence poplatků	Nominální (3)	String	
Klienti	Rodné číslo	Interval	Integer	5 369
	ID regionu klienta	Interval	Integer	
Dispoziční práva	ID klienta	Interval	Integer	5 369
	ID účtu	Interval	Integer	
	Typ dispozičního práva	Binární (2)	String	
Regiony	Název regionu	Nominální (77)	String	77
	Kraj	Nominální (8)	String	
	Počet obyvatel regionu	Interval	Integer	
	Počet obcí do 499 obyvatel	Interval	Integer	
	Počet obcí s 500 až 1 999 obyvateli	Interval	Integer	
	Počet obcí se 2 000 až 9 999 obyvateli	Interval	Integer	
	Počet obcí s více než 10 000 obyvateli	Interval	Integer	
	Počet měst nad 100 000 obyvatel	Interval	Integer	
	Podíl městského obyvatelstva	Interval	Real	
	Průměrná mzda	Interval	Real	
	Míra nezaměstnanosti (1995)	Interval	Real	
	Míra nezaměstnanosti (1996)	Interval	Real	
	Počet podnikatelů na 1000 obyv.	Interval	Integer	
	Počet spáchaných trestných činů (1995)	Interval	Integer	
	Počet spáchaných trestných činů (1996)	Interval	Integer	
Úvěry	ID účtu	Interval	Integer	682
	Datum schválení úvěru	Interval	Date	
	Půjčená částka	Interval	Integer	
	Doba splácení v měsících	Interval	Integer	
	Měsíční splátka	Interval	Integer	
	Status splácení	Nominální (4)	String	
Karty	ID dispozičního práva	Interval	Integer	892
	Typ karty	Nominální (3)	String	
	Datum zavedení karty	Interval	Date	
Trvalé příkazy	ID účtu	Interval	Integer	6 471
	Odepsaná částka	Interval	Real	
	Druh platby	Nominální (4)	String	
Transakce	ID účtu	Interval	Integer	1 056 320
	Datum transakce	Interval	Date	
	Označení příjmové či výdajové položky	Binární (2)	String	
	Mód transakce – např. výběr kartou	Nominální (5)	String	
	Částka	Interval	Real	
	Zůstatek po transakci	Interval	Real	
	Charakter transakce – např. SIPO	Nominální (7)	String	

Tabulka s popsanými atributy bankovních dat tedy tvoří vstupní data, která budou následně předzpracována. Celkový počet atributů poté významně naroste, protože řada charakteristik klientů se ukrývá v transakčních údajích. Ty představují veškeré bankovní operace nad účty klientů, z nichž budou různými způsoby odvozeny, přesněji vypočteny, potenciálně použitelné proměnné.

Data o *nebankovních* půjčkách tvoří po sloučení všech záznamů jednu matici o cca půl milionu úvěrů. Názvy atributů byly podle dostupného popisu v angličtině přeloženy za účelem usnadnění procesu porozumění datům. Na rozdíl od bankovních dat se počet atributů příliš nezmění, případně pouze sníží.

Tab. 3 Atributy nebankovních půjček

Atributy	Rozsah hodnot	Datový typ	Počet záznamů
ID klienta	Interval	Integer	
Půjčená částka	Interval	Integer	
Počet splátek	Interval	Integer	
Úroková míra	Interval	Real	
Měsíční splátka	Interval	Real	
Rating úvěru	Nominální (7)	String	
Délka zaměstnání	Interval	Integer	
Vlastnictví nemovitosti	Nominální (5)	String	
Roční příjem	Interval	Integer	
Splátkový kalendář	Binární (2)	String	
Účel úvěru	Nominální (14)	String	
Stát USA	Nominální (50)	String	
ZIP kód	Interval	Integer	
Koeficient měsíční splátky dluhů/měsíční mzda	Interval	Real	512 340
Počet kreditních prohřešků za poslední 2 roky	Interval	Integer	
Počet přihoršujících veřejných záznamů	Interval	Integer	
Kreditní skóre klienta – dolní mez	Interval	Integer	
Kreditní skóre klienta – horní mez	Interval	Integer	
Počet měsíců od poslední delikvence	Interval	Integer	
Počet měsíců od posledního veřejného záznamu	Interval	Integer	
Počet měsíců od hlavní nepříznivé události	Interval	Integer	
Počet otevřených kreditních linií	Interval	Integer	
Celkový počet kreditních linií	Interval	Integer	
Celkový objem revolvingových úvěrů	Interval	Integer	
Míra využití revolvingových úvěrů	Interval	Real	
Datum schválení úvěru	Interval	Date	
Datum zavedení první kreditní karty	Interval	Date	
Status úvěru	Nominální (7)	String	

4.2.3 Kontrola kvality dat

Nezbytným krokem, který má za cíl eliminovat možné chyby ve vystavěných modelech, je prověření kvality dat. V SPSS se používá pro základní a rychlý způsob kontroly kvality dat uzlu Data audit, díky kterému je možné sledovat charakteristiky polohy a variability, šikmosti, množství chybějících, odlehlých a extrémních hodnot. Dále je možné pomocí tohoto uzlu sledovat četnosti hodnot v jednotlivých atributech, což může vypovídat o jejich použitelnosti. Řadu „nekvalitních" a nesouvisejících atributů lze odfiltrovat pomocí algoritmu Feature selection.

Chybějící a prázdné hodnoty

V případě obou množin nebylo zapotřebí odstraňovat řádky s chybějícími hodnotami. U *bankovní* sady chyběly některé údaje pouze mezi demografickými daty. Po dohledání konkrétních případů bylo zjištěno, že nejlepší cestou bude doplnění chybějících hodnot pomocí metod odhadu. K tomuto účelu lze použít například lineární regresi, regresní stromy, neuronové sítě atd.

V případě *nebankovních* dat nebyly nalezeny žádné chybějící hodnoty v tom smyslu, že by údaje skutečně chyběly, protože se jednalo spíše o prázdný údaj. Pokud údaje neexistovaly, nejčastěji se jednalo o počty prohřešků klientů, kde v případě včasného splácení tento údaj samozřejmě chyběl. Nedefinované hodnoty byly v těchto případech nejčastěji nahrazeny nulou nebo upraveny tak, aby hodnoty zůstaly konzistentní.

Nekonzistence dat

U *bankovních* dat byly objeveny mírné nekonzistence v hodnotách atributů transakčních údajů. Typicky šlo například o záměnu hodnoty „výdaj" za „výběr", přičemž význam těchto hodnot u konkrétního atributu byl nadále zjevný, tudíž bylo chybu jednoduché napravit.

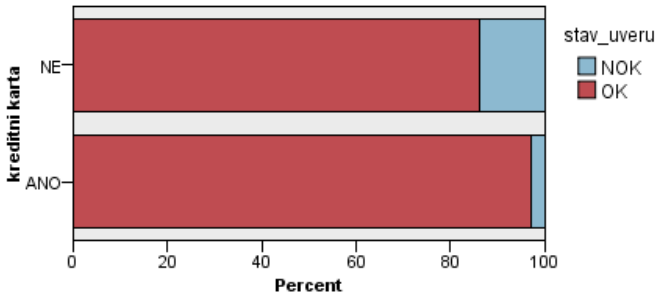
V případě *nebankovních půjček* bylo nekonzistentních zadání hodnot více. Za možnou příčinu nekonzistence je považována změna v kódovacím schématu hodnot v průběhu času[2]. Problémy se týkaly atributů o počtu měsíců od posledních prohřešků. U úvěrů z let 2007 byly prázdné hodnoty označeny jako NULL, tedy hodnotou nedefinovanou, zatímco u novějších úvěrů byly tyto hodnoty značeny takto - "". Díky tomu nebylo možné záznamy z let 2007 sloučit s ostatními bez potřebných úprav.

4.2.4 Průzkum dat

Nejjednodušší způsob, jakým lze získat co nejvíce informací o datech a urychlit tak celý proces přípravy, je data vizualizovat či agregovat. Kromě popisu atributů lze získat prvotní informace pomocí různých agregačních tabulek či grafů. SPSS nabízí řadu nástrojů pro podporu této fáze.

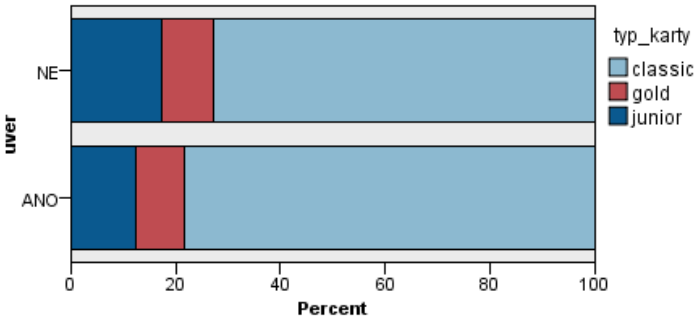
[2] Informace o úvěrech pochází z období 2007 až 2015

Z hlediska vytyčených cílů nás může zajímat co nejvíce informací o úvěrech a kartách. Vzhledem k tomu, že řada atributů ještě nebyla v této etapě známá, musíme pracovat pouze s původními. V první fázi nás může zajímat, zda je vhodné v případě žádosti o půjčku uvažovat atributy, týkajících se karet a opačně. Tato povrchní analýza byla provedena nad sadou bankovních dat.



Obr. 9 Podíl problémových úvěrů u klientů s kartou a bez ní

Na obrázku Obr. 9 je vidět, že u klientů s kartou se vyskytuje méně problémových úvěrů. Může to tedy znamenat, že pokud klient kartu vlastní, bude více pravděpodobné, že úvěr splatí. Jinými slovy, atribut týkající se používaní či nepoužívání karet, může mít roli při posuzování žádosti o poskytnutí úvěru. Podívejme se nyní na obrácenou situaci, zda při výběru vhodného typu karty má smysl uvažovat atribut týkající se úvěru.

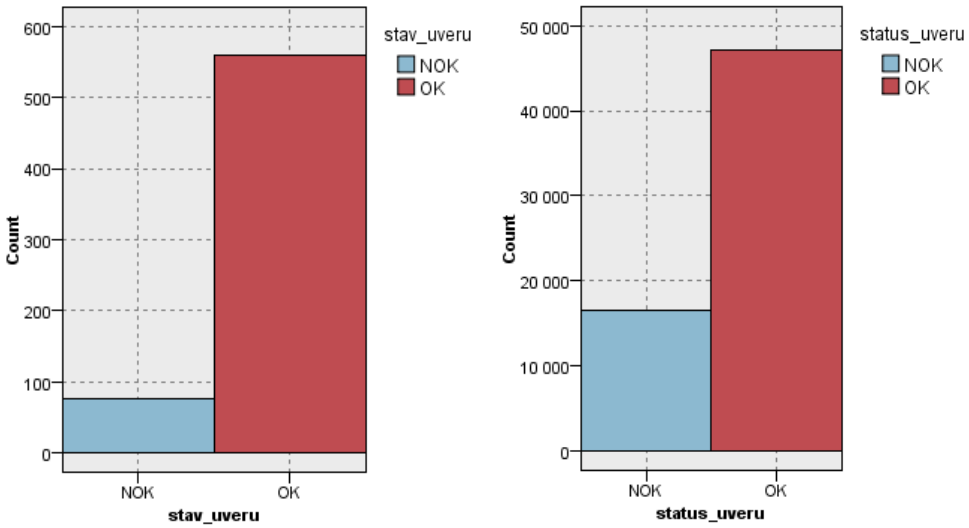


Obr. 10 Podíly typů karet v závislosti na poskytnutí úvěru

Z Obr. 10 je vidět, že rozložení typů karet mezi klienty s úvěrem se skoro neliší od klientů bez něj. Existence úvěru tedy nemusí mít takřka žádný vliv na přidělení vhodného typu karty. Před zahájením modelování bude provedena redukce atributů pomocí algoritmu Feature selection, který zajistí vyřazení atributů týkajících se např. úvěrů, ale nesouvisejících s kartami.

Klasifikace úvěrů

Pro tuto úlohu budou použita data bankovní i nebankovní. V první řadě bude zkoumáno rozložení záznamu do tříd cílových proměnných v kontextu vytyčených dolovacích úloh. Počty problémových (NOK) a bezproblémových (OK) úvěrů u obou sad jsou následující:



Obr. 11 Výchozí počty úvěrů bankovních (vlevo) a nebankovních (vpravo)

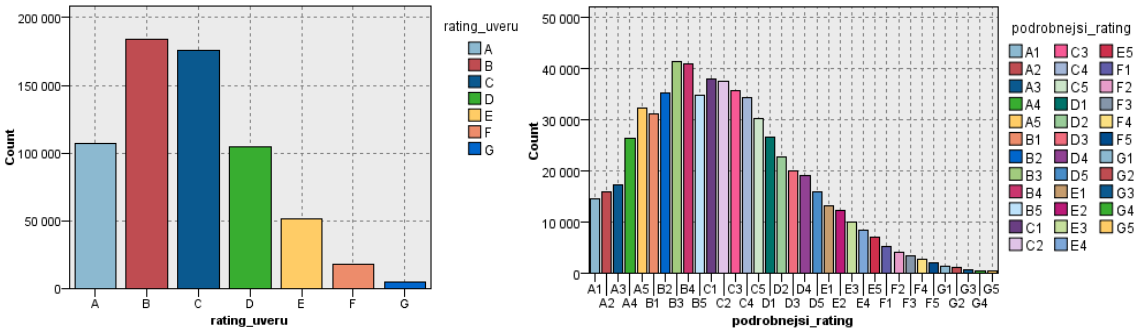
U bankovních půjček byl problém s malým počtem problémových úvěrů. Nebylo tedy možné vyrovnat počty záznamů ve třídách redukcí. Pro odhalování vzorů v datech je často doporučováno mít vyrovnaný počet záznamů ve všech třídách, z toho důvodu bude dobré počet problémových úvěrů zvýšit duplikováním záznamů. Model tak nebude mít tendenci vyhodnocovat podle třídy, ve které je více záznamů. Toto zvýšení počtu záznamů může mít kladný vliv na přesnost modelu, která by byla v případě redukce obvykle nižší z důvodu ztráty části trénovacích dat. Vyšší přesnosti je očekáváno i z důvodu výrazně převyšující množství atributů oproti nebankovním půjčkám.

U nebankovních úvěrů rovněž převažuje počet bezproblémových. Záznamů je však dost v obou třídách, a proto není problém záznamy zredukovat pro dosažení vyrovnaného počtu v obou třídách. Tato data se liší od bankovních úvěrů v počtu různých délky úvěrů dle splácení. V sadě nebankovních dat jsou k dispozici pouze úvěry tříleté nebo pětileté. Zde se nabízí otázka, zda nevytvořit souběžně dva klasifikační modely pro oba typy úvěrů, vybrat pouze jeden nebo zahrnutou oba typy najednou. Počty úvěrů na Obr. 11 nejsou původní. V sadě se nacházelo sedm stavů splácení úvěru, kdy čtyři z nich byly pro tuto úlohu označeny jako problémové. Jednalo se o úvěry, u kterých se klienti dopustili k pochybení v podobě zpožděného splácení, či k úplné neschopnosti půjčku platit. Jeden typ úvěrů, tzv. odepsaných byl vyloučen zcela, protože neměl v kontextu této úlohy význam. Zbyly úvěry splacené

a neukončené. Běžící úvěry byly podle potřeby označeny za bezproblémové, či vyloučeny. Více o řešení způsobu rozložení záznamů do tříd je ve fázi přípravy dat.

Určení rizikovosti úvěru

Dalším cílem je zpětně zjistit, jakým způsobem byla určována rizikovost úvěru z předem známého zatřídění. Rizikovost jednotlivých úvěrů je značena atributem *rating úvěru* a je obvykle vyjádřena konkrétní výší *úrokové míry*. Rating je značen písmeny A až G a představuje jakési pásmo hodnot úrokových měr v širším rozsahu. Užší pásmo je specifikováno v atributu *podrobnější rating*. Samotné hodnoty úrokových měr jsou však stanoveny pro konkrétní úvěry, které náleží do jednoho pásma. Protože úroková míra má spojitý charakter, nabízí možnost využití regrese. Při konkrétním počtu hodnot cílového atributu však lze předpokládat vyšší pravděpodobnost, že se natrénovaný model „trefí“ do správné hodnoty známé v testovací množině. Pro pomyslné určení úrokové míry bude zvolena klasifikace s použitím jednoho z ratingových atributů jako cílové proměnné. Dále lze předpokládat, že s nižším počtem hodnot cílové proměnné bude snadnější postavit přesnější model. Atribut *rating úvěru* by byl z tohoto pohledu vhodnější, protože má nižší počet hodnot než *podrobnější rating*. Dalším důvodem pro volbu jednoduššího ratingu je možnost jeho použití pro nastavení rovného počtu úvěrů ve všech třídách. Je celkem pravděpodobné, že by se nám nepodařilo rozdělit záznamy ve všech úrokových třídách rovnoměrně, aniž by počet záznamů v jednotlivých třídách nebyl příliš malý, pokud bychom vybrali podrobnější variantu. O původním rozložení úvěrů dle ratingu vypovídá Obr. 12[3].



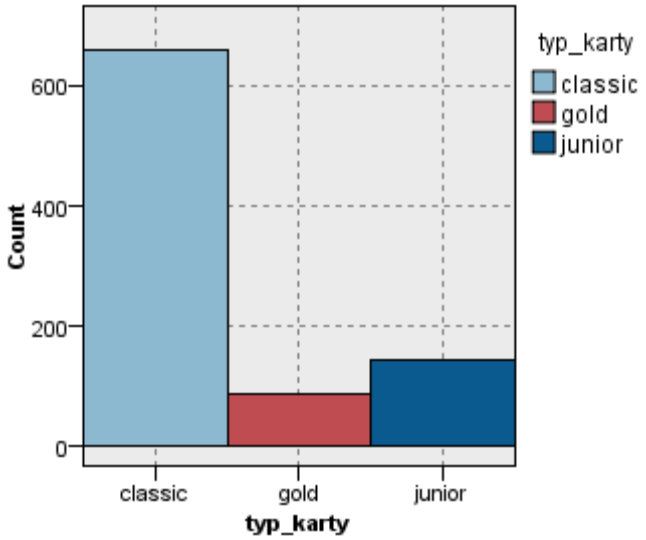
Obr. 12 Počty úvěrů dle atributu „rating úvěru“ a „podrobnější rating“.

Rovnoměrné rozložení úvěrů se bude řídit počtem záznamů v nejméně četné třídě. Z grafů je vidět, že v tomto souboru dat výrazně převažují méně rizikové úvěry, které jsou častěji také bezproblémové. Pro úlohu klasifikace rizikovosti úvěru budou použita data nebankovních půjček.

Přidělení typu karet

[3] Jsou uvažovány veškeré úvěry bez ohledu na status splácení.

K naplnění cíle této úlohy budou použita data bankovní. Po průzkumu dat bylo zjištěno, že typy karet, které banka rozdělovala, jsou *junior*, *classic* a *gold*. U karty junior je dle názvu evidentní, že přidělení bude souviset s věkem, a po dosažení určité věkové hranice už nebude možné tuto kartu získat. Je tedy možné, že atribut věk vyjde jako určující pro přidělování typu karty, což by nebylo příliš vypovídající. Hlavní kritérium pro doporučení typu *classic* nebo *gold* pravděpodobně nebude souviset s věkem. Z tohoto důvodu bude tento atribut vyloučen při sestavování klasifikačního modelu. Model bude více zaměřen na chování klienta z hlediska operací na účtu. Vzhledem k malému počtu záznamů o kartách bude výchozí rozložení záznamů dle typu karty doplněno o duplikované záznamy (viz Obr. 13).



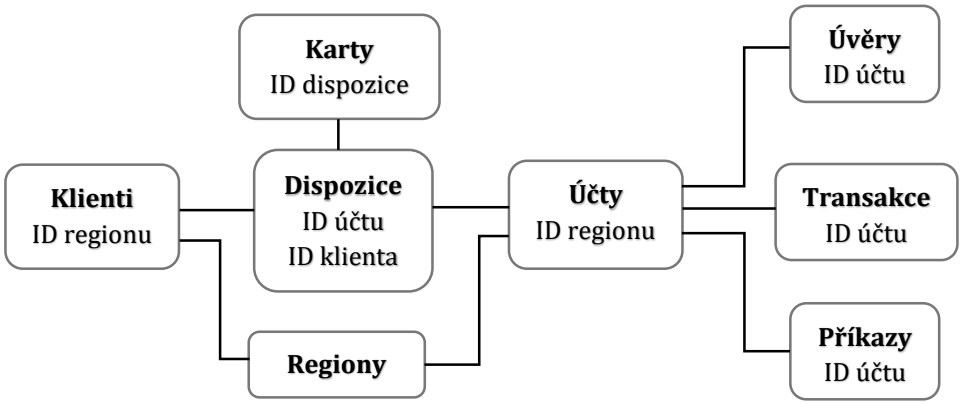
Obr. 13 Rozdělení záznamů o kartách dle typu karty

Na závěr kapitoly je ještě potřeba zmínit znění cílů jednotlivých dataminingových úloh:

1. *Vytvoření klasifikačních modelů pro přidělování či zamítání úvěrů a zjištění klíčových atributů.*
   - *1.1. Vytvoření shluků klientů a vizualizace v grafu dle počtu problémových úvěrů.*
2. *Vytvoření klasifikačních modelů pro rating úvěrů a zjištění klíčových atributů.*
   - *2.1. Vytvoření shluků klientů a vizualizace v grafu podle úrovně úrokové míry.*
3. *Vytvoření klasifikačních modelů pro přidělování typu karty a zjištění klíčových atributů.*
   - *3.1. Vytvoření shluků klientů a vizualizace v grafu dle průměrné vybrané částky.*

4.3 Příprava dat

4.3.1 Bankovní data - účty, úvěry a karty



Obr. 14 Schéma propojení původních tabulek bankovních dat

U bankovních dat se většina sledovaných atributů vztahuje k účtu. Například transakční údaje obvykle nelze přidružit k jednotlivým klientům, ale právě k účtu. Důkazem toho je atribut o průměrném zůstatku na účtu, který vychází z výpočtu nad operacemi zahrnující změny na účtu podle toho, zda přichází příjmová nebo výdajová položka. U jednotlivých položek nelze obvykle určit, zda se týkají vlastníka účtu či disponenta. Výsledkem je tedy jedna velká matice záznamů účtů se stovkami atributů, ze kterých se vybere potřebná množina dat vzhledem k vytyčenému cíli. Vybraná množina pak představuje klienty a jejich parametry.

Každý úvěr je vztažen k právě jednomu účtu a na každém účtu je maximálně jeden úvěr. Stejná situace je u karet. Pokud by bylo úvěrů nebo karet k jednotlivým účtům více, musely by se např. stejné informace o jednom účtu připojit vícekrát k jinému záznamu o úvěru či kartě. V takovém případě by vznikly např. dva záznamy o úvěrech, které by se až na požadovaný obnos peněz či typ karty nijak nelišily. Tato situace však nenastala.

U úvěrů bylo potřeba nejdříve pomocí uzlu Reclassify změnit původní třídy na OK a NOK. Výhodou SPSS je, že umí pracovat s datovým typem DATE či DATETIME, díky čemuž lze prostřednictvím CLEM[4] výrazů například vypočítat časový rozdíl mezi dvěma okamžiky ve dnech, měsících, rocích, ale třeba i v sekundách. Z informace o datu schválení půjčky a délce úvěru bylo určeno plánované datum, dokdy má být úvěr splacen. Podobně bylo postupováno při výpočtu doby od zavedení karty, která může mít význam např. při žádosti o úvěr. Informace o kartách a úvěrech poté byly sloučeny dohromady.

[4] Zkratka CLEM - The Control Language for Expression Manipulation

4.3.2 Bankovní data – transakce

V první řadě bylo nutné upravit nekonzistence v pojmenování účtovaných položek, tak aby šly od sebe jednoznačně odlišit kvůli výpočtu atributů. Soubor byl rozdělen na příjmovou a výdajovou část, odkud byly odstraněny tyto nekonzistence v pojmenování typů položek. Následně byly záznamy opět sloučeny pomocí Append a seřazeny podle ID účtu a data. U každé položky jsou sledovány následující atributy:

- ID účtu
- datum transakce
- obnos
- zůstatek po operaci
- příjem/výdaj
- mód transakce
- symbol operace

Tab. 4 Typy příjmových a výdajových položek u transakcí

Příjem/Výdaj	**Typ položky**
Příjem	Vklad na účet Převod ve prospěch účtu – mzda či jiný příjem Důchod Připsaný bonusový úrok
Výdaj	Poplatek za služby Sankční úrok – opožděné splácení SIPO – Soustředění inkaso plateb obyvatelstva Pojistné Výběr z účtu u přepážky Výběr kartou Převod na vrub Splátka úvěru

Zůstatek

Při zjišťování charakteristik kolem zůstatku bylo v první řadě potřeba vyloučit z výpočtu položky nevhodné pro kombinaci s možnými cílovými proměnnými. Pokud má klasifikační model sloužit k rozhodování o poskytování půjček klientů, kteří o úvěr teprve žádají, pak nelze do výpočtu průměrného zůstatku historických klientů zahrnout i splátky úvěru, případně sankční úrok. Zůstatek by se totiž u žádajícího klienta doposud bez úvěru mohl lišit, přičemž by teoreticky mohlo jít o stejný případ, jako byl trénovací. V případě klasifikační úlohy o udělení typu karty nás však může zajímat zůstatek včetně sankčních úroků za pozdní splácení úvěru, které ani tehdy v zůstatku nebudou započteny. Jsou evidovány samostatně v atributu týkajícího se

sankcí. V případě *úvěrů* jsou vypuštěny *splátky úvěru* a *sankční úroky*. U *karet* se jedná o veškeré atributy týkající se *výběru kartou*. Tyto položky tedy vůbec nebudou zahrnuty do výpočtu průměrného zůstatku.

Nyní přejděme k samotnému výpočtu. V první řadě je potřeba vypočítat zůstatek pro dny, ve kterých proběhly nějaké operace. Po průzkumu dat bylo zjištěno, že funkční metodou pro stanovení zůstatku v konkrétní den, je tento výraz[5]:

```
if prijmy_Sum>0 then
        balance_Max-vydaje_Sum
else balance_Min endif
```

prijmy_Sum	součet všech příjmových položek v daný den
balance_Max	maximální zůstatek po přičtení všech příjmových položek
vydaje_Sum	součet všech výdajových položek v daný den
balance_Min	minimální zůstatek po odečtení všech výdajových položek

Po zjištění zůstatku v daných dnech máme více možností, jak pokračovat. Můžeme počítat *zůstatek za celé období aktivity* klienta nebo jen např. *za poslední dva roky*. Bylo rozhodnuto, že zůstatek bude vypočten pro obě délky období s využitím dvou odlišných přístupů, protože oba přináší jiné možnosti využití.

```
if @OFFSET(account_id__1,1)=undef then 0 else
        elseif account_id=account_id__1 then
                date_days_difference(datum_transakce__1,datum_transakce)
        else 0 endif
```

Pro postup výpočtu *zůstatku za celé období aktivity* byl napsán uvedený výraz sloužící k určení časového rozdílu mezi dvěma po sobě jdoucími dny, ve kterých proběhly změny zůstatku. Pro zkrácení zápisu tohoto výrazu byl použit uzel History, který umožňuje duplikovat hodnoty vybraného atributu přechozího záznamu do připravené proměnné v následujícím záznamu. Výpočet časového rozdílu lze realizovat u výpočtů nad různými operacemi, tedy nejen u změn zůstatků. Lze jej použít jako atribut ve formě průměrného časového rozdílu, vyjadřujícího periodicitu operací daného typu – např. každých 30 dní.

```
if @OFFSET(account_id__1,1)=undef then balance_Mean else
        elseif account_id=account_id__1 then
                balance_Mean+(casovy_rozdil-1)*balance_Mean__1
        else balance_Mean endif
```

Tento výraz násobí poslední známý zůstatek počtem dnů od poslední změny zůstatku a přičte aktuální stav zůstatku v daném dni. Poté se pomocí uzlu Aggregate

[5] Tento výraz počítá s variantou, že nejdříve jsou evidovány všechny položky příjmové.

sečtou hodnoty pomocného atributu u všech účtů. Průměrný *zůstatek za celé období aktivity* získáme vydělením výsledku počtem dnů sledovaného období (suma časových rozdílů).

Nevýhodou této metody je nemožnost určit např. minimální, maximální zůstatek, směrodatnou odchylku či průměr samotný napříč měsíci v roce. V SPSS však existuje uzel Time intervals (dále TI), který umožňuje rozprostřít časově provázané záznamy do tabulky se stanoveným krokem (např. vteřina, den, rok) v určitém časovém intervalu. Tento postup bude použit pro *výpočet zůstatku za poslední dva roky.*

Můžeme navázat na část týkající se výpočtu průměrného zůstatku v konkrétní dny. Před použitím uzlu TI je však potřeba provést několik kroků. Nejprve odfiltrujeme sloupce, které nebudeme pro výpočet potřebovat. Musíme totiž matici dostat do podoby, ve které jediný identifikátor záznamu v řádku bude datum a kde všechny účty budou převedeny na sloupce. Pomocí uzlu Restructure je nedříve odděleno datum transakce od záznamu a převedeno na sloupce. Důvodem proč nyní převádíme na sloupce datum místo účtů je, že uzel Restructure neumí najednou vytvořit více než 255 atributů. Počet dnů za dva roky je nižší než počet účtu v tabulce, bude tedy rychlejší převést postupně všechna data transakci na sloupce a poté matici transponovat pomocí Transpose. Nyní máme matici, kde jsou na řádcích data transakcí a ve sloupcích účty. Po provedení několika dalších úprav je možné použít uzel TI. Výsledkem je rozprostření změn zůstatků u všech účtů ve zvoleném časovém měřítku pro jednotlivé dny. Nyní stačí duplikovat poslední známé zůstatky směrem dolů. Poté je možné začít agregovat podle měsíců, let a postupně počítat požadované charakteristiky. Po několika dalších úpravách je možné transponovat matici tak, že na řádcích budou opět identifikátory účtu a ve sloupcích nové atributy.

Funkce Time intervals je jistě přínosná, ale je potřeba zvážit, zda je její použití opodstatněné. Vzhledem k multiplikování počtu atributů lze očekávat vysoké nároky na systémové zdroje, které se ještě zvýší s větším množstvím záznamů.

Tab. 5 Atributy týkající se zůstatku

Délka období	**Zůstatek – atributy**
Od počátku aktivity	Průměr Celkem Počet operací celkem Průměrný počet operací najednou (v určitý den) Délka sledovaného období (ve dnech) Průměrné časové rozmezí (ve dnech) Aktivní dny (počet)
Měsíčně za 2 roky	Měsíčně průměr Měsíčně maximum Měsíčně minimum Měsíčně odchylka

Ostatní typy operací

Mezi ostatní typy operací budeme řadit různé typy položek a navíc celkové příjmy a výdaje. U výdajů bude zavedeno podobné omezení, jako u zůstatků, tedy že do nich nebudou zahrnuty položky týkající se úvěrů a karet. U některých typů operací budou sledovány pouze atributy za celé období aktivity, nikoliv měsíční za poslední dva roky. Týká se to především těch operací, jejichž průměrná hodnota zůstávala po celou dobu stejná. Dobrým kritériem pro rozhodnutí, zda počítat u daného typu operace navíc ještě průběh v měsících byla směrodatná odchylka od průměru. Jestliže byla ve většině případů nulová, pak nemělo smysl počítat průměrné měsíční atributy. Třeba minimum a maximum nemá při nulové odchylce žádný význam.

Pokud je vhodné měsíční charakteristiky vypočítat, situace je oproti zůstatku snazší v tom, že není potřeba použít komplikovanou metodu typu Time intervals. Pokud záznamy v daném měsíci nejsou, pak se automaticky počítá s nulou a měsíce se sčítají. Následně se pomocí uzlu Aggregate spočítá průměr, minimum, maximum, směrodatná odchylka a počet operací za dva roky. Hodnoty je pak nutno upravit v případě, že záznamy za některé měsíce chybí. Průměr se nahradí sumou záznamů za rok vydělenou dvanácti místo původního počtu měsíců, ve kterých existovaly údaje. Za minimum se uvažuje nula. Dále je možné *přepočítat směrodatnou odchylku* na počet měsíců kalendářního roku. SPSS používá směrodatnou odchylku s výběrem n – 1. Lze ji přepočítat podle následující formule[6]:

*sqrt((amount_Sum_SDev*amount_Sum_SDev*(pocet_mesicu-1)+*
*amount_Sum_Mean*amount_Sum_Mean*pocet_mesicu-amount_Sum_Sum*amount_Sum_Sum*12)/11)*

Je potřeba zmínit, že u těchto operací je oproti zůstatku velký rozdíl mezi průměrnými hodnotami atributů za celé období a průměry měsíčně za dva roky. U zůstatku se totiž v případě obou atributů jedná o průměr z průměru, čili výsledky jsou v podstatě stejné. Rozdíl v hodnotách je dán hlavně odlišnou délkou sledovaného období. U ostatních položek se však jedná o odlišné atributy. Název *průměr* představuje aritmetický průměr nad všemi záznamy dané operace na účtu. *Průměry měsíčně* jsou vypočteny jako průměry celkových částek v jednotlivých měsících.

Tab. 6 Rozdíl hodnot atributů u zůstatku a výběrů

Zůstatek průměr	Zůstatek měsíčně průměr	Výběry průměr	Výběry měsíčně průměr	Výběry měsíčně počet
16 269	15 181	1 481	1 445	1.125
36 707	41 736	5 759	16 077	2.916
27 882	29 881	2 996	4 561	1.291
20 781	22 718	1 795	1 900	1.166
22 978	26 205	1 508	2 706	1.791

[6] Proměnná *pocet_mesicu* ve vzorci odpovídá počtu měsíců, kdy proběhla alespoň jedna operace daného typu

Tab. 7 Přehled sledovaných atributů u ostatních operací

Délka období	Typ atributu	Operace
Od začátku aktivity	Celkem Průměr Odchylka Počet operací Průměrný počet Sledované období Průměrné časové rozmezí Aktivní dny	Vklady na účet Převody ve prospěch účtu Důchody Připsané bonusové úroky Poplatky za služby Sankc. úrok SIPO Pojistné Výběry z účtu Výběry kartou Převody na vrub Splátky úvěru Příjmy Výdaje
Měsíčně za 2 roky	Měsíčně průměr Měsíčně minimum Měsíčně maximum Měsíčně odchylka Měsíčně počet	Vklady na účet Převody ve prospěch účtu Připsané bonusové úroky Výběry z účtu Výběry kartou Příjmy Výdaje

Dle Tab. 7 je možné vidět, že měsíční atributy byly uvažovány pouze u sedmi typů operací kromě zůstatku. Jednalo se o operace, které probíhaly často a v nepravidelných periodách. Obnos částky byl až na výjimky pokaždé jiný. U převodů ve prospěch se obvykle jednalo o pravidelné příchozí platby, které se občas opakovaly.

4.3.3 Bankovní data - další tabulky

Klienti a dispoziční práva

Z tabulky klientů bylo na základě rodného čísla určeno datum narození a pohlaví klienta. Podle tabulky s dispozičními právy byla ke klientům přidána informace, zda k účtu, ke kterému klient náleží, existuje disponent. Disponent je druhý uživatel účtu, např. manželka či manžel, který má pouze disponentská práva.

Trvalé příkazy

Tabulka obsahuje záznamy o zřízených trvalých příkazech k účtům. Byly identifikovány čtyři typy příkazů: SIPO, pojistné, úvěr a leasing. Ostatní příkazy byly nespecifikované. Pokud bylo k některému účtu více trvalých příkazů stejného typu, pak byl spočten průměr pomocí uzlu Aggregate. Opět se jednalo o charakteristiky související s účty.

Demografické údaje

Demografické údaje zahrnovaly šestnáct atributů. Nacházely se mezi nimi chybějící hodnoty, a to u dvou atributů. Pro doplnění byla zvolena metoda odhadu. Pomocí Feature selection byly vybrány atributy, které mají vliv na odhadovanou cílovou proměnnou. Konkrétně šlo například o chybějící údaj v míře nezaměstnanosti za rok 1995. Feature selection vybral v kontextu této proměnné následující čtyři atributy:

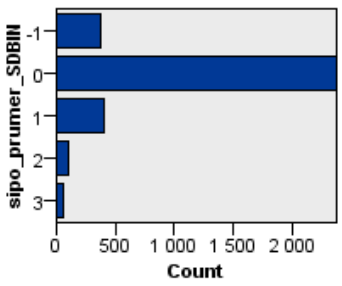
- míra nezaměstnanosti v roce 1996
- kraj
- počet živnostníků na 1 000 obyvatel
- počet obcí s více než 10 000 obyvateli

Pro odhad chybějících hodnot byl použit boostovaný regresní strom s nastavením zmíněných vstupních prediktorů. Demografické údaje byly po odstranění chybějících hodnot připojeny k jednotlivým účtům. Byly však párovány dvakrát. V prvním případě jako atribut vlastníka účtu, podruhé jako místo, kde byl zřízen bankovní účet, či kde se nachází bankovní poradce.

4.3.4 Další úpravy

U obou datových množin, tj. u bankovní i nebankovní, byly pro vybrané číselné intervalové hodnoty vytvořeny jejich kategorizované varianty. Z těchto vytvořených atributů mohou být některé z nich označeny jako cílové proměnné. Vzhledem k velkému množství atributů lze data do budoucna použít k dalším analýzám prostřednictvím kategorizovaných proměnných. Tyto atributy byly dále použity při shlukování pro vyznačení případů u jednotlivých shluků v grafu. Kategorizované atributy však nebyly použity pří výstavbě modelů, protože jsme chtěli pracovat pouze s původními hodnotami.

Nové hodnoty	Dolní hranice	Horní hranice
-3		< -17108,82482962
-2	>= -17108,82482962	< -7819,48988641
-1	>= -7819,48988641	< 1469,84505679
0	>= 1469,84505679	<= 20048,51494321
1	> 20048,51494321	<= 29337,84988641
2	> 29337,84988641	<= 38627,18482962



Obr. 15 Kategorizace proměnné SIPO průměr

Obr. 15 znázorňuje příklad kategorizace odvozené proměnné SIPO průměr. Většinou byla zvolena metoda diskretizace podle průměrné hodnoty atributu a jeho směrodatných odchylek. Z kategorizace byly vyloučeny nedefinované hodnoty, čímž se

zabezpečí nezkreslení průměrných hodnot, ke kterému by došlo případným nahrazením prázdných hodnot nulou. Proměnné kategorizované metodou směrodatných odchylek jsou v názvu doplněny příznakem SDBIN (Standard Deviation Binning). Pomocí takto vytvořených hodnot lze jednoduše určit, v jaké pozici je hodnota daného atributu vzhledem k ostatním případům, tj. zda je průměrná, nadprůměrná či podprůměrná.

Bankovní data

Zbývající úpravy spočívaly v odvozování dalších atributů. Byla zavedena proměnná *podíl aktivních dnů* u jednotlivých typů operací. Vyjadřuje relativní množství počtů dnů, ve kterých probíhala daná operace, vzhledem ke stáří účtu. Tento atribut vyjadřuje míru aktivity na účtu dle typu operace. Tab. 8 obsahuje záznamy z pěti různých účtů, kde je možné porovnávat aktivity na účtech dle typu operace. Podíl aktivity zůstatku v sobě obsahuje všechny ostatní aktivity, kromě výběrů karet, sankčních úroků a splátek úvěru. Zajímavým zjištěním je, že více často probíhaly výdajové operace místo příjmových.

Tab. 8 Aktivita na účtech dle typu operace (v %)

Zůstatek	Výdaje	Příjmy	Výběry	Úroky bonus.	Služby	SIPO	Vklady	Důchody
13,488	9,500	7,324	3,916	3,191	2,828	2,828	0,798	0,000
17,283	13,677	7,026	7,869	3,279	3,044	3,044	0,562	0,000
18,416	12,707	8,287	3,315	3,315	2,394	2,394	5,157	0,000
14,450	11,196	6,220	3,062	2,871	2,775	2,775	0,096	3,158
11,876	8,262	6,024	3,442	2,582	2,410	2,410	0,172	3,270

Následně je možné použít Data audit pro kontrolu extrémních a odlehlých hodnot, který zjistí jejich počet. Problémy může řešit několika způsoby. Nevyhovující záznamy je možné odstranit nebo např. zaměnit extrémní či odlehlé hodnoty za nejbližší, které nejsou považovány za extrém. Před klasifikací nebudou použity žádné podobné úpravy v podobě vylučování záznamů, či doplňování jiných hodnot z důvodu možných ztrát směrodatných informací. Před shlukováním je však použito algoritmu pro detekci anomálií, které by mohly mít negativní vliv na vytváření shluků.

Nebankovní data

Po odstranění nekonzistencí v hodnotách atributů bylo možné sloučit dohromady čtyři CSV soubory. Poté následovalo vylučování nerelevantní atributů. Data obsahovala mnoho informací, kterých se nedalo nijak využít. Jednalo se o různé atributy týkající se stavu splácení půjčky vztažené ke konkrétnímu datu (červen 2015), jako například množství zaplacených splátek. Klienti žádající o úvěr doplňovali kromě výběru kategorie úvěru detailnější popis účelu. Zde se nabízela možnost vytvoření kódovacího schématu, ale většina sdělení oproti vybrané kategorii nepřinášela no-

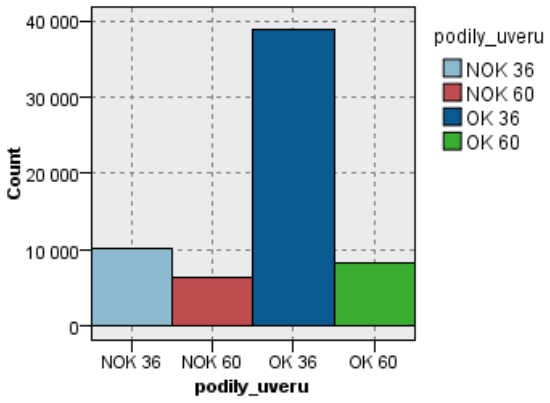
vou informaci nebo byla dokonce v rozporu. Poměrně užitečný atribut mohl vzniknout z popisu pracovní pozice, nicméně množství hodnot bylo pro zakódování příliš velké.

Veškeré hodnoty včetně atributů byly přeloženy do češtiny. Dále byly doplněny prázdné hodnoty. Například u počtu měsíců od posledního prohřešku byla zadána hodnota „-1“, pokud žádný prohřešek neexistoval. Chybějící hodnoty v pravém slova smyslu se v souboru nevyskytovaly. Na konci úprav byla provedena kategorizace hodnot podobným způsobem, jako u transakčních údajů v případě bankovních dat.

Z data schválení úvěru bylo odvozeno datum, do kdy má být úvěr uhrazen. Následně bylo možné určit zbývající počet měsíců do konce splácení. Tento atribut posloužil k omezení „mladých“ běžících úvěrů a dále také ke snížení počtu předčasně splacených úvěrů. Omezení bylo provedeno pomocí uzlu Select v režimu Discart s následující formulí:

(data="pujcky" and (term-pocet_mesicu_do_konce_uveru)<term and loan_status="Fully Paid" and term=36) or
(data="pujcky" and loan_status="Current" and term=36) or
*(data="pujcky" and (term-pocet_mesicu_do_konce_uveru)<(term*0.7) and loan_status="Current" and term=60) or*
*(data="pujcky" and (term-pocet_mesicu_do_konce_uveru)<(term*0.7) and loan_status="Fully Paid" and term=60)*

Výše uvedený výraz odstraňuje úvěry s délkou 36 měsíců, které byly splaceny před dohodnutým datem. Dále v případě úvěrů na 36 měsíců jsou vypuštěny veškeré běžící půjčky, protože jsou upřednostněny dokončené splacené v dohodnutém termínu, kterých byl dostatek. U půjček na pět let počet problémových úvěrů výrazně převyšoval nad počtem OK úvěrů. Byly tedy vybrány i předčasně splacené a běžící úvěry, v obou případech po 70 % uhrazených splátek. Důvod, proč omezujeme počet úvěrů nedokončených je jasný, ale omezení předčasně splacených úvěrů má rovněž svůj význam. Věřitelé nemají zájem, aby se úvěr splatil dříve, než v dohodnutém termínu, protože mohou očekávat nižší úrok, a tedy nižší zisk. Obrázek níže znázorňuje počtu úvěrů ve třídách po provedení těchto úprav.



Obr. 16 Rozložení úvěrů dle stavu a délky splácení v měsících

Vyvažování záznamů lze provést automaticky po vygenerování uzlu Balance z distribučního grafu, a to pomocí boostingu nebo redukce. V případě boostingu bude doplněn počet záznamů v ostatních třídách podle třídy s největším počtem případů. U redukce se naopak počty sníží, a to podle třídy s nejméně záznamy. Velmi vysoký počet OK úvěrů o délce 36 měsíců bude snížen na úroveň NOK 36. Počet problémových a bezproblémových úvěrů na 60 měsíců můžeme také zredukovat dle méně početné třídy.

Pokud by se počet záznamů vyvažoval pouze podle stavu úvěru OK či NOK bez ohledu na dobu splácení, mohlo by se stát, že model bude klasifikovat převážně podle úvěrů kratších. Dále je třeba zvážit, zda úvěry s různou délkou vůbec slučovat dohromady, protože se jedná o odlišný produkt. Dá se přepokládat, že model postavený pouze na datech týkajícího se jednoho produktu, bude přesnější než ten, který zahrnuje oba typy. Aby se však počet záznamů ještě více nesnížil, sloučíme oba typy úvěrů a pokusíme se udělat jakýsi kompromis v podobě nalezení hranice mezi problémovým a bezproblémovým úvěrem bez ohledu na dobu splácení s očekáváním nižší přesnosti. Při přípravě dat pro rating úvěrů se postupovalo obdobným způsobem, kdy počet úvěrů v jednotlivých třídách byl nastaven stejně.

5 Modelování

5.1 Klasifikace úvěrů

U *bankovních* dat bylo nejprve potřeba vyřadit atributy týkající se průběhu splácení úvěru, které vznikly z transakčních údajů. Jedná se o veškeré atributy obsahující ve svém názvu „úvěr“ nebo „sankce“. Týkají se zmíněných operací, které nebyly zahrnuty do výpočtu průměrného zůstatku. U *nebankovních* dat bylo potřeba vyloučit především prohřešky a počty měsíců od posledních nepříznivých událostí. Zde však bylo potřeba dát pozor, zda atribut neobsahuje v názvu slovo „veřejný“. U takových atributů lze předpokládat, že mohou být dostupné v okamžiku žádosti o úvěr z nějaké dostupné databáze. Proto je dobré vzít je v úvahu, i když samozřejmě nemusí hrát významnou roli. Může jít o historii splácení jiných úvěrů nebo o tzv. revolvingový úvěr, který se týká kreditních karet. Vzhledem k tomu, že v případě těchto nebankovních dat, které pochází z amerického trhu, byl význam některých atributů odhadován až v průběhu modelování, což jistě plní svůj účel. Před modelováním byla provedena selekce dimenzí pomocí Feature selection, který vyřadil ty nejméně důležité z hlediska klasifikace úvěrů. V uzlu Partition byla nastavena hodnota rozdělovacího semínka, při kterém byl podíl záznamů v jednotlivých třídách v trénovací i testovací množině přibližně stejný.

Tab. 9 Výchozí nastavení streamu pro klasifikaci úvěrů

Data	Počet záznamů	Počet atributů	Dělící semínko	Trénovací m.	Testovací m.
Bankovní	1116	135	5330371	50%	50%
Nebankovní	32 868	20	295244	50%	50%

Výše uvedená tabulka shrnuje výchozí nastavení společné pro veškeré algoritmy použité pro stavbu modelů v rámci úlohy. Dále bylo potřeba nastavit prostřednictvím uzlu Type odpovídající typy jednotlivých proměnných včetně určení cíle. Algoritmy následně převezmou vlastnosti streamu a v kombinaci s vlastními zadanými parametry vybudují klasifikační modely. Pomocí uzlu Analyze je vytvořený model testován nad množinami trénovacích a testovacích dat za účelem zjištění klasifikační přesnosti. V závislosti na použití manuálního testování či skriptu byly výsledné přesnosti klasifikace uloženy do souborů, u kterých jsou v názvu uvedeny parametry modelu. Pro vysvětlení významu některých parametrů byla použita online příručka IBM Knowledge Centre. (IBM CORPORATION, 2012)

5.1.1 Rozhodovací stromy

V případě rozhodovacích stromů byl pro klasifikaci úvěrů zvolen algoritmus C5.0 v režimu stromu. Pomocí skriptu bylo vytvořeno několik modelů (cca 1000), které se od sebe lišily různým nastavením parametrů. Testovanými parametry byly:

- míra lokálního prořezávání,

- minimální počet záznamů ve větvích.

U všech modelů byl zapnut *boosting* za účelem zvýšení přesnosti s počtem opakování deset. Dále bylo u všech modelů povoleno globální prořezávání, které také může přesnost zvýšit. Tato volba bývá ve výchozím stavu vždy zapnutá. Redukce počtu atributů tzv. winnowing nebyla využita z důvodu předchozí selekce. U rozhodovacích stromů je možné použít i metodu křížové validace místo dělení záznamů na trénovací a testovací. Křížová validace je doporučená, pokud je vzorek dat malý. V tomto případě by tedy mohla být vhodná, nicméně chceme použít podobné nastavení parametrů pro obě datové množiny. Druhá datová sada s nebankovními půjčkami má mnohem více záznamů, a proto by nemělo použití křížové validace význam.

V případě bankovních dat se dále nabízí možnost využití metody hlavních komponent pro redukci počtu atributů. U rozhodovacího stromu se budeme pokoušet i o interpretaci informací, kterou by aktivní zapojení skrytých faktorů mohlo zkomplikovat, nehledě na případnou nižší přesnost. Dalším důvodem je paralelní porovnávání klasifikačních modelů nad dvěma sadami dat, kde by v případě nebankovních úvěrů redukce dosavadního počtu dvacet atributů neměla význam. V obou případech chceme pracovat s původními proměnnými.

Bankovní úvěry

Nejpřesnější modely nad testovací sadou dosáhly výsledku 99,06 %, kterého bylo dosaženo při:

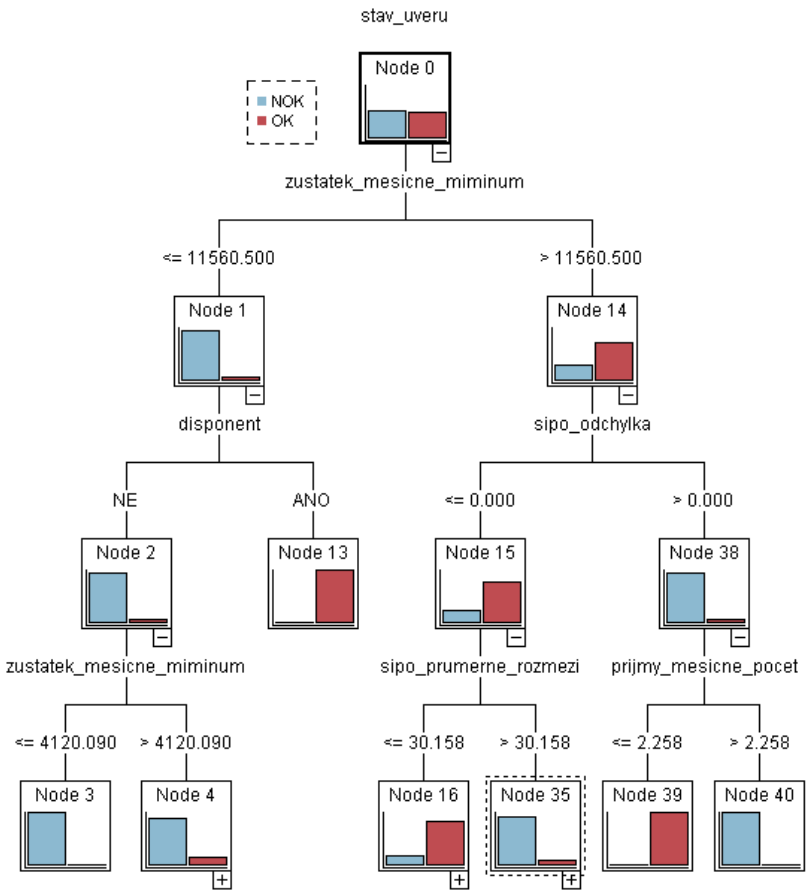
- míře lokálního prořezávání 50 až 65
- minimálním počtu záznamů ve větvích 1 až 2.

Testované míry lokálního prořezávání se pohybovaly mezi 50 až 100 a minimální počty záznamů na větev mezi 1 až 20. Pro další analýzu volíme model s nastavením 65 a 2 podle pořadí uvedených parametrů. Výsledná přesnost je vysoká, i přes takto malý počet záznamů. U neboostovaného stromu bylo dosaženo o něco nižší přesnosti 95,09 % s jinak stejným nastavením parametrů. Pro interpretaci nejvýznamnějších proměnných se jeví jednodušší použít neposílený strom, ve kterém jsou patrnější rozdíly mezi důležitostmi jednotlivých proměnných. Tento model rozhodovacího stromu označil za nejvýznamnější z hlediska klasifikace úvěrů tyto atributy:

- *minimální měsíční zůstatek*
- *disponent k účtu – ANO/NE*
- *SIPO odchylka* – vyjadřuje kolísání průměrných výdajů na SIPO.

Celkový počet hladin u boostovaného stromu je šestnáct a u základního dvanáct. Oba modely jsou ve vyobrazených hladinách stejné, kde větvení a počty záznamů v uzlech se nijak neliší. Část vytvořeného stromu (Obr. 17) říká, že pokud *minimální měsíční zůstatek* byl vyšší, než uvedená hranice, byla pak většina úvěrů bezproblémových. Následně s nulovou *odchylkou od průměrného SIPA* byla rovněž většina úvěrů

bezproblémových. Pokud byl zůstatek nižší nebo roven stanovené hranici, pak rozhodoval atribut o počtu uživatelů účtu. V případě existence *disponenta*, tedy druhého uživatele, zůstaly úvěry bezproblémové.



Obr. 17 Část stromu pro klasifikaci bankovních úvěrů

Nebankovní úvěry

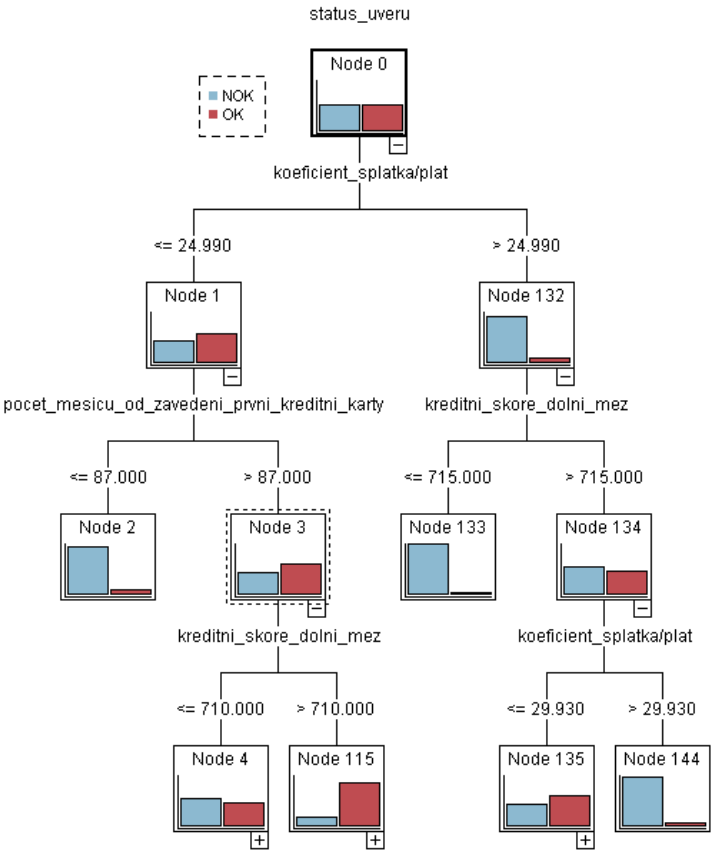
Nejpřesnější model dosáhl výsledku 80,54 %. Jeho nastavení parametrů bylo následující:

- míra lokálního prořezávání 83
- minimální počet záznamů ve větvích 18

Rozdíl přesností cca 20 % mezi bankovními a nebankovními úvěry je velký. Může to být dáno rozdílným počtem vstupních proměnných nebo i odlišností sledovaných charakteristik. U nebankovních dat označil neboostovaný model rozhodovacího stromu tyto atributy za nejvýznamnější:

- *kreditní skóre* – hodnotí bonitu klienta. Čím vyšší skóre, tím vyšší bonita.

- *koeficient splátka/plat* – vyjadřuje poměr mezi měsíční splátkou ostatních dluhů a průměrnou měsíční mzdou. Čím nižší hodnota, tím lepší.
- *měsíční splátka úvěru*



Obr. 18 Část stromu pro klasifikaci nebankovních úvěrů

Dle Obr. 18 máme možnost zachytit, že klienti s *koeficientem splátka/plat* vyšším jak 25 % měli v drtivé většině problém se splácením. Další klienti v pravém podstromu měli rovněž problém se splácením, pokud *skóre* bylo nižší než 715.

Nyní je potřeba zmínit, že v datech nebankovních půjček jsou v tuto chvíli zahrnuty obě délky úvěrů, tj. 3 a 5 let. Pokud bychom postavili dva modely se zmíněným nastavením parametrů pro oba typy zvlášť, dostali bychom se na následující přesnosti:

- 81,31 % pro model s 5letými úvěry o cca 12 000 záznamech
- 78,4 % pro model s 3letými úvěry o cca 20 000 záznamech

Zajímavé je, že i přes vyšší počet záznamů o kratších úvěrech je vypočtená přesnost nižší než u delších. Dále jsou kratší úvěry tvořeny záznamy o zcela ukončených úvěrech, na rozdíl od 5letých, čili predikce by měly být u kratších úvěrů i proto spolehlivější. Na základě těchto zjištění se lze domnívat, že informace o delších úvěrech jsou stabilnější a lze od nich očekávat přesnější modely. Pozitivním zjištěním je, že

u obou těchto modelů byly označeny za nejvýznamnější stejné atributy, jako u modelu obsahujícího oba typy úvěrů.

Tab. 10 Porovnání klasifikátorů rozhodovacího stromu pro oba typy úvěrů

Typ úvěrů	Parametry[7]	Nejvyšší přesnost
Bankovní úvěry	65 a 2	99,06 %
Nebankovní úvěry	83 a 18	80,54 %

5.1.2 Neuronové sítě

Pro stavbu modelu pomocí neuronových sítí byl zvolen typ vícevrstevného perceptronu (MLP). Kromě MLP je dostupný typ RBF, který ovšem nebude testován vzhledem k očekávané nižší přesnosti. Nevýhodou MLP je vyšší potřebná doba pro trénování a skórování záznamů. Vzhledem k vlastnostem MLP bylo zapotřebí nastavovat různé počty perceptronů v obou skrytých vrstvách, za účelem dosažení co nejvyšší přesnosti. Ta byla zjištěna pomocí skriptu obsahujícího for cykly, které testovaly přesnost nejdříve s různým počtem neuronů u první skryté vrstvy. Nastavení první vrstvy s nejlepším výsledkem bylo poté použito ve druhém cyklu, který měnil počet neuronů ve druhé vrstvě a zapisoval výsledky do souboru. V SPSS je možné určit vhodný počet perceptronů i automatickou volbou. Této možnosti však nebylo využito, protože tímto způsobem lze sestavit model pouze s jednou skrytou vrstvou. Počet neuronů v jednotlivých vrstvách by neměl přesáhnout počet vstupních prediktorů. (IBM CORPORATION, 2012)

Ve všech případech byl zapnut *boosting* za účelem dalšího zvýšení přesnosti. Tato metoda vytvořila určitý počet modelů, z nichž se sestavil výsledný s nejvyšší přesností. Počet modelů komponent byl 10, což byla výchozí hodnota. Z omezujících kritérií byly zvoleny pouze maximální doba trénování pro každou komponentu. Tento parametr byl nastaven na 15 minut, což byla také výchozí hodnota. Trénovací čas se však vždy mírně lišil, protože se čekalo na dokončení probíhajícího cyklu. Dalším možný parametrem pro nastavování je stanovení podílu množiny záznamů určených pro předcházení tzv. přetrénování. Výchozí hodnota byla 30, která byla zvolena.

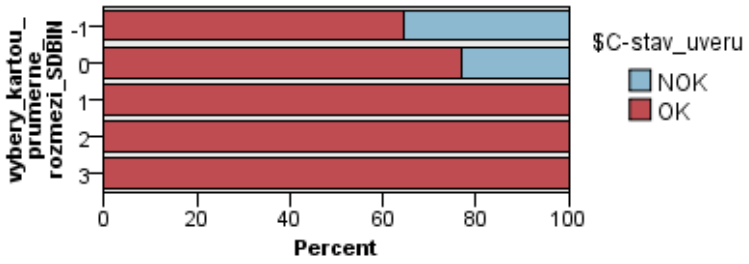
Bankovní úvěry

U neuronové sítě s *jednou skrytou vrstvou* byla nejvyšší naměřená přesnost nad trénovací množinou 95,85 %. Počet neuronů v této vrstvě byl 14. Po přidání *druhé skryté vrstvy* stoupla přesnost na 97,17 % při počtu 7 neuronů v této vrstvě. S tímto počtem neuronů v obou skrytých vrstvách dosáhl *neboostovaný* model přesnosti 92,08 %. U základního modelu bylo možné zjistit typy použitých aktivačních funkcí v jednotlivých vrstvách. V případě obou skrytých vrstev byla použita funkce *hyperbolický tangens*. Jedná se o nelineární typ funkce, která má svým tvarem blízko

[7] Míra lokálního prořezávání a minimální počet potomků ve větvi

k funkci *sigmoidální*. Liší se však výstupními hodnotami, které se pohybují v intervalu (-1,1), zatímco u sigmoidální funkce v intervalu (0,1). U neuronů ve výstupní vrstvě byla použitá prostá *identita*, tedy nejjednodušší typ lineární funkce.

U *posíleného* modelu byly oproti základnímu výraznější rozdíly mezi významnostmi atributů. Za nejdůležitější model označil skupiny atributů týkajících *připsaných bonusových úroků* a *průměrného zůstatku*. Základní model se lišil od posíleného označením atributu *průměrné rozmezí výběrů kartou ve dnech* za nejdůležitější. Ve fázi prostoupení do problematiky úvěrů bylo zmíněno, že atributy týkající se používání karet mohou mít vliv na rozhodnutí o poskytnutí či zamítnutí úvěru, což výsledek domněnku potvrzuje. Níže je uvedena zmíněná proměnná v kontextu rozložení predikovaných výsledků dosažených pomocí nejpřesnějšího modelu[8]. Z obrázku je patrné, že zamítnuté úvěry pocházely pouze od klientů, kteří vybírali méně často až průměrně. Graf se týká pouze klientů, kteří kartu měli, což je v případě těchto dat menšina. To může být důvod, proč tyto atributy sehrály úlohu v případě jen některých modelů.



Obr. 19 Rozložení klasifikovaných úvěrů dle průměrného časového rozmezí mezi výběry kartou

Nebankovní úvěry

U nebankovních půjček se podařilo v případě použití *jedné skryté vrstvy* dosáhnout přesnosti nad testovací množinou 78,83 %. Počet neuronů v této vrstvě byl 4. Pokud se přidala *druhá skrytá vrstva*, přesnost stoupla na 79,74 % při počtu 3 neuronů v této vrstvě. Vzhledem k výrazně nižšímu počtu vstupních prediktorů oproti bankovním půjčkám byla podle očekávání nalezena nejvyšší přesnost u menšího počtu neuronů než u bankovních půjček.

Tab. 11 Porovnání klasifikátorů neuronových sítí pro oba typy úvěrů

Typ úvěrů	Počty neuronů[9]	Nejvyšší přesnost
Bankovní úvěry	14 a 7	97,17 %
Nebankovní úvěry	4 a 3	79,74 %

[8] Jedná se o model rozhodovacího stromu s přesností 99,06 %.

[9] V první a druhé skryté vrstvě.

Neuronová síť označila za nejvýznamnější stejné atributy, jako model rozhodovacího stromu. *Základní neboostovaný model* se stejným počtem neuronů ve skrytých vrstvách dosáhl přesnosti 76,48 %. Neurony ve skryté vrstvě rovněž obsahovaly aktivační funkci *hyperbolický tangens* a ve výstupní vrstvě *identitu*.

5.1.3 Support vector machines

V SPSS Modeleru se v případě SVM rozlišují 4 jádrové funkce, kdy každá se hodí na jiný typ dat. Pokud máme lineárně neseparabilní data, pak funkce nelineární záznamy obvykle lépe transformují do jednodušeji oddělitelné podoby. Typy těchto jádrových funkcí jsou:

- lineární
- RBF
- sigmoidální
- polynomiální

Vzhledem k tomu, že nevíme, jaké vlastnosti data mají, vyzkoušíme všechny typy funkcí s různým nastavením parametrů a výsledky porovnáme. Parametry, které budeme nastavovat a sledovat, jsou (IBM CORPORATION, 2012):

- *Stop kritérium* – nižší hodnota znamená vyšší přesnost, ale delší čas pro natrénování. Může vést k přeučení.
- *C Regularizační parametr* – hodnoty by měly být v rozsahu 1 až 10. Smyslem regularizačního parametru je řídit kompromis mezi maximalizací dělicí nadroviny a minimalizací chyby
- *γ gamma / RBF gamma* - vyšší hodnota zvyšuje přesnost, také může vést k přeučení. V případě RBF gamma by měla hodnota parametru běžně spadat do intervalu 3/k až 6/k, kde k představuje počet vstupních proměnných.
- *D stupeň složitosti polynomu* – pouze polynomiální fce.

Některé parametry se týkaly použití spojité cílové proměnné místo kategoriální, a proto nebyly relevantní. Atribut *bias* byl ve většině případů ponechán dle doporučení s výchozím nastavením 0. Následně byl vytvořen skript za účelem zjištění nejvyšších přesností daných typů funkcí s různým nastavením parametrů u obou datových sad. Volba určení významnosti jednotlivých proměnných byla vypnuta, protože tato možnost značně prodlužovala dobu generování modelů.

Bankovní úvěry

Dle výsledků v tabulce níže byl nejpřesnější model SVM s jádrovou funkcí *RBF*. U těchto dat lze použít i lineární funkci, pravděpodobně proto, že transformovat data lze pomocí lineárně oddělitelné hranice. Regularizační parametr měl nejvýraznější vliv na změnu přesnosti klasifikace.

Tab. 12 Porovnání klasifikátorů SVM bankovních úvěrů

Typ funkce	Stop kritérium	Regularizační parametr C	(RBF) Gamma γ	Stupeň složitosti polynomu D	Přesnost
RBF	1.0E-1	7	0,03	-	97,74 %
Polynomiální	1.0E-1	5	0,5	4 - 5	96,79 %
Lineární	1.0E-1	3	-	-	96,42 %
Sigmoidální	1.0E-1	1	0,1	-	79,25 %

Metoda testování parametrů vypadala následovně. Například u *polynomiální funkce* se pomocí skriptu zjistilo, jaké nastavení regularizačního parametru zajistí nejvyšší přesnost. Poté se podle vybrané hodnoty C ladily výsledky ostatních parametrů, tj. nastavení γ a D. Nejjednodušší bylo ladění modelu s *lineární funkcí*, u kterého stačilo sledovat nárůst přesnosti, v důsledku změny hodnoty parametru C. U *sigmoidální* funkce bylo jako u jediné zapotřebí testovat parametr *bias*, kdy nejlepšího výsledku bylo dosaženo při hodnotě chyby 1. V ostatních případech stačilo nastavení výchozích hodnot. Tento typ funkce nebude uvažován při závěrečném zhodnocení.

Nebankovní úvěry

Tab. 13 Porovnání klasifikátorů SVM nebankovních úvěrů

Typ funkce	Stop kritérium	Regularizační parametr C	(RBF) Gamma γ	Stupeň složitosti polynomu D	Přesnost
RBF	1.0E-1	7	0,15	-	78,39 %
Polynomiální	1.0E-1	1	0,5	2	78,02 %
Lineární	1.0E-1	10	-	-	77,39 %
Sigmoidální	1.0E-1	1	2,5	-	76,03 %

I v případě modelů SVM s různými typy jádrových funkcí se přesnosti významně lišily od bankovních úvěrů. U sigmoidální funkce byl zjištěn nejlepší výsledek při nastavení parametru *bias* v hodnotě 2.

5.1.4 Shrnutí klasifikace bankovních a nebankovních úvěrů

Na základě porovnání modelů postavených nad oběma množinami lze říci, že úvěry skutečně mohou být klasifikovány s velkou přesností. Záleží však na počtu a typu sledovaných atributů. Lze předpokládat, že velkou úlohu v případě bankovních úvěrů sehrály atributy pocházející z transakčních údajů, jako je například průměrný zůstatek na účtu. Atribut kreditní skóre klienta u nebankovních úvěrů v kombinaci s ostatními atributy ani zdaleka nezajistil tak vysokou přesnost modelů, a to ani vzhledem k mnohonásobně většímu počtu dostupných záznamů.

Pořadí přesností klasifikačních modelů je v případě obou datových množin stejné. Nejvyšší přesnosti dosáhly modely algoritmu C5.0, druhé byly neuronové sítě

a jako třetí SVM. Vzhledem k podobných přesnostem jednotlivých modelů v rámci jedné sady dat nelze říci, že by některý ze tří typů modelů byl výrazně přesnější. Důležité je především nastavení parametrů u jednotlivých algoritmů. V případě obou typů úvěrů byl rozdíl mezi nejlepším a nejhorší modelem cca tři procenta.

5.2 Navazující úloha ke klasifikaci úvěrů

V této úloze půjde o shlukování na základě určité vlastnosti, kterou je počet OK/NOK úvěrů mezi jednotlivými shluky. Pro účely této úlohy budou použity obě datové množiny, které jsou připraveny stejným způsobem, jako u klasifikace. Pro vytvoření shluků bude použito dvou metod, které na sebe budou navazovat. V první fázi je užito *faktorové analýzy*, která podle korelací mezi atributy vytvoří skupiny skrytých atributů. S použitím rotace vznikne několik ortogonální faktorů, které vystihují rozptyl v datovém souboru. Vybrané faktory poté budou použity pro shlukování a vizualizaci v grafu, kde ve zvolených dimenzích jednotlivé shluky vyniknou. Shluky budou v podstatě vytvořeny ještě před použitím clusterovacího algoritmu. Algoritmus *k – means* zde bude hrát roli v případě zatřídění záznamů do jednotlivých skupin, které bude možné popsat pomocí agregační tabulky. Z této tabulky budou následně odvozeny sledované charakteristiky. Hlavní zkoumaná vlastnost bude sloužit k vyznačení a popisu shluků záznamů v bodovém diagramu. Pro obarvení záznamů tedy bude použita třída vzniklá z algoritmu k – means, jejíž popis bude nahrazen vypočteným podílem problémových úvěrů v daném shluku. Díky těmto metodám bude možné osvětlit význam některých atributů, ke kterému nebyl věnován prostor v předchozích kapitolách. Za účelem potvrzení závislosti mezi určitými proměnnými v rámci jednoho faktoru může být použit klasifikační algoritmus, například C5.0. Na druhou stranu použití samotné faktorové analýzy (dále FA) je z pohledu zjištění závislostí mezi atributy vhodnější, protože z ní můžeme odvodit, zda mezi atributy panuje přímá či nepřímá korelace.

5.2.1 Nebankovní úvěry

Nejdříve byl proveden experiment u nebankovních půjček, které obsahují menší počet atributů, na němž půjde problém FA přehledněji demonstrovat[10]. Před jejím provedením bylo zapotřebí najít a odstranit případné anomálie pomocí Anomally detection, což by mělo zamezit tvorbě příliš malých shluků obsahující právě tyto odlehlé hodnoty. Do procesu FA byly zahrnuty stejné atributy jako při klasifikaci úvěrů, kromě cílové proměnné, která byla vypuštěna. FA patří mezi metody učení bez učitele, resp. redukci atributů bez učitele. Nominální proměnné bylo potřeba převést na spojité pomocí uzlu Reclassify, aby mohly být rovněž zapojeny do FA. Pro vytvoření modelu byly sledovány tyto nejdůležitější parametry:

- *korelační matice* – používá se pro faktorovou analýzu

[10] Do analýzy byly zahrnuty úvěry s délkou 5 let.

- *parametr extrakce faktorů* – eigen value > 1. To znamená, že budou vysvětleny atributy, jejichž eigen value je větší než jedna. Nebudou tedy extrahovány faktory, které vysvětlují méně než jednu proměnnou.
- *metoda rotace* – Varimax. Chceme rovnoměrnou zátěž na jednotlivé faktory. Například první faktor tak nebude přetížen několika atributy s vysokou zátěží. Faktor půjde lépe interpretovat v kontextu vstupních atributů.

Jedním z výstupů vytvořeného modelu je vysvětlovaný rozptyl (viz Obr. 20), kde extrakcí původních 19 dimenzí se počet zredukoval na 9. Tyto skryté faktory vysvětlují 77 % původního rozptylu datového souboru.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,214	16,917	16,917	3,214	16,917	16,917	2,576	13,559	13,559
2	2,657	13,986	30,903	2,657	13,986	30,903	2,445	12,868	26,427
3	1,826	9,613	40,516	1,826	9,613	40,516	1,960	10,316	36,742
4	1,543	8,120	48,636	1,543	8,120	48,636	1,897	9,986	46,728
5	1,201	6,322	54,958	1,201	6,322	54,958	1,355	7,134	53,862
6	1,154	6,075	61,033	1,154	6,075	61,033	1,159	6,099	59,961
7	1,055	5,550	66,583	1,055	5,550	66,583	1,149	6,046	66,007
8	1,030	5,420	72,003	1,030	5,420	72,003	1,126	5,928	71,935
9	1,012	5,327	77,330	1,012	5,327	77,330	1,025	5,396	77,330
10	,891	4,690	82,020						

Obr. 20 Celkový vysvětlovaný rozptyl pomocí vytvořených faktorů

Z devatenácti inicializačních faktorů byly vybrány ty, jejichž eigen value bylo vyšší než 1. Suma eigen values těchto faktorů odpovídá počtu původních dimenzí. Jednotlivé faktory vysvětlují zatížení rozptylem po extrakci pomocí PCA. Následně po provedení rotace došlo ke změně zatížení rozptylem na jednotlivých faktorech. Z výstupu je patrné, že po rotaci se značná část rozptylu z prvního faktoru přenesla na další. Nadále však zůstává platným pravidlem, že množství rozptylu na jednotlivých faktorech se postupně snižuje směrem od prvního faktoru.

Výsledkem je tedy rotovaná matice podle metody Varimax (viz Obr. 21), kde je upřednostněno rozložení zatížení tak, aby v jednotlivých komponentách (či faktorech) bylo co nejméně proměnných s velkým vlivem. Z matice jsou patrné skupiny původních atributů, které jsou nyní popsány skrytým faktorem. Z devíti faktorů byly zahrnuty do shlukování pomocí k - means první tři, jejichž počet byl dostatečný na vytvoření definovaného počtu odlišných shluků. Jednotlivé clustery neměly příliš rozdílnou velikost. Poměr mezi nejmenším a největších shlukem byl cca dvojnásobný. Počet shluků je šest. V SPSS se měří kvalita shlukování podle míry koheze a separace. V podstatě se jedná o to, že případy musí být co nejblíže centroidu svého clusteru a co nejdále od centroidu nejbližšího clusteru. Kvalita shlukování byla v SPSS hodnocena jako „Fair“. Následně byly s pomocí agregační tabulky a dalších

operací zjištěny podíly problémových úvěrů v jednotlivých shlucích, které byly použity jako popis skupin případů představující shluky.

Rotated Component Matrix(a)

	Component								
	1	2	3	4	5	6	7	8	9
kreditni_skore_dolni_mez	-,895								
kreditni_skore_horni_mez	-,895								
mira_vyuziti_revolvingovych_uveru	,781								
uver_castka		,930							
mesicni_splatka		,911							
rocni_prijem		,664						-,359	
celkovy_objem_revolvingovych_uveru		,427						-,349	
pocet_otevrenych_kreditnich_linii			,884						
celkovy_pocet_kreditnich_linii			,831						
koeficient_splatka/plat	,407		,549			,322			
pocet_mesicu_od _posledniho_verejneho_zaznamu				,952					
pocet_prihorsujicich_verejnych_zaznamu				,952					
delka_zamestnani					,729				
vlastnictvi_nemovitosti_Reclassify7					-,580				,341
pocet_mesicu_od_zavedeni_prvni_kreditni_karty					,569			-,302	
pocet_mesicu_od_posledni_delikvence						-,888			
pocet_kreditnich_prohresku_posledni_2_roky							-,898		
ucel_uveru_Reclassify7								,738	
stat_usa_Reclassify7									,924

Obr. 21 Rotovaná matice faktorových zátěží dle metody Varimax

Hodnotu prvního faktoru lze u jednotlivých případů vypočítat takto:

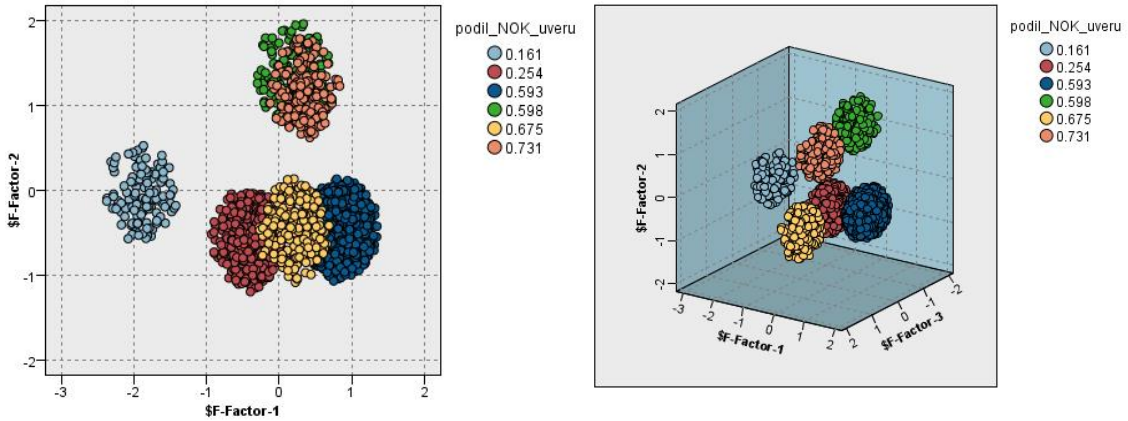
$$F_1 = -0{,}895 \times X_1 - \ 0{,}895 \times X_2 + 0{,}781 \times X_3 + 0{,}407 \times X_4$$

X_n vyjadřuje hodnotu případu u daného atributu. Tyto hodnoty původních proměnných jsou vynásobeny svými zátěžemi uvnitř faktoru a sečteny. Výsledek je vyjádřen v normalizovaném měřítku.

Na Obr. 22 jsou po vynesení dvou faktorů vidět jednotlivé shluky včetně přibližné pozice jejich centroidů. Při vynášení případů do grafu byla nastavena omezující podmínka pro zobrazení záznamů se vzdáleností alespoň 0.08 od svého k – středu. Světle modrý shluk je díky prvnímu faktoru nejvíce oddělen od ostatních. Jedná se o shluk s nejmenším počtem problémových úvěrů. Podívejme se nyní na vztah proměnných ve *faktoru 1*. Největší vliv na hodnotu tohoto faktoru má *kreditní skóre klient*a, které je v silné nepřímé korelaci. Se zvyšujícím se skóre klienta obvykle klesá míra rizika. Pojmenujme tedy *první faktor* jako *míru rizika*. Dále pak proměnná

míra využití revolvingových úvěrů je také v korelaci, ale ve slabší přímé. Jedná se vyjádření míry využívání disponibilních revolvingových úvěrů. Míra závisí na maximu, které si klient může půjčit dle rizikového ohodnocení, a aktuální výši čerpání. Tyto úvěry jsou spojeny s čerpáním prostřednictvím kreditních karet.

Pokud srovnáme světle modrý a oranžový shluk, zjistíme, že oranžový je podle podílu problémových úvěrů na tom nejhůře ze všech. V případě *druhého faktoru* je možné postupovat stejně. Faktor 2 lze pojmenujme jako *velikost úvěru*. Pokud opět srovnáme podle našeho kritéria dva nejvzdálenější shluky, pak vidíme, že velikost úvěru opět hraje roli. Na druhou stranu pokud do srovnání zahrneme všechny shluky, pak tento faktor nám říká, že sklon ke splácení úvěru přímo nezávisí pouze na velikosti půjčky. Zní to logicky, protože především rizikovost klienta udává, při jaké hranici velikosti úvěru bude více pravděpodobné, že úvěr nesplatí. Každý klient má tuto pomyslnou hranici nastavenou jinak.



Obr. 22 Vizualizace shluků ve dvou a tří rozměrném prostoru

Pokud vyneseme do grafu ještě třetí faktor, uvidíme vizuální oddělení dvou nejhorších shluků, tedy žlutého a oranžového. Tento faktor můžeme nazvat jako *počet kreditních karet*. Odtud lze odvodit, že vyšší počet kreditních linií zvyšuje pravděpodobnost nesplacení úvěru. Vyšší počet kreditních karet opět zvyšuje zatížení klienta celkovým objemem úvěrů včetně revolvingových, což nahrává trendu splácení půjček úvěrem.

Dle Tab. 14 je vidět, že shluk s nejmenším počtem problémových úvěrů byl charakteristický nejnižší mírou rizika, průměrně vysokým požadovaným úvěrem a průměrným počtem kreditních karet. Pro srovnání jsou u každého faktoru uvedeny související kategorizované proměnné vzniklé ve fázi přípravy dat.

Tab. 14 Charakteristiky shluků nebankovních úvěrů

Podíl NOK úvěrů	Faktor 1 Riziko	Kreditní skóre dolní mez SDBIN	Faktor 2 Požadovaný úvěr	Úvěr částka SDBIN	Faktor 3 Počet kreditních karet	Počet otevřených kreditních linií SDBIN
0,161	-1,888	1,578	0,002	0,299	0,000	-0,068
0,254	-0,426	0,289	-0,576	-0,109	-0,484	-0,289
0,593	0,858	-0,296	-0,481	-0,049	-0,424	-0,222
0,598	0,193	0,014	1,395	1,114	-0,810	-0,249
0,675	0,191	-0,089	-0,470	0,028	1,387	0,642
0,731	0,330	-0,010	1,196	1,126	0,786	0,368

5.2.2 Bankovní úvěry

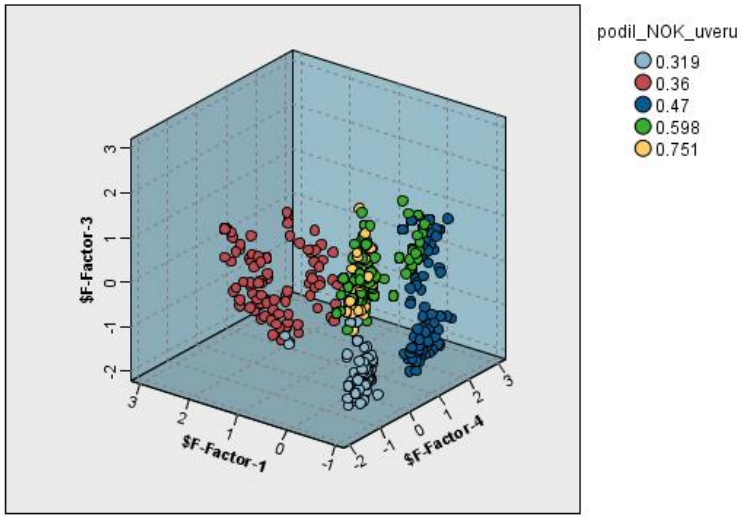
Obdobným způsobem bylo přistupováno k bankovním úvěrům. Vytvořené faktory jsou umístěny v příloze diplomové práce ve formě vygenerovaného HTML souboru. Výsledek shlukování je vidět na obrázku níže. Účastnilo se jej prvních osm faktorů. Tento počet byl uvážen na základně posouzení metrik pro hodnocení shlukování v SPSS podobným způsobem, jako u předchozích dat.

Do vizualizace byly zahrnuty faktory 1, 3 a 4. Podle rotované matice jsou v prvním faktoru zastoupeny atributy zahrnující příjmy, tvořené převody ve prospěch a vklady. Převody ve prospěch jsou v silné přímé korelaci s faktorem, zatímco vklady převážně v méně silné korelaci nepřímé. To znamená, že pokud přichází na daný účet prostředky v elektronické formě, pak množství prostředků vložených fyzicky u přepážky obvykle klesá. Pokud příjmy přichází formou převodu na účet, pak se může jednat o pravidelné měsíční výplaty ze zaměstnání. V opačném případě může jít o nepravidelné vklady, které nemusí představovat trvalý příjem. Pojmenujme *první faktor* například jako *pravidelné příjmy*, kde je pro jeho vysvětlení zvolena proměnná shrnující minimální měsíční příjmy. *Třetí faktor* je popsán pomocí odchylky od měsíčních výdajů a vyjadřuje *čerpání prostředků*. *Čtvrtý* faktor zachycuje *aktivitu na účtu*. Popis jednotlivých shluků podle faktorů v kombinaci s vybranými proměnnými je k dispozici v Tab. 15. Tabulka obsahuje agregované hodnoty představující průměrné hodnoty u daných charakteristik nad všemi případy.

Tab. 15 Charakteristiky shluků bankovních úvěrů

Podíl NOK úvěrů	Faktor 1 Pravidelný příjem	Příjmy měsíčně minimum SDBIN	Faktor 3 Čerpání prostředků	Výdaje měsíčně odchylka SDBIN	Faktor 4 Aktivita	Zůstatek podíl aktivních dnů SDBIN
0,319	-0,631	0,113	-1,294	0,000	-0,783	-0,019
0,360	1,686	1,258	-0,200	0,258	-0,061	0,266
0,470	-0,520	-0,286	-0,278	0,493	1,530	0,908
0,598	-0,518	-0,781	0,848	1,285	-0,268	0,055
0,751	-0,506	-0,683	0,506	1,102	-0,594	0,024

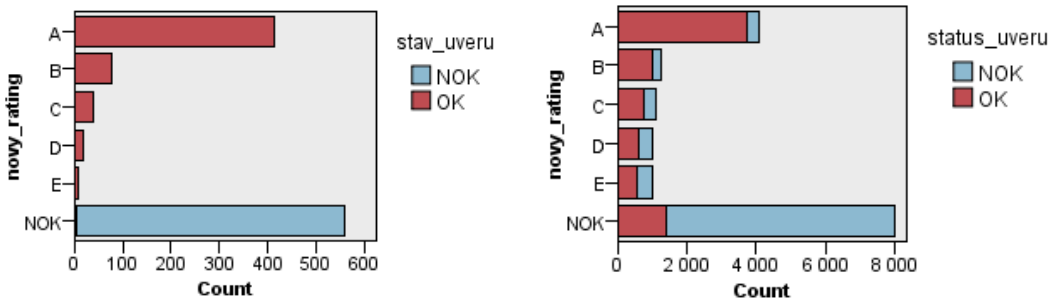
Na Obr. 23 má světlemodrý shluk nejméně problémových úvěrů. S pomocí tabulky můžeme říci, že se jedná se o klienty, kteří mají příjmy F1 průměrné až podprůměrné vzhledem ke všem klientům. Míra čerpání prostředků F3 je v průměru nejnižší. Aktivita na účtu F4, kterou nejlépe vyjadřuje změna zůstatku, je rovněž nejnižší. Takovýto shluk však nemusí být z pohledu banky příliš zajímavý, protože klienti v něm přináší málo prostředků, kterými banka může kooperovat při poskytování úvěrů. Z tohoto pohledu se zdá být výhodnější shluk červený, ve kterém klienti mají nadprůměrné příjmy. Za nejvíce rizikový můžeme označit žlutý shluk, který má stejně jako světlemodrý podprůměrné příjmy a nízkou aktivitu. Liší se však nadprůměrnou mírou čerpání prostředků.



Obr. 23 Shluky klientů s bankovními úvěry

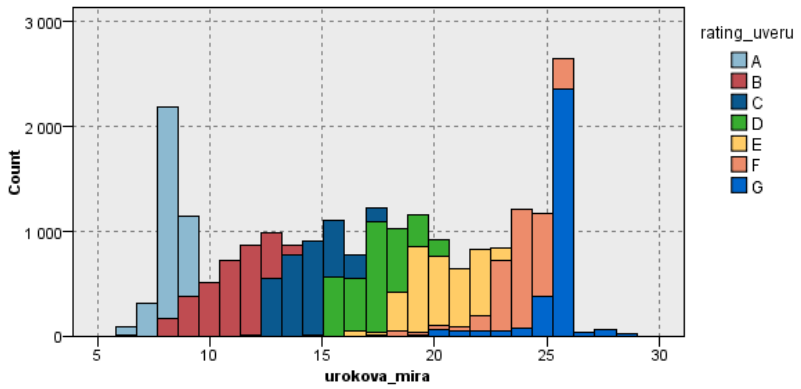
5.3 Rating úvěrů

Kromě předpovědi, zda klient vůbec bude úvěr splácet, můžeme predikovat, s jakou pravděpodobností se mu to podaří. Tento proces lze nazvat jako skórování, k čemuž může sloužit výsledek předchozího modelu pro klasifikaci. Ke skórování můžeme použít v podstatě každý klasifikační model v SPSS. Obecně je potřeba, aby cílový atribut byl binární, u kterého jedna z hodnot je označena jako „true“. Následně se u každého záznamu určuje pravděpodobnost, s jakou bude cílová hodnota pravdivá. V konkrétním případě to znamená, že pokud úvěr bude modelem označen za problémový, sklon ke spolehlivosti splácení se bude pohybovat mezi 0 až 50 procenty. V pomyslné skórovací tabulce tak vždy bude zařazen do kategorie NOK. Pokud vyjde sklon ke splácení nad 50 procent, pak predikovaný úvěr je považován bezproblémový. Hodnoty tohoto skóre je poté možné roztřídit do několika pásem, které budou představovat rating. Na Obr. 24 níže je k dispozici realizovaný rating u obou typů úvěru. Třídy A až E představují intervaly sklonu v šíři 10 %, kde např. rating A zahrnuje úvěry se sklonem 100 až 90 %.



Obr. 24 Vlastní rating bankovních a nebankovních úvěrů

U nebankovních půjček k dispozici rating máme, čehož můžeme využít pro pokus zpětného určení vzorce pro skórování úvěrů. Tento vzorec či model nám může posloužit pro predikci ratingové třídy v případě, že bychom chtěli žádat o úvěr u dané společnosti. Pokud zadáme příslušné parametry úvěru a předáme je modelu, můžeme s určitou pravděpodobností predikovat příslušnost našeho úvěru do dané třídy. Také můžeme parametry měnit tak, abychom dosáhli lepšího ratingu. Jako výchozí informace, která bude určující pro zdůvodnění určitých postupů v této úloze, bude rozložení jednotlivých ratingových tříd podle výše úrokové míry. Domněnka o původním rozložení dat byla taková, že úroková míra je u všech úvěrů v rámci ratingové třídy přesně vymezena intervalem a je tedy možné ji zařadit do pásma vyhrazeného ratingem. Tento obrázek to částečně vyvrací, nicméně v datech neexistuje více provázaná proměnná s úrokovou míru, než je právě rating.



Obr. 25 Počty úvěrů dle výše úrokové míry a ratingu

5.3.1 Klasifikační modely pro rating úvěrů

Tato úloha slouží k vytvoření klasifikačního modelu pro určování *stupně rizika úvěrů.* Podle zatřídění do určitého pásma lze stanovit přibližnou výši *úrokové míry.* Pro vytvoření modelů byly aplikovány stejné algoritmy, jako v předchozí klasifikační úloze. Byly použity nebankovní úvěry s délkou pěti let z důvodu nepřijatelně nízké přesnosti v případě zahrnutí obou délek.

Tab. 16 Výchozí nastavení streamu pro rating úvěrů

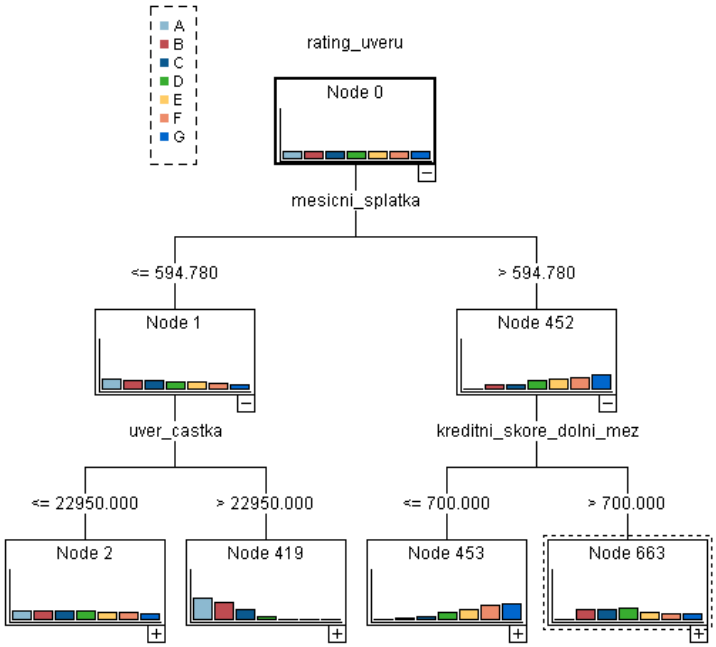
Počet záznamů	Počet atributů	Dělící semínko	Trénovací m.	Testovací m.
22 196	20	140598	50%	50%

Rozhodovací stromy

Nejvyšší přesnost u boostovaného stromu byla naměřena ve výši 77,96 %. Míra prořezávání byla nastavena na 68 a minimální počet potomků ve větvi na 5. Základní neposílený model označil za nejvýznamnější atribut výši *měsíční splátky.* Další atributy výsledky zpřesňují a patří mezi ně např.:

- *Kreditní skóre*
- *Počet měsíců od poslední delikvence* – údaj se týká splácení ostatních úvěrů včetně kreditních

Přesnost základního modelu byla 73,65 %.



Obr. 26 Část rozhodovacího stromu pro rating nebankovních úvěrů

Dle Obr. 26 vidíme, že pokud výše *měsíční splátky* překročila hranici cca 600 USD, pak úvěr častěji spadal do více rizikových kategorií. Následně *kreditní skóre* s hodnotou nižší než 700 vedlo více opět k problémovým ratingům. V levém podstromu ve třetí hladině došlo překročením *částky úvěru* nad 23 000 USD k výraznější změně v rozložení úvěrů do tříd. Od této částky byla více rozlišována rizikovost úvěrů, které byly v tomto případě zatím méně rizikové.

Neuronové sítě

Nejpřesnější model měl 17 neuronů v první skryté vrstvě a 0 ve druhé. Naměřená přesnost byla 80,68 %. Zbylé nastavení parametrů koresponduje s nastavením neuronových sítí u předchozí úlohy. Boosting měl rovněž vliv na zvýšení přesnosti s využitím ensemble modelu. Neboostovaný model se stejným nastavením dosáhl přesnosti 62,12 %. Pokud bychom použili úrokovou míru jako cílovou proměnnou pro odhad formou regrese, volba *boostingu* či *baggingu* by neměla význam, protože je automaticky v SPSS pro tyto případy ignorována. Samozřejmě i možnost regrese úrokové míry byla uvažována místo použití ratingových tříd, nicméně výsledky nebyly uspokojivé. Odhady jednotlivých sazeb se významně lišily od původních hodnot, proto bylo od tohoto řešení úlohy upuštěno.

Support vector machines

V případě použití algoritmu SVM byly vyzkoušeny všechny dostupné jádrové funkce, nicméně přijatelných výsledků bylo dosaženo pouze při použití funkcí lineárních. Uspokojivé výsledky poskytla *funkce polynomiální* se stupněm složitosti 1 a funkce *lineární*. Výsledné přesnosti byly u těchto typů následující:

- *polynomiální fce* s přesností 79,15 %, C = 20, γ = 6,5
- *lineární fce* s přesností 77,91 %, C = 20

Polynomiální funkce stupně jedna má rovněž tvar přímky. Kromě testování vlivu parametru C, byla sledována hodnota γ, která při vyšších hodnotách rovněž zvyšovala přesnost. Stop kritérium bylo v obou případech nastaveno na nejnižší hodnotu. Z výsledků klasifikátorů postavených pomocí SVM lze usoudit, že data jsou pravděpodobně lineárně separabilní a s použitím nelineární mapovací funkce nelze očekávat vyšší přesnost, spíše naopak.

5.3.2 Shrnutí výsledků ratingu úvěrů

Smyslem této úlohy byl pokus o vytvoření modelu, který bude schopen replikovat vzorce pro odhad rizikovosti úvěrů, které používá daná společnost. Rozdělit úvěry podle rizika lze tak v tomto případě takřka lineárně, což je z pohledu vytvoření spolehlivého klasifikátoru pozitivní zjištění, nicméně je potřeba dostatečné množství dat. Nejvyšší přesnosti bylo dosaženo použitím neuronových sítí. Rozhodovací stromy dosáhly o pár desetin horšího výsledku. Nejhorší model u SVM měl opět pouze o cca tři procenta horší výsledek ve srovnání s nejlepším modelem. U tohoto algoritmu se projevila výhodnější volba lineární mapovací funkce do prostoru vyššího počtu dimenzí. Použití nejspíše jakékoliv nelineární transformační funkce by tak vedlo k horším výsledků, protože by byla z pohledu těchto dat příliš složitá pro klasifikování.

5.3.3 Shlukování dle průměrné výše úrokové míry

Pro tuto úlohu byla použita podobně připravená data s rovným počtem záznamů v ratingových třídách s tím rozdílem, že byl vypuštěn atribut o kreditním skóre klienta. Analýza tak bude založena na atributech, které jsou jednodušeji zjistitelné. Výsledný počet shluků byl stanoven na 7, podle počtu ratingových tříd.

Tab. 17 Rotovaná matice faktorů pro rating úvěrů

	Faktor							
	1	2	3	4	5	6	7	8
uver_castka	0,928							
mesicni_splatka	0,918							
rocni_prijem	0,63			-0,392				
pocet_otevrenych_kreditnich_linii		0,891						
celkovy_pocet_kreditnich_linii		0,877						
pocet_mesicu_od _posledniho_verej_zaznamu			0,944					
pocet_prihorsujicich_verejnych_zaznamu			0,941					
koeficient_splatka/plat		0,375		0,793				
mira_vyuziti_revolvingovych_uveru				0,664				
delka_zamestnani					0,733			
pocet_mesicu_od_zavedeni_prvni_kreditni_karty					0,599			
vlastnictvi_nemovitosti_Reclassify					-0,479			0,435
ucel_uveru_Reclassify						0,692		
celkovy_objem_revolvingovych_uveru	0,383					-0,496		
pocet_mesicu_od_posledni_delikvence							0,822	
pocet_kreditnich_prohresku_posledni_2_roky						0,446	-0,602	
stat_usa_Reclassify								0,906

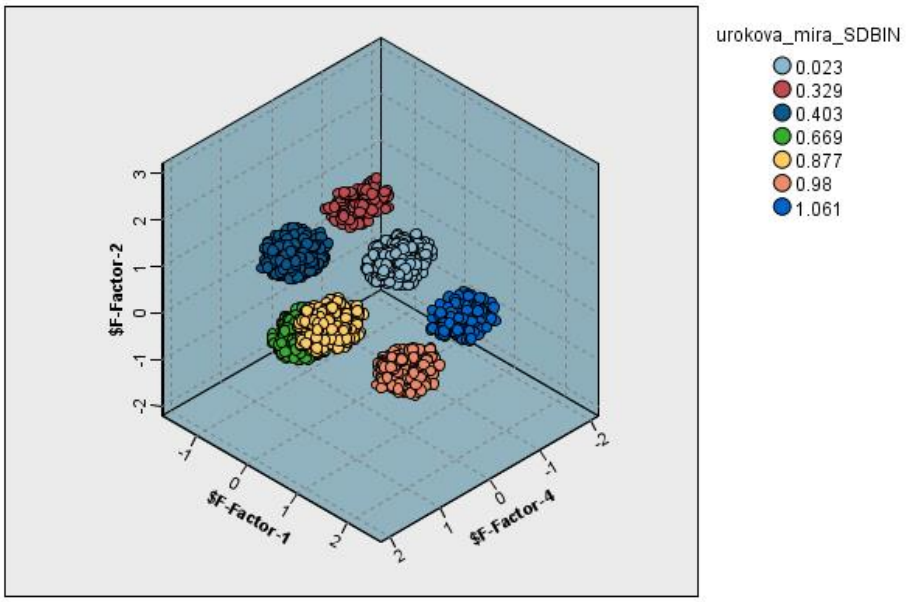
Tab. 18 Charakteristiky shluků klientů podle výše úrokové míry

Úrok. míra SDBIN	Faktor 1 Velikost úvěru	Úvěr částka SDBIN	Faktor 2 Počet kredit. karet	Počet otevř. kredit. linií SDBIN	Faktor 4 Zatížení ostat. úv.	Koeficient měsíční splátka/příjem SDBIN
0,023	-0,381	0,127	-0,198	-0,077	-1,239	-0,442
0,329	-0,098	0,407	1,935	1,529	-0,117	0,351
0,403	-0,961	-0,033	0,534	0,253	0,255	0,429
0,669	-0,640	-0,007	-0,933	-0,288	0,366	0,055
0,877	0,350	0,607	0,168	0,175	0,852	0,480
0,980	1,046	1,063	-0,939	-0,230	-0,009	-0,196
1,061	1,760	1,595	0,505	0,400	-0,368	-0,025

Faktor číslo jedna souvisí převážně s *velikostí úvěru*, ať už máme na mysli měsíční splátku, či celkovou půjčenou částku. Tento faktor obsahuje nejvíce vysvětlovaného rozptylu původních dat, a proto byl použit. *Druhý faktor* je zastoupen informacemi o počtu kreditních karet bez ohledu na to, zda se jedná o karty aktivní nebo neaktivní. *Třetí faktor* zahrnuje záznamy o prohřešcích klientů. *Čtvrtý faktor* představuje zatížení klienta ostatními úvěry. Nejlepšího výsledku shlukování bylo dosaženo po

zahrnutí faktorů 1, 2 a 4, díky nimž vzniklo 7 shluků, u kterých rozdíl mezi největším a nejmenším činil cca dvojnásobek. Pokud byl např. místo faktoru 2 zahrnut faktor 3 o prohřešcích, pak poměr stoupnul přibližně na devítinásobek. Vzniklo by tak několik malých skupin s problémovými úvěry, jejichž existence lze předpokládat ještě před zahájením shlukování.

Světlemodrý shluk (viz Obr. 27) má nejnižší průměrnou úrokovou míru. Velikost úvěru F1 byla jedna z nejnižších, stejně tak průměrný počet kreditních karet F2 byl blízko střední hodnoty. F4 vypovídá o zatížení ostatními úvěry (tj. i úvěry z kreditních karet), které bylo v případě tohoto shluku rovněž nejnižší. Protiklad tohoto shluku se liší především ve větším obnosu požadovaného úvěru, ale i počet karet byl vyšší včetně celkového zatížení ostatními úvěry.



Obr. 27 Shlukování úvěrů a zobrazení průměrné úrokové míry

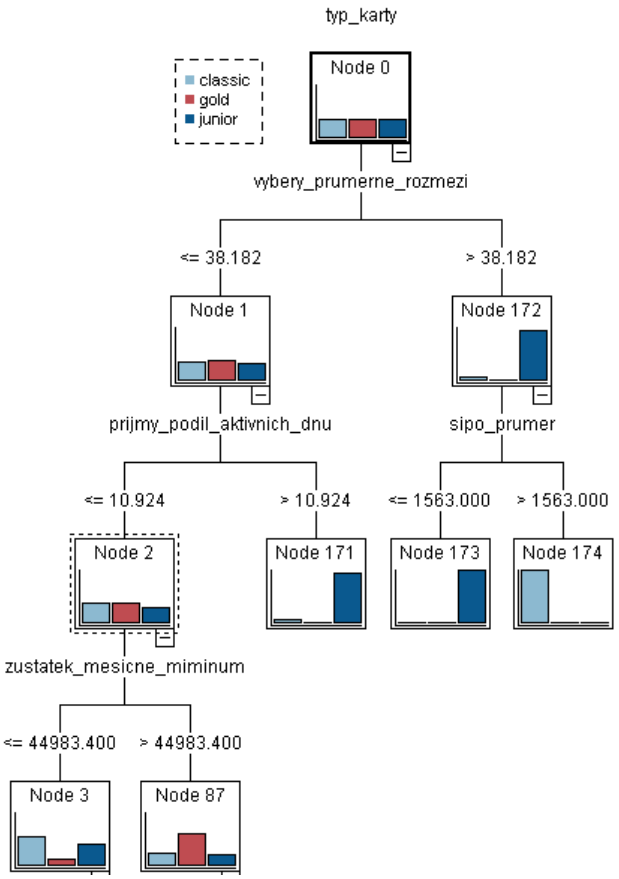
5.4 Přidělování typů karet

V bankovních datech jsou dostupné údaje o klientech, kteří vlastní jeden ze tří typů bankovních karet. Karta *junior* byla poskytována klientům mezi 16 a 24 lety. Po uplynutí horní věkové hranice bylo nutné přejít na typ *classic*. V případě dosažení určitého kritéria bylo možné získat kartu *gold*. U mladých žadatelů rovněž šlo zvolit jiný typ karty než junior, přestože věková hranice nebyla překročena. Záleželo především na chování klienta vyjádřeného pomocí atributů z transakčních údajů. Během klasifikace bylo vhodné atribut věku vypustit právě z důvodu nalezení lepšího pravidla pro přidělení karty. Věk zde hraje roli omezení, které nemusí být přímo zakomponováno do modelu. Před klasifikací bylo samozřejmě nutné vypustit atributy týkající se výběrů kartou. Výchozí nastavení streamu se nachází v Tab. 19.

Tab. 19 Výchozí nastavení streamu pro klasifikaci typů karet

Počet záznamů	Počet atributů	Dělicí semínko	Trénovací m.	Testovací m.
1 982	134	4551136	50%	50%

Rozhodovací stromy

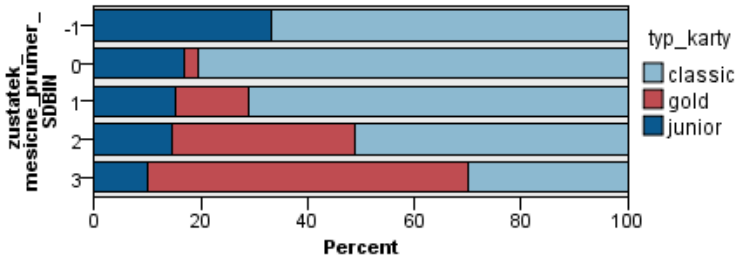


Obr. 28 Strom pro klasifikaci typů karet

Ve druhé hladině stromu (viz Obr. 28) vidíme, že pokud klienti v průměru vybírali u přepážky po 38 dnech, pak šlo v drtivé většině o juniory. Třetí hladina levého podstromu rozděluje klienti podle podílu aktivních dnů příjmových operací. Pokud tedy klient dostával příjem v 11 % dnů z celkového počtu sledovaných dnů od začátku existence účtu, pak šlo o juniora. Čtvrtá hladina rozděluje klienty podle zůstatku, kdy v případě minimálního měsíčního zůstatku většího než 45 000 korun o klienty s kartou gold.

Nejvyšší dosažená přesnost nad testovací množinou byla 92,11 % při míře lokálního prořezávání 88 a minimálním počtu 2 potomků ve větvi. Nejdůležitějším kritériem pro přidělení jednotlivých typů karet je výše *minimálního* či *maximálního průměrného měsíčního zůstatku*. Na Obr. 29 je znázorněno rozložení klientů dle typů karet a kategorizovaného průměrného měsíčního zůstatku. Je patrné, že klienti s

měsíčním zůstatkem pohybujícím se v průměrném (0) či podprůměrném (-1) pásmu vlastnili kartu classic případně junior. Na druhou stranu převažující většina klientů s kartou gold měla měsíční zůstatek výrazně nadprůměrný (3).



Obr. 29 Typy karet podle průměrného měsíčního zůstatku

Neuronové sítě

Základní neboostovaný model dosáhl přesnosti 70,88 %. Neurony ve skrytých vrstvách obsahovaly aktivační funkci *hyperbolický tangens*, zatímco ve výstupní vrstvě byla použita *identita*. Nejlepšího výsledku s přesností 91,41 % dosáhl boostovaný model se sedmi neurony v první skryté vrstvě a třemi neurony ve druhé.

Podle modelu neuronových sítí měly největší význam atributy týkající se *připsaných bonusových úroků* a *průměrného zůstatku*. Níže jsou uvedeny průměry vybraných kategorizovaných atributů týkajících se připsaných úroků podle typu karty. Nejvyšší průměrný měsíční obnos bonusových úroků byl připsán klientům s kartou *gold*, naopak nejméně klientům s kartou *junior*.

Tab. 20 Skupina atributů připsaných úroků nad všemi typy karet[11]

Karta	PÚ měsíčně průměr	PÚ měsíčně minimum	PÚ měsíčně Maximum	PÚ měsíčně odchylka
Junior	0,418	0,342	0,317	0,324
Classic	0,422	0,419	0,352	0,335
Gold	1,311	1,405	0,856	0,345

Support vector machines

Nejvyšší přesnosti bylo dosaženo pomocí nelineární funkce RBF. Lineární jádrová funkce poskytla zhruba o 15 % horší výsledek. V případě této úlohy by bylo vhodnější vybrat model s některou z nelineárních transformačních funkcí. Nejvyšších přesností bylo u jednotlivých funkcí dosaženo při nastavení parametrů uvedených v Tab. 21.

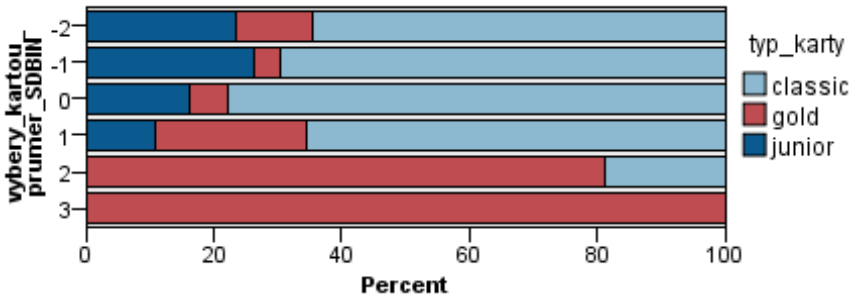
[11] Hodnoty je vhodné srovnávat vertikálně

Tab. 21 Srovnání klasifikátorů SVM pro přidělování typů karet

Typ funkce	Stop kritérium	C	γ	D	Přesnost
RBF	1.0E-1	24	0,4	-	96,06 %
Polynomiální	1.0E-1	35	4,5	5	89,08 %
Lineární	1.0E-1	28	-	-	80,18 %

5.4.1 Shrnutí klasifikace karet

Nejvyšší přesnosti tedy dosáhl model SVM s kernel funkcí RBF, za ním se umístil model rozhodovacího stromu a nakonec neuronové sítě. Vytvořený klasifikátor můžeme použít pro doporučení vhodného typu karty pro jednotlivé klienty. Novému klientovi však může být automaticky přidělen nejčastěji používaný typ karty *classic.* Pokud dosáhne určitého kritéria a bude modelem klasifikován jako klient, pro kterého se více hodí např. typ *gold*, pak může být bankou osloven z důvodu nabídky výměny. Tímto hlavním kritériem může být například výše měsíčního zůstatku, která vyšla u dvou modelů jako klíčová. Při tvorbě modelu byly také vyloučeny atributy týkající se používání karet, jako například průměrná vybraná částka. Důvodem bylo vytvoření modelu upřednostněného pro klienty doposud nevlastnící karty. Model zahrnující tento atribut by šel použít např. při změně typu stávající karty, kdy banka má u daného klienta k dispozici historii o používání karty. Na Obr. 30 je možné vidět, že naprůměrně vysoké výběry z bankomatů byly realizovány hlavně klienty se zlatou kartou. Toto kritérium tedy může být směrodatné pro doporučení změny typu karty.

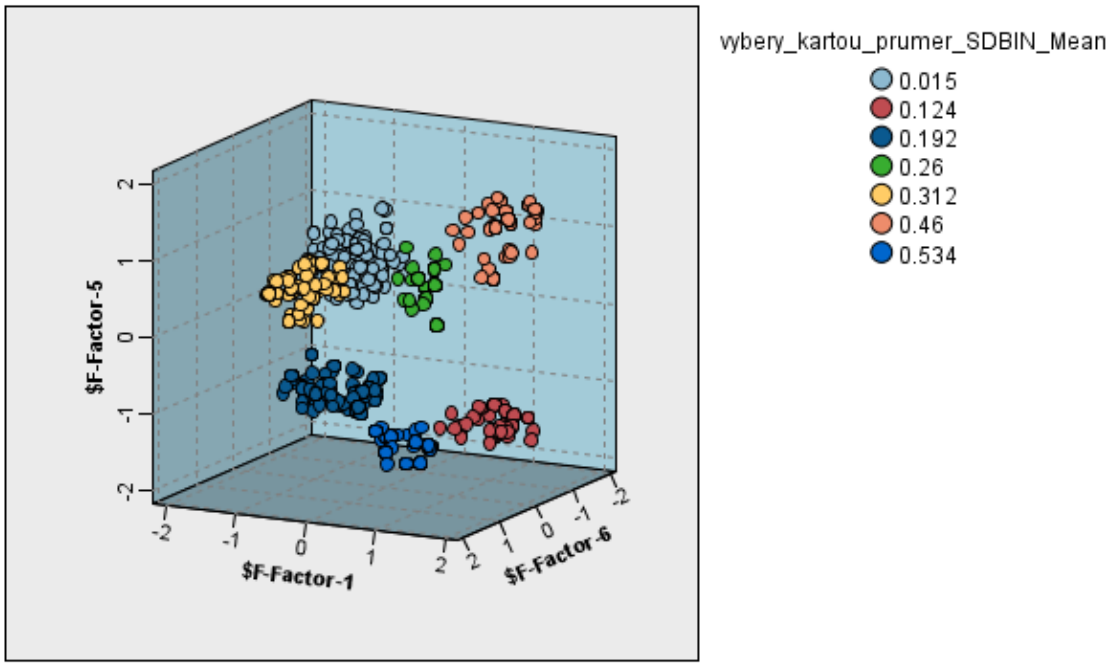


Obr. 30 Rozdělení klientů podle typu karty a průměrných výběrů kartou

5.4.2 Shlukování dle průměrné výše výběru kartou

Výše zmíněné kritérium bude použito pro popis sledované vlastnosti odlišující jednotlivé shluky. Z rotované matice faktorů byly vybrány osy 1, 5 a 6, které byly následně zapojeny do shlukování pomocí k-means. *Faktor první* představuje *délku sledovaného období* vyjádřenou rozmezím mezi první a poslední operací. Takovou operací může být např. změna zůstatku, u které rozmezí vyjadřuje, jak dlouho klient svůj účet aktivně používá. *Faktor pátý* je obsazen proměnnými týkající se sdružených inkasních plateb (SIPO), tedy jejich výší, pokud byly prováděny. Poslední zahrnutý *fak-*

tor s číslem 6 obsahuje proměnné týkajících se *výše průměrného zůstatku* a *výše připsaných bonusových úroků*, které spolu podle výsledku korelují. Toto bylo možné předpokládat již od výsledku modelu vytvořeného pomocí neuronových sítí, které označily zmíněné skupiny atributů za nejdůležitější. Jiné faktory, z nichž některé se týkaly např. příjmů, výdajů nebo důchodů nebyly zahrnuty. Příjmy a výdaje lze nahradit faktorem zůstatku, čímž nám vznikne další prostor pro umístění jiné osy. Důchody se týkají jen určité části klientů, čili by nebyly dobrým podkladem pro vysvětlení zkoumaného jevu z hlediska všech klientů. Respektive zahrnutí tohoto faktoru do shlukování by nemělo smysl, protože skupina důchodců se da určit přímo na základě nenulového důchodu. Výsledek shlukování je v pásmu „fair" a poměr mezi největším a nejmenším shlukem je cca 2,5násobný. Lepšího výsledku shlukování může být dosaženo volbou prvních tří faktorů, které nesou nejvíce vysvětlovaného rozptylu. Kvalita shlukování by v tomto případě byla vyjádřena co nejvíce odlišnou velikostí průměrných výběrů kartou napříč shluky. Na druhou stranu je také možné použít takovou kombinaci faktorů, která problém objasní s použitím ne na první pohled důležitých přitom významných skrytých dimenzí, a to i za cenu horší kvality shlukování.



Obr. 31 Shlukování klientů podle průměrné výše výběru kartou

Na Obr. 31 středně tmavě modrý shluk představuje klienty s největším průměrným výběrem kartou. Jejich průměrný zůstatek F6 byl ze všech skupin nejvyšší. Průměrná výše jejich nájmu F5 byla jedna z nejnižších nebo neplatili nájem vůbec. Jednalo se o klienty, kteří byli v bance aktivní již delší dobu F1. Protikladem je světle modrý shluk, ve kterém se klienti liší podprůměrným zůstatkem, kratší dobou aktivity a nadprůměrně vysokým nájmem.

Tab. 22 Charakteristiky shluků klientů podle výše výběru

Výběry kartou průměr SDBIN	**Faktor 1 Délka aktivity**	**Zůstatek sledované období BIN**	**Faktor 5 Výše nájmu**	**SIPO průměr SDBIN**	**Faktor 6 Výše zůstatku**	**Zůstatek měsíčně průměr SDBIN**
0,015	-0,745	2,514	0,661	0,296	-0,667	0,106
0,124	1,059	5,183	-1,345	-1,000	-0,822	0,360
0,192	-0,806	2,345	-0,985	-0,300	-0,147	0,451
0,260	0,754	4,923	0,661	0,026	0,945	1,678
0,312	-0,892	2,398	0,416	0,046	0,744	1,014
0,460	1,208	5,432	1,073	0,573	-0,811	0,456
0,534	0,842	4,926	-1,313	-	1,060	1,839

6 Závěr

Diplomová práce se zabývala problematikou přípravy dat pro modelování ve vybraném dataminingovém nástroji. Důvodem k výběru tohoto nástroje byla předchozí zkušenost získaná během studií. Nástroj je ideální na seznámení se s širokým rozsahem metod pro práci s daty během celého dataminingového procesu. Metody jsou přístupné intuitivním způsobem, což značně zefektivňuje celý proces, a je tedy eliminována nutnost zjišťovat, jak metodu ovládat. Přesto je však potřeba porozumět jednotlivým algoritmům či postupům do hloubky a pochopit, kdy a jak je použít. SPSS Modeler obsahuje řadu účinných nástrojů pro podporu fáze předzpracování dat. Velmi zjednodušeně řečeno, jednotlivé operace s daty jsou prováděny pomocí tzv. uzlů, které se navzájem propojují a s nastavením parametrů či speciálních výrazů ovlivňují proud procházejících dat. Tento postup se opakuje až do samotné výstavby modelů.

V praktické části byly k dispozici dvě sady dat, z nichž převážná část práce byla věnována datům bankovním. Společně s analýzou dat o nebankovních půjčkách vyplynulo řešení tří dataminingových úloh. Jednalo se o klasifikaci úvěrů, rating a přidělování různých typů karet k účtům. Pro klasifikaci byly použity metody strojového učení, z nichž dvě se podle odborných zdrojů běžně v praxi používají pro automatizované schvalování a zamítání úvěrů. Kromě neuronových sítí a SVM byly testovány modely rozhodovacího stromu C5.0. Vynikaly v lepší interpretovatelnosti, ale především v rychlosti poskytnutí výsledku, který se dostavil i u objemnějších dat během pár sekund.

Pro *klasifikaci úvěrů* byly sledovány odděleně sady bankovních a nebankovních úvěrů, u kterých byly porovnávány klasifikační schopnosti modelů vytvořených zmíněnými typy algoritmů. Testování nejvyšší přesnosti bylo dosaženo pomocí skriptování, které je podporované přímo v SPSS. Umožňuje automatizovat generování modelů s různým nastavením parametrů včetně vypsání výsledků do souborů. V případě bankovních úvěrů bylo dosaženo přesnosti vyšší řádově v desítkách procent oproti nebankovních úvěrům. Téměř stoprocentní predikci má na svědomí dostatek atributů, které pocházejí z transakčních údajů banky, a které spolehlivě vypovídají o chování klienta s daným typem úvěru. Modely jsou nastaveny, tak aby byly připraveny na obě situace stejně, tj. na problematické i spolehlivé klienty. Pokud se reálně očekává např. pouze deset procent problémových úvěrů, pak je poměrně pravděpodobné, že podíl se podaří udržovat i nadále minimální. Pokud máme v obou třídách dostatek záznamů, pak by nemělo docházet ani k opačné situaci, tj. že potenciálně výnosný úvěr by byl modelem eventuálně zamítnut.

Rating úvěrů do více než dvou skupin lze provést již při výstavbě předchozího modelu. Kritériem pro zatřídění úvěru do určitého ratingového pásma může být výsledek ve formě skóre, které poskytuje v SPSS každý klasifikační model predikující hodnoty binární proměnné. Skóre neboli sklon se počítá jako výsledek spolehlivosti (confidence) příslušnosti do modelem klasifikované třídy, který se odečte od hodnoty 1. Hodnota 1 představuje pravděpodobnost příslušnosti do hlavní třídy, která v tomto případě představuje schválený úvěr. U nebankovních úvěrů byl k dispozici

rating, který byl použit jako cílová proměnná této dolovací úlohy. Jednalo se o pokus zpětného určení vzorce či vytvoření modelu pro rating úvěrů poskytovaných u dané společnosti. Model může být použit pro predikci rizikové třídy po zadání parametru úvěru. Můžeme tak například sledovat, při jaké výši by již úvěr spadal do více rizikové třídy, odkud by se odvíjel vyšší úrok. Nejlepšího výsledku bylo dosaženo pomocí MLP, z čehož lze předpokládat, že pro rating byl použit algoritmus bližší neuronovým sítím. Podle výsledku SVM je možné říci, že v případě ratingu lze záznamy mapovat pomocí lineární funkce, která je pro následnou separaci dat vhodnější.

Model pro *klasifikaci karet* slouží k automatickému přidělování druhů karet klientům poprvé žádajících nebo kartu již vlastnících. Po testování modelů se ukázalo, že výše průměrného zůstatku v kombinaci s dalšími atributy je dostatečná pro přidělování vhodného typu karet s vysokou přesností. Z modelů byl vypuštěn atribut věku klienta z důvodu přidělení důležitější role jiným atributům. Typ klienta takto bude odhadnut na základě jiných kombinací vlastností, než je právě věk. V případě, že by banka posunula věkovou hranici pro přidělování karty typu *junior*, modely vystavěné na tomto atributu by přestaly být validní. Pokud bude modelem vyhodnocen typ karty, který klient nemůže vlastnit např. kvůli věku, pak bude přidělen výchozí typ *classic*. Díky SVM se ukázalo, že data jsou poměrně výrazně lineárně neseparabilní. V takovém případě je potřeba použít nelineární mapovací funkci, z nichž model s RBF dosáhl nejlepšího výsledku nad všemi typy modelů.

Dalším postupem v praktické části bylo použití připravených množin určených původně pro klasifikaci k deskriptivní úloze realizované pomocí shlukování. Pro clusterování byl použit algoritmus k – means. K výběru atributů pro zapojení do shlukování bylo použito faktorové analýzy, která se liší od použití metody analýzy hlavních komponent mimo jiné v odvození většího počtu skrytých faktorů než dvou. FA seskupuje původní proměnné do nových, které zachovávají maximum rozptylu původního datového souboru. Kromě toho FA může sloužit k vysvětlení souvislostí mezi jednotlivými proměnnými uvnitř skupiny popsané skrytým faktorem, kde jsou zjišťovány korelace mezi konkrétní proměnnou a hodnotou faktoru. Provádí se tedy i za účelem nacházení vzorů v datech, které prostřednictvím faktorů vysvětlují souvislosti mezi proměnnými. Do shlukování byly zapojeny faktory, které obsahují převážnou část rozptylu, čímž lze pozitivně ovlivnit výsledky k - means. V některých případech mohou shluky vyniknout ještě před použitím k – means, protože samotná FA či PCA vede k rozptylování záznamů dle vytvořených os. Rozdíl je patrný mezi bankovními a bankovními daty, kde ve druhém případě vedl menší počet záznamů k viditelnějšímu seskupení v trojrozměrném prostoru, aniž bychom museli omezovat vyobrazení záznamů v určité vzdálenosti od středu svého shluku. Pokud bychom FA či jinou podobnou metodu nepoužili, museli bychom náhodně zkoušet různé proměnné, abychom dosáhli optimálního výsledku shlukování. Je vhodné vybírat takové faktory, které lze sledovat u většiny případů, v opačném případě by skupiny mohly být zřejmé bez použití shlukování. Pozitivně ovlivnit clustrování lze i zahrnutím skóre vygenerovaného klasifikačním modelem, nicméně tento postup je aplikovatelný pouze u úloh s binární cílovou proměnnou. Jako možné rozšíření práce se nabízí tvorba konkrétních opatření pro vytvořené segmenty klientů.

7 Literatura

BERRY, Michael J. a Gordon LINOFF, c2004. *Data mining techniques: for marketing, sales, and customer relationship management.* 2nd ed. Indianapolis, Ind.: Wiley, xxv, 643 p. ISBN 978-0-471-47064-9.

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

ČESKÁ SPOŘITELNA, 2010*. Řízení úvěrových rizik v praxi* [online]. 23. 4. 2010 [cit. 2015-11-28]. Dostupné z:

http://www.csas.cz/static_internet/cs/Komunikace/Tiskove_centrum/Prezentace_novinari/Prilohy/100423_Credit_risk.pdf

CLEFF Thomas, 2014. *Exploratory Data Analysis in Business and Economics An Introduction Using SPSS, Stata, and Excel.* Cham: Springer. ISBN 978-3-319-01517-0.

DING, Chris a Xiaofeng HE, 2006. *K-means Clustering via Principal Component Analysis* [online]. 2015-11-22 [cit. 2015-11-22]. Dostupné z:

www.vision.caltech.edu/wikis/EE148/images/c/c2/KmeansPCA1.pdf

GIUDICI, Paolo, 2003. *Applied data mining: statistical methods for business and industry.* Chichester: Wiley, xii, 364 p. ISBN 0-470-84679-8.

GUPTA, Gaurav a Himanshu AGGARWAL, 2012. Improving Customer Relationship Management Using Data Mining. *International Journal of Machine Learning and Computing* [online]. 2012, vol. 2, issue. 6, p. 874-877 [cit. 2015-05-03]. DOI: 10.7763/ijmlc.2012.v2.256. Dostupné z:

http://www.ijmlc.org/papers/256-L40070.pdf

HAN, Jiawei, Micheline KAMBER a Jian PEI, c2012. *Data mining: concepts and techniques*. 3rd ed. Boston: Elsevier, xxxv, 703 p. Morgan Kaufmann series in data management systems. ISBN 978-0-12-381479-1.

HSSINA, Badr et al., 2014. A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications* [online]. vol. 4, issue 2 [cit. 2015-11-27]. DOI: 10.14569/SpecialIssue.2014.040203. ISSN 2158107x. Dostupné z:

http://thesai.org/Publications/ViewPaper?Volume=4&Issue=2&Code=SpecialIssue&SerialNo=3

IBM CORPORATION, 2011. *IBM SPSS Modeler CRISP-DM Guide* [online]. [cit. 2015-11-17]. Dostupné z:

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

IBM CORPORATION, 2012. *IBM Knowledge Center* [online]. [cit. 2015-11-17]. Dostupné z: http://www-01.ibm.com/support/knowledgecenter/

JOLLIFFE, I. T., c2002*. Principal component analysis.* 2nd ed. New York: Springer, xxix, 487 p. Springer series in statistics. ISBN 0-387-95442-2.

KREDITECH, 2015. *Kredito24: jak se pozná dlužník, aneb co to je skóring* [online]. [cit. 2015-11-28]. Dostupné z:

https://www.kredito24.cz/content/jak-se-pozna-dluznik-aneb-co-to-je-skoring/

LENDING CLUB, c2006-2015. *Lending Club Statistics* [online]. San Francisco [cit. 2015-11-09]. Dostupné z: https://www.lendingclub.com/info/download-data.action

MAIMON, Oded a Lior ROKACH, c2005. *Data mining and knowledge discovery handbook.* New York: Springer, 1383 p. ISBN 0-387-2546-5.

MOIN, Kazi Imran a Qazi Baseer AHMED, 2012. Use of Data Mining in Banking. *International Journal of Engineering Research and Applications* [online]. vol. 2, issue 2, p. 738-742 [cit. 2015-11-28]. ISSN 2248-9622. Dostupné z:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.7821&rep=rep1&type=pdf

LAROSE, Daniel T, c2005. *Discovering knowledge in data: an introduction to data mining.* Hoboken, N.J.: Wiley-Interscience, xv, 222 p. ISBN 0-471-66657-2.

OLSON, David L. a Dursun DELEN, c2008. *Advanced data mining techniques.* Berlin: Springer, xii, 180 p. ISBN 978-3-540-76917-0.

PIATETSKY, Gregory, 2015. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDnuggets* [online]. [cit. 2015-04-22]. Dostupné z: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

PETR, Pavel, 2006. *Data Mining.* Vyd. 1. Pardubice: Univerzita Pardubice. ISBN 80-7194-886-1.

PKDD'99 Discovery Challenge: A Collaborative Effort in Knowledge Discovery from Databases [online]. Praha, 1999 [cit. 2015-11-09]. Dostupné z: http://lisp.vse.cz/pkdd99/Challenge/chall.htm

ROHANIZADEH, Seyed Soroush, 2009. A Proposed Data Mining Methodology and its Application to Industrial Procedures. *Journal of Optimization in Industrial Engineering* [online]. vol. 2, issue 4 [cit. 2015-04-22]. Dostupné z:

http://www.qjie.ir/?_action=showPDF&article=31&_ob=2e9f779810eaef02d9bcc00959616080&fileName=full_text.pdf

ROKACH, Lior a Oded MAIMON, 2015. *Data mining with decision trees: theory and applications.* 2nd ed. Hackensack, New Jersey: World Scientific, xxi, 305 p. ISBN 978-9814590075.

RYCHLÝ, Marek, 2003. *Klasifikace a predikce* [online]. Brno: Vysoké učení technické v Brně, Ústav informačních systémů [cit. 2015-04-22]. Dostupné z:

http://www.fit.vutbr.cz/~rychly/public/docs/classification-and-prediction/classification-and-prediction.pdf

SADATRASOUL, S. M. et al., 2013. Credit scoring in banks and financial institutions via data mining technique. *Journal of AI and Data Mining* [online]. vol. 1, issue 2, p. 119-129 [cit. 2015-11-28]. Dostupné z:

http://jad.shahroodut.ac.ir/article_124_0.html

TAN, Pang-Ning, Michael STEINBACH a Vipin KUMAR, c2006. *Introduction to data mining.* 1st ed. Boston: Pearson Addison Wesley, xxi, 769 p. ISBN 0321321367.

TUFFÉRY, Stéphane, 2011. *Data mining and statistics for decision making.* Hoboken, NJ.: Wiley, xxiv, 689 p. ISBN 978-0-470-68829-8.

WANG, John, c2009. *Encyclopedia of data warehousing and mining*. 2nd ed. Hershey: Information Science Reference, 4 v. ISBN 9781605660110.

WITTEN, I, Eibe FRANK a Mark A HALL, c2011. *Data mining: practical machine learning tools and techniques*. 3rd ed. Burlington: Morgan Kaufmann, xxxiii, 629 p. Morgan Kaufmann series in data management systems. ISBN 978-0-1237-4856-0.

Přílohy

A Ukázka testovacího skriptu v SPSS

Tento skript byl použit pro generování modelů SVM s různým nastavením parametrů a ukládání výsledků do souboru.

```
var svmNode
...
var regmax

set svmNode = svmA
set modelName = "svmA"
set svmAnalysis = AnalyseA
set inputNode = PartitionA

set typ="Polynomial"                    # výběr kernel funkce
set stop=1                              # stop kritérium - nastaveno na nejnižší hodnotu
set regmin = 1                          # dolní hranice regularizačního parametru
set regmax = 10                         # horní hranice regularizačního parametru

for I from regmin to regmax

 set ^svmNode:svmnode.mode=Expert
 set ^svmNode:svmnode.stopping_criteria = "1.0E-"><stop
 set ^svmNode:svmnode.regularization = I

 if typ="RBF" then

  for J from 0 to 3

  set rbgmm = 0.02*J+0.01               # rbf gamma testována v rozmezí 3/k a k/6, tedy např. 0,02 až 0,04
  set ^svmNode:svmnode.kernel = typ
  set ^svmNode:svmnode.rbf_gamma = rbgmm

  execute ^svmNode                      # spuštění generování modelu algoritmem SVM

  insert model ^modelName connected between ^inputNode and ^svmAnalysis
                                        # vložení modelu mezi uzly Partition a Analysis z palety vygenerovaných modelů
  set ^svmAnalysis.output_mode = File
  set ^svmAnalysis.output_format = text
          set ^svmAnalysis.full_filename = "D:/vysledky/banka/svm/svm_rbf_regpar_"><I><"_gamma_"><rbgmm><"_stop-
ping_"><stop><".txt"
                                        # nastavení parametrů uzlu Analysis, který uloží výsledky do souboru

  execute ^svmAnalysis                  # spuštění uzlu Analysis

  delete ^svmNode:applysvmnode          # vymazání vloženého modelu
  clear generated palette               # vymazání palety s vygenerovanými modely – uvolnění RAM

  endfor

 else

  if typ="Polynomial" or typ="Sigmoid" then
   for J from 1 to 10

      set gmm=0.5*J
      set ^svmNode:svmnode.kernel = typ
      set ^svmNode:svmnode.gamma = gmm

      if typ="Polynomial" then

       for K from 1 to 5                # stupeň polynomu
         set ^svmNode:svmnode.degree = K
```

```
            execute ^svmNode

            insert model ^modelName connected between ^inputNode and ^svmAnalysis

            set ^svmAnalysis.output_mode = File
            set ^svmAnalysis.output_format = text
              set ^svmAnalysis.full_filename = "D:/vysledky/banka/svm/svm_"><typ><"_reg-
par_"><I><"_gamma_"><gmm><"_stopping_"><stop><"_degree_"><K><".txt"

            execute ^svmAnalysis

            delete ^svmNode:applysvmnode
            clear generated palette

          endfor

          else

           execute ^svmNode

           insert model ^modelName connected between ^inputNode and ^svmAnalysis

           set ^svmAnalysis.output_mode = File
           set ^svmAnalysis.output_format = text
              set ^svmAnalysis.full_filename = "D:/vysledky/banka/svm/svm_"><typ><"_reg-
par_"><I><"_gamma_"><gmm><"_stopping_"><stop><".txt"

           execute ^svmAnalysis

           delete ^svmNode:applysvmnode
           clear generated palette

         endif
      endfor
    else
         set ^svmNode:svmnode.kernel = typ

         execute ^svmNode

         insert model ^modelName connected between ^inputNode and ^svmAnalysis

         set ^svmAnalysis.output_mode = File
         set ^svmAnalysis.output_format = text
              set ^svmAnalysis.full_filename = "D:/vysledky/banka/svm/svm_"><typ><"_regpar_"><I><"_stopping_"><stop><".txt"

         execute ^svmAnalysis

         delete ^svmNode:applysvmnode
         clear generated palette
    endif
   endif
endfor
```

B Ostatní přílohy

Přiložený disk obsahuje následující položky:

- výstupy z modelů pro jednotlivé úlohy
    - výkony modelů
    - evaluační grafy
    - nastavené parametry
- projektové složky s množinami „bankovních“ a „nebankovních“ dat obsahující:
    - data a jejich původní popis
    - SPSS soubor s projektem. Projekt je s použitím SuperNode hierarchie rozdělen na několik specifických streamů.
    - testovací skripty