Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra Statistiky

Bakalářská práce

Big Data - výzva pro statistickou analýzu

Josef Štec

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Josef Štec

Informatika

Název práce

Big Data – výzva pro statistickou analýzu

Název anglicky

Big Data – challange for statistical analysis

Cíle práce

Cílem bakalářské práce bude vyhodnocení vybrané zákaznické databáze statistickými metodami užívanými v Big Data. Snahou bude identifikovat faktory, které chování zákazníka/klienta ovlivňují.

Metodika

Těžiště práce bude spočívat v analýze a vyhodnocení rozsáhlejší databáze. K řešení budou využity statistické metody mající uplatnění v Big Data, tj. metody vícerozměrné statistiky a metody prediktivního modelování.

Doporučený rozsah práce

30 – 40 stran

Klíčová slova

Big data, klient, faktor, chování, statistická analýza

Doporučené zdroje informací

ABBOT, D.: Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, 2014. ISBN: 978-1-118-72796-6

AGRESTI, A. *Categorical data analysis.* Hoboken: John Wiley & Sons, 2013. ISBN 978-0-470-46363-5.

HENDL, J. *Přehled statistických metod : analýza a metaanalýza dat.* Praha: Portál, 2012. ISBN 978-80-262-0200-4.

RUD, P., O.: Data Mining. Praha: Computer Press, 2002. ISBN 8072265776

SIEGEL, E. Predictive Analytics. Hoboken: John Wiley & Sons, 2013. ISBN 978-1-118-35685-2

WHITE, T.: Hadoop: The Definitive Guide, 3rd Edition, 2012. ISBN: 978-1-4493-1152-0

Předběžný termín obhajoby

2016/17 LS – PEF

Vedoucí práce

Ing. Tomáš Hlavsa, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 25. 11. 2016

prof. Ing. Libuše Svatošová, CSc.

Vedoucí katedry

Elektronicky schváleno dne 25. 11. 2016

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 04. 03. 2017

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Big Data - výzva pro statistickou analýzu" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 15.3. _____

Poděkování

Rád bych poděkoval doktoru Tomáši Hlavsovi za skvělé vedení práce, pevné nervy, pomoc s výběrem odborné literatury a tématu, dále také za průběžné konzultace. Také bych rád poděkoval Petru Nečesalovi z ČSOB za poskytnutí dat k analýze.

Big Data - výzva pro statistickou analýzu

Souhrn

Obsah bakalářské práce analyzuje Big Data čili velký objemem dat, který pochází z transakční databáze ČSOB. Práce je rešeršního charakteru a obsahuje analýzu transakcí prováděných kreditními kartami za období tří měsíců. Cílem teoretické práce je přiblížit problematiku Big Data, jak z pohledu technologického myšleno nástroje pro skladování, zprávu dat a jejich zpracování, tak z pohledu statistické analýzy a metod ke zpracování dat používaných. V praktické části je pak realizována analýza vzájemné podobnosti skupin MCC kódů daných transakcí. A případné spojení transakcí v určité kategorii s rizikovostí klienta z pohledu na historická data. Teoretická část je rozdělena do čtyř pododdílů první definuje pojem Big Data z technologického pohledu, což zahrnuje technologie k uchovávání a ke zpracování dat. Druhá část popisuje práci s daty z pohledu statistické analýzy. Ve třetí a čtvrté části jsou objasněny pojmy MCC a Basilejský model.

Klíčová slova: Big Data, klient, faktor, chování, statistická analýza

Big Data - challange for statistical analysis

Summary

The content of this bachelor‘s thesis is to analyze Big Data which means a large amount of data which comes from a CSOB transaction database. The thesis contains an analysis of credit card transactions within a period of three months. The goal of theoretical part is to approach the problematic of Big Data, from both technological angle, meaning tools for storing, administering and processing of data and the angle of statistical analysis and methods used for the data processing. The practical part of this thesis implements analysis of mutual similarity between groups of MCC codes on given transactions, and the possibility of connection between certain categories and the riskiness of clients by looking at a historical data. Theoretical part is split into four parts first part defines the Term Big Data from a technological standpoint, including technologies used for storing and processing data. Second Part describes statistical methods for analysis of Big Data. Third and fourth parts clarify the terms MCC and Basel model.

Keywords: Big data, client, factor, behavior, statistical analysis

Obsah

Seznam tabulek

Odkazovaný seznam tabulek

1 Úvod

Dnes žijeme v době, kdy jednou z nejmocnějších zbraní jsou informace. Ať už se jedná o armádu, vládu nebo nadnárodní korporace pro všechny jsou nejcennější informace, které jim pomohou dostat se před jejich konkurenty. A kde se tyto informace berou? Informace se nacházejí v datech. Když si jdete koupit rohlíky, jedete do práce, telefonujete, cestujete nebo jen sedíte doma a trávíte čas na sociálních sítích, zanecháváte za sebou datovou stopu. A ten kdo s těmito daty umí pracovat, získává nad ostatními obrovské výhody. Obzvláště velké korporace se snaží vyvíjet a inovovat způsoby, jak data uchovávat, zpracovávat a získávat pro co nejlepší uspokojení svých zákazníků a navýšení svých zisků. A tak se správná datová analýza stala jednou z největších nutností, bez které nelze nikomu v dnešní době konkurovat. Dle IBM se každým dnem vygeneruje 2,5 quintillion bytů dat většina těchto dat je nestrukturalizovaných a neurčitých. Dle IBM se také 90% těchto dat vygenerovalo za poslední dva roky. (IBM 2017)

IDC (International Data Corporation) předpovědělo, že veškerá data v digitální sféře bude k roku 2011 obsahovat 1.8 ZettaBytů to je 10^{21} Bytů nebo také 1mil. PetaBytů. (White 2012).

Jedná se o veškerá data, která člověk generuje, záznamy senzorů, GPS a telefonní signály, data ze sociálních sítí, obrázky, videa, statusy, komentáře a mnoho dalších. Takové množství dat nelze uchovávat ani zpracovávat běžnými prostředky. A tak je důležité v případě Big Data věnovat velkou pozornost skladování dat, a to zejména na systém a metodiku ukládání, které jsou zapotřebí pro takto velké objemy dat. Další velmi důležitou částí práce s Big Data je kvalifikace odborníků pro co největší efektivitu práce s danými daty.

Práce se v následujících kapitolách bude těmto tématům včetně samotné analýzy data blíže věnovat, a to jak po teoretické, tak po praktické stránce. Česká odborná literatura termín Big Data nijak nepřekládá a pracuje se s ním, jako označením technické kategorie je možné ho psát, jak s velkými, tak malými písmeny. Pro tuto práci byla zvolena varianta s velkými písmeny.

2 Cíl práce a metodika

2.1 Cíl práce

Cílem bakalářské práce bude vyhodnocení vybrané zákaznické databáze statistickými metodami užívanými v Big Data. Snahou bude identifikovat, zda kategorie MCC, které klienta ovlivňují, jsou vzájemně slučitelné pro usnadnění následné rizikové analýzy klienta.

2.2 Metodika

Tvorba a zpracování teoretické části, probíhala v několika fázích. V první fázi budou vybrány odborné zdroje pro nastudování a popsání problematiky, kterou se práce zabývá. Jako zdroje budou vybrána jak odborná literatura v podobě knih, tak různé elektronické zdroje v podobě článků a příspěvků zabývajících se danou problematikou, a to na českých i zahraničních portálech. Po nastudování materiálů bude připravena osnova a rešerše, která bude následně na základě uvedených zdrojů sepsána.

Těžiště práce bude spočívat v analýze a vyhodnocení rozsáhlejší transakční databáze ČSOB. Nejprve bude extrahován dostatečně velký vzorek, který bude očištěn o případné chyby v datech. Následně bude vzorek analyzován pomocí softwarových nástrojů od společnosti SAS Institue, zejména SAS Enterprise Guide. Vzorek bude následně prozkoumán základními statistickými metodami, jako je například procedura univariate, která odhalí momenty a základní statistická měřítka jako je velikost vzorku, průměrná hodnota, směrodatná odchylka šikmost a špičatost, variační koeficient, rozptyl.

Tyto hodnoty se vypočítávají na základě následujících vzorců:

Velikost vzorku je jednoduchý součet řádků v tabulce.

$$N = \sum záznamů \quad (1)$$

Aritmetický průměr je statistická veličina, vyjadřující typickou hodnotu soubory. Vypočte se součtem všech hodnot souboru a následným vydělením jejich počtem.

$$\mu = \frac{\sum X}{N} \quad (2)$$

Rozptyl je jedním z centrálních momentů náhodné veličiny. Jedná se o střední hodnotu kvadrátů odchylek od střední hodnoty.

$$\sigma^2 = \frac{1}{n}\sum_{i=0}^{n}(x_i - E(x))^2 \quad (3)$$

Směrodatná odchylka je kvadratický průměr odchylek hodnot od aritmetického průměru. Vypovídá tedy o tom, jak moc se liší vzájemně liší typické hodnoty v soboru zkoumaných čísel.

$$\sigma = \sqrt{\sigma^2} \tag{4}$$

Další dvě popisné charakteristiky jsou míry šikmosti a špičatosti. Tyto charakteristiky pomáhají určit, jak moc se rozdělení dat v souboru, podobá nebo naopak odlišuje od normálního rozdělení tedy Gaussovy křivky. K jejich určení se používají tzv. momenty. Moment Ktého stupně se vypočte následovně.

$$m_k = \frac{\sum(x_i - \mu)^k}{n} \tag{5}$$

Šikmost poté určuje, kterým směrem je proměnná asymetricky rozložená. Šikmost se dělí na kladno neboli pravostrannou a to v případě, že se většina získaných hodnot nachází pod průměrem. A šikmost zápornou neboli levostrannou, v případě, že většina hodnot je naopak větší než průměr. Pokud vyjde šikmost nulová vypovídá to o symetričnosti souboru.

$$\gamma_1 = \frac{m_3}{\sqrt{m_2^3}} \tag{6}$$

Špičatost určuje, jak se v rozdělení četností vyskytují vysoké a nízké hodnoty. Podle výsledku se pak dělí na více špičaté než normální rozdělení nebo na méně špičaté než normální rozdělení. Podobně jako u šikmosti nulová hodnota výsledku značí rozdělení normální.

$$\gamma_2 = \frac{m_4}{m_2^2} - 3 \tag{7}$$

Variační koeficient se používá k porovnání variability dvou nebo více souborů s odlišnou úrovní hodnot.

$$C_v \frac{\sigma}{\mu} \tag{8}$$

Jedno faktorová analýza rozptylu, ověřuje statistický význam hodnoty některého znaku, který se dá pozorovat. Tento znak musí nabývat konečného počtu hodnot minimálně však dvou.

$$F = \frac{\sigma^2"mezi\ skupinami"}{\sigma^2"uvnitř\ skupin"} \tag{9}$$

Kruskal-Wallisův test je neparametrickou alternativou jedno faktorové analýzy rozptylu. To znamená, že se testuje, zda vzorky pocházejí ze stejného rozložení.

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\overline{r_i}-\overline{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij}-\overline{r})^2} \tag{10}$$

Wilcoxonův test je neparametrický párový test diferencí. Tento test se používá pro testování hypotéz pro srovnání párů nebo závislostí v případě, že základní soubor není normálně rozdělen.

$$W = \sum_{i=1}^{N_r}[sgn(x_{2,i} - x_{1,i}) * R_i] \tag{11}$$

Dunnova metoda mnohonásobného srovnání zamítá nulovou hypotézu, pokud platí následující vztah:

$$|T_i - T_j| > u_{\frac{2\alpha}{k(k-1)}}\sqrt{\frac{N(N+1)}{12} * (\frac{1}{n_i} + \frac{1}{n_j})} \tag{12}$$

3 Teoretická východiska

3.1 Big Data

Jak bylo již dříve zmíněno, žijeme v době, ve které se generují stále větší množství dat a jejich správné využití je také zcela zásadní problém. Big Data se snaží tato data určitým způsobem uchopit, uchovat, provést nad nimi analýzu, prohledávat je a vizualizovat je. To ovšem přináší spoustu komplikací s jejich uchováním.

Big Data má mnoho definic jednou z nich (nejvíce známou a často uváděnou) je technologicko-výzkumné společnosti Gartner: „Big Data je termín aplikovaný na soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými nástroji v rozumném čase.“ (Dolák 2011) od technologicko-výzkumné společnosti Gartner.

Nebo také:

„Big data je termín popisující velké množství dat – jak strukturované, tak nestrukturované – které zaplavují společnosti každý den. Na čem záleží, je způsob, jak společnosti s daty nakládají. Big Data mohou být analyzována pro lepší náhled, který vede k lepším, rozhodnutím a strategickým obchodním rozhodnutím.“ (SAS 2016)

3.1.1 Úvod do Big Data

Jde o obrovské objemy dat, které se skládají z různých druhů, typů a původu. Pokud by se jednalo o prostá číselná data, nad kterými lze provádět běžné matematické a statistické operace jako sčítání, odčítání, násobení, průměrování atd. stačilo by k jejich zpracování využití běžných nástrojů. Ale o běžná číselná data se bohužel nejedná, jedná se o data nestrukturovaná, tedy taková, se kterými si běžné nástroje neporadí. Příkladem takových dat jsou data ze sociálních sítí. Taková data se mohou lišit v několika aspektech například v jazyce, použitých zkratek, zdrobnělin a slangu v závislosti na tom, kdo s kým komunikuje. Velké množství dat na sociálních sítích je totiž volně přístupné a poskytují velmi zajímavé informace o tom, kdo, s kým komunikuje, jaké má potenciální zájmy a jaký produkt by bylo vhodné mu nabídnout. Tato data ovšem, jak již bylo zmíněno, nejsou strukturovaná, a tak je nelze analyzovat běžnými počítačovými algoritmy. Zpráva velkých dat je pro firmy v dnešní době esenciální, a proto se klade velký důraz na technologie,

které je dokážou sbírat, uchovávat a analyzovat, neboť to jim dává nezanedbatelnou výhodu oproti konkurenci. V oblasti vědy, sociologie a byznysu existuje široká nabídka aplikací. (Arthur 2013, Rouse 2014, SAS 2016).

Ve zkratce se tedy jedná o obrovské a stále rychle rostoucí zdroje dat nebo informací, které také představují široký záběr komplexních problémů a analýz. Které mohou zahrnovat zavádění nové infrastruktury, sběr, zpracování a analýzu dat v téměř reálném čase. A tedy technologie Big Data popisují novou generaci technologií a architektur pro co nejefektivnější práci a analýzu shluku informací v podobě ohromného objemu různorodých dat (Villars aj. 2011).

3.1.2 **Vznik Big Data**

Zdrojů takovýchto velkých a nestrukturalizovaných dat je mnoho. Počítače, herní konzole, gps systémy, telefony, tablety a v dnešní době už i tzv. chytré hodinky a televize si ukládají data ve formě logů, a to jak samy v rámci svého operačního systému, tak jednotlivé programy a aplikace, které si vytváří vlastní loga. (Dolák 2011)

Dalšími sběrači dat jsou internetové prohlížeče a vyhledávací nástroje jako je například google nebo seznam. Tyto nástroje si tato data uchovávají pro zlepšení služeb, kdy nám díky historii vyhledávání dokážou lépe napovědět při dalším vyhledávání, ale také proto, aby věděly, jaké reklamy nám zobrazit. Důležité je si také uvědomit, že mnoho organizací a společností počínaje zdravotními či vládními organizacemi, obchody poskytující věrnostní karty a e-shopy konče o nás také sbírají data. Toto má také za důsledek nejen zlepšení poskytovaných služeb, ale také zlepšení vlastních potřeb dané organizace, jako například marketing. Jedním z největších benefaktorů jsou kreditní společnosti, neboť tyto společnosti uchovávají transakční data klientů podle nichž lze následně odhadovat jejich chování. Tyto společnosti dostávají informace vždy, když se použije karta. Jsou to informace o poloze, ceně, obchodě, zboží a frekvenci použití karty. To poskytuje informace jak kreditním společnostem, které na základě těchto údajů dokážou udělat analýzu rizik, ale také opět spoustu užitečných informací pro marketingové společnosti (Arthur 2013, Rouse 2014).

3.1.3 **Využití Big Data**

Experti zastávají názor, že je důležité mít jednak správná data a také správné prostředky pro jejich zpracování. Dále také je důležitou částí práce s Big Data nebát se vyhledávat nová řešení a nové metod myšlení. Big Data jsou v očích odborníků zásadní pro inovaci, konkurenci a produktivitu (McKinsey aj. 2016). Pro co nejvyšší efektivitu využití dat je zapotřebí používat správné prostředky a mít dostatečně kvalifikovaný personál. A to vyžaduje značné finanční investice.

Většina společností má představu, že bude schopna vzít jakákoli data z jakéhokoli zdroje, vytěžit z daného zdroje potřebná data a ty dále analyzovat. Z analýzy pak vyvodit patřičné závěry – ušetření peněz, času, zdrojů, případná optimalizace produktů a zejména napomoci k lepšímu obchodnímu rozhodování. Díky všem datům, která tyto společnosti získávají od zákazníků, jak už stávajících, tak potencionálních pomocí různých průzkumů a marketingových kampaní, tak mnoha dalšími způsoby. Mohou tyto společnosti lépe predikovat a do jisté míry i ovlivňovat budoucí vývoj a chování trhu. To je největším potenciálem využití Big Data. Těchto možností nejlépe a nejvíce využívají kreditní, marketingové a farmaceutické společnosti. Dále je ovšem mohou využívat i vládní instituce, například pro snazší evidenci obyvatel, odhalit potenciální příčiny poruch a závad a včasně je řešit a šetřit tak velké množství peněz. Dodávkové služby mohou lépe optimalizovat své trasy a šetřit tak náklady s dodáním zboží. A také je mohou využívat vysoké školy pro výzkum a vlastní univerzitní projekty (McAfee 2012, McKinsey aj. 2016).

Velikost objemu Big Data, který má daná společnost k dispozici je sama o sobě nepoužitelná. Bez správných analytických metod a prostředků pro zpracování a kvalitního personálu jsou Big Data téměř bez hodnoty a dalo by se říci, že jsou v podstatě pro společnost bez významu. A tedy to, že společnost má k dispozici velký objem dat nemusí ve finále znamenat onu podstatnou výhodu.

„Pokud jsou data nekompletní, nedávají smysl či obsahují chyby, mohou vést ke špatným rozhodnutím, která mohou podkopat konkurenceschopnost firmy či poškodit životy jednotlivců. Jeden z klasických případů toho, jak nesouvisející data mohou vyvolat nechtěné závěry“, popsal profesor Gary King z Harvardova ústavu pro kvantitativní společenské vědy. V rámci projektu s využitím velkého objemu dat měly být využity tweety a příspěvky z jiných sociálních sítí k tomu, aby předpovídaly míru nezaměstnanosti

v USA pomocí monitorování klíčových slov jako „zaměstnání“, „nezaměstnanost“ a „inzerát“. Byla využita analytická technika, při níž byly do skupin sdružovány příspěvky obsahující tato slova, které byly dále zkoumány. Při monitoringu výzkumníci narazili na vysoký počet tweetů obsahujících jedno z těchto klíčových slov. Avšak, jak King zjistil později, nemělo to žádnou souvislost s nezaměstnaností. „Jednoduše jsme si nevšimli toho, že zemřel Steve Jobs,“ řekl King. Nebýt smrti legendárního zakladatele Apple, byl by zvýšený výskyt slova „jobs“ (anglicky „zaměstnání“) úsměvnou historkou. „Podobným problémům se můžete pokusit vyhnout přidáním výjimek, nikdy si však nemůžete být jistí,“ dodal King s tím, že relevanci určitých pojmů si může vyzkoušet každý sám. „Stačí je zadat do webového vyhledávače. Zobrazí se jak relevantní výsledky, tak ty očividně o něčem úplně jiném.“ (Stančík, 2013).

3.1.4 **Technologie skladování**

Jak již bylo zmíněno při práci s Big Data je zapotřebí inovovat a investovat do nových metodik, organizačních struktur a zejména do technologií, nástrojů a personálu, a to jak najímání již zkušených a talentovaných lidí, tak případné rekvalifikace současných zaměstnanců.

Bez řádného uložení a tedy i uskladnění dat by se práce s Big Data značně znepříjemnila a je tedy potřeba data správně skladovat. Toto platí o datech obecně, nejen pouze o Big Data. Data se tedy obecně skladují do tak zvaných datových skladů (Data Warehouses, zkráceně DWH). Těm, které umožňují analyticky zpracovávat data, se říká OLAP(Online Analitical Processing). První potřeba DWH vznikla při hromadném nasazování informačních systémů (IS), ty generovaly velké množství dat, které bylo potřeba někde uchovat pro další využití k analýze a získání různých statistických údajů. Pokud je DWH správně implementovaný, samostatně a pravidelně si bere data z IS a ukládá si je ve své databázi. Zde jsou data ukládána relačním způsobem a umožňuje tedy náhled na určité části provozního IS (Adshead 2013, Dolák 2011, Matoušek 2014).

DWH využívají tří vrstvové architektury. Na spodní vrstvě této architektury se nachází servery skladu, na nichž jsou uloženy relační databáze. Uprostřed se nachází OLAP server a nad ním v poslední vrstvě je klient. Tento systém lze rozdělit na dvě části a to OLAP a OLPT(Online Transaction Processing). Na rozdíl od OLAP, OLPT ukládá

záznamy jednotlivých transakcí a realizují se použitím relačních databázových technologií. V případě klasického OLPT jsou data agregována a následně ukládána do DWH, kde se později provádí potřebně a okamžité analýzy za užití OLAP. DWH jako takový slouží pouze ke čtení a analyzování dat. Do DWH se zapisuje pouze při pravidelné údržbě v podobě aktualizací a mazání neaktuálních záznamů. Během této údržby nelze vytvářet požadavky na analýzu a dotazy. Tyto rozdíly značně ovlivňují jak samotnou datovou implementaci, tak návrh a tvorbu konceptuálních modelů pro optimalizaci zpracování dotazů směřovaných na OLAP vrstvu DWH. Základní podstatou OLAP je co nejrychlejší poskytnutí požadované agregace daných dat. Pro naplnění DWH se používá proces ETL (Extract-Transform-Load). Jak již název napovídá proces je rozdělen na tři fáze. První extrakce dat z primárních datových zdrojů, následuje transformace dat do unifikovaných datových typů a převod na datový model, nad kterým lze vytvářet požadované agregace. V třetí a poslední fázi se agregovaná data uloží (Vítek 2002, Schiller 2003).

Klasické zpracování DWH pracuje s objemy dat ve velikosti, která málo kdy přesáhne řád TB (TeraByte) zdrojem takovýchto dat jsou zejména CRM (Customer Relationship Management), ERP(Enterprise Resource Planning) a finanční aplikace. Tento přístup v případě Big Data není zcela vyhovující. Zde již periodické aktualizace v podobě ETL nevyhovují zejména proto, že je zapotřebí zvládat obrovské objemy dat okamžitě a přizpůsobovat se toku. Zajistit potřebný vstup a výstup operací za vteřinu k včasnému doručení dat analytickým nástrojům.

Kvůli velkému objemu a rozsahu dat většina společností využívá tzv. „Hyperscale compitng environment“. Tato prostředí využívají úložišť, která jsou schopná velmi rychle a účinně expandovat pro pojetí obrovského toku dat, jako například z webu, tak i databázových systému a analýz, high-performance computingu a dalších systémů a aplikací (Villars 2011).

Další možnou technologií skladování je Hadoop. Hadoop je open source software Framework pro distribuovaná uložiště a zpracování velmi velkého množství dat. Hadoop se skládá z několika samostatných počítačů sestavených z běžného snadno dostupného HW. Všechny moduly v Hadoop jsou sestaveny a navrženy tak, aby počítaly s fundamentálním předpokladem selhání HW. Takovéto selhání by pak měl Framework automaticky začít řešit. Jádr Hadoop se skládá ze tří částí. První část je datové úložiště druhá je sjednocený souborový systém pro všechny moduly, a to Hadoop Distributed File

Systém (HDFS) a poslední částí je MapReduce, jinak také znám jako MapRed. MapRed je programovací model umožňující rychlé a paralelní zpracování a generování velkého objemu dat (IBM 2017, SAS 2016,White 2012).

Hadoop kromě nezměrných výhod v podobě velkého odolného úložiště schopného zpracovávat data, má také řadu nevýhod. MapRed ačkoli schopný zpracovávat velké množství dat rychle a efektivně, se příliš nehodí na řešení složitějších problémů a využívá se převážně pro jednoduché požadavky a problémy. Dalším problémem je zajistit dostatečné zabezpečení dat. A zcela zásadní problém je získat kvalitní zaměstnance, kteří dokážou s Hadoopem efektivně pomocí Javy a MapRed pracovat (SAS 2016).

3.1.5 Zdroje dat

Mezi zdroji dat a technologiemi skladování je významný rozdíl. Technologií je myšleno jakými technickými prostředky se data ukládají za použití hardwaru, softwaru a celkové struktury. Zdrojem dat je pak myšleno, odkud k nám data přicházejí nikoli to, jak jsou skladovaná, ačkoli se toto nemusí nutně vylučovat, jako například u DWH.

Zdroje dat se dělí do dvou kategorií interní a externí. Interní zdroje si vytváří firma sama svou vlastní aktivitou. A to jak formou záznamů o zákaznících, firemním webem, tak různými telefonními či poštovními kampaněmi. Externí data pochází pak od různých firem s velkými databázemi zákazníků, jako například úvěrové společnosti (Marr 2016, Rud 2002).

Interní zdroje

Jedná se o data, která si uchovává přímo jedna firma nebo organizace na základě vlastní činnosti. Tato data mají pro firmu největší prediktivní hodnotu, neboť se jedná o data přímo spojená s jejich výrobky, službami a zákazníky. Typickým zdrojem takových dat jsou databáze zákazníků, transakcí, historie nabídek nebo již zmiňované datové sklady (Hylbak 2014, Rud 2002).

Zákaznická databáze

Tato databáze většinou obsahuje jeden záznam pro každého zákazníka, v některých firmách může toto být jediná databáze. To znamená, že může obsahovat veškeré záznamy o prodejích a aktivitách pro všechny zákazníky. Tyto informace následně poskytují náhled na návyky zákazník a mohou do jisté míry tvořit vzor chování pro modelování prediktivních modelů. Častějším případem však je, že zákaznická databáze obsahuje pouze

aktuální informace k danému zákazníkovi. Tyto údaje se pak propojují pomocí různých identifikátorů s jinými databázemi například transakčními a získává se tak snímek aktivity daného zákazníka.

Každá firma má svůj vlastní návrh takové databáze, toto jsou však nejčastější prvky, které by databáze mohla obsahovat:

ID zákazníka – jedná se o jedinečný alfanumerický či číselný kód jasně označující konkrétního zákazníka.

Jméno zákazníka – Jméno osoby, popřípadě firmy. Nejčastěji se dělí do více polí například „Jméno“ a „Příjmení“. Nicméně to není podmínkou a jedno pole může obsahovat celé jméno.

Podrobnosti o nabídce – Například datum, kdy byla nabídka podána, o jaký typ nabídky se jedná, odpovědný pracovník.

Hodnocení modelu – Rizikovost, pravděpodobnost odchodu, odezva na nabídky.

Dále může obsahovat adresu, telefonní číslo, emailovou adresu, různé demografické údaje například věk (mbasKool 2017, Marketing Teacher 2017, getbase 2016, Rud 2002).

Transakční databáze

Transakční databáze obsahují informace o aktivitě zákazníků. Jedná se o velmi obsáhlou databázi, neboť každá transakce má vlastní záznam, tedy řádek. Každý zákazník má tedy v takové databázi několik záznamů. Ve většině případů databáze obsahují veškeré transakce provedené zákazníky. To poskytuje velmi kvalitní data pro prediktivní modelování, ale může být velmi problematické, tato data zužitkovat. Aby data byla využitelná je zapotřebí je napřed buď sumarizovat, nebo agregovat na úrovní zákazníka. Tato databáze má některé společné prvky se zákaznickou databází například: ID zákazníka nebo jméno zákazníka. Ale také vlastní jako je například:

Datum transakce – datum kdy byla transakce uskutečněna

Velikost transakce – například kolik peněz bylo odesláno či přijato

Typ transakce – zdali došlo k odeslání či přijetí peněz

Například databáze kreditních karet může také obsahovat, velikost poplatku za transakci, penále, ostatní poplatky, MCC, ID obchodníka (Rud 2002).

Databáze historie nabídek

Tato databáze se snaží pro každého zákazníka vytvořit unikátní záznam. Tento záznam obsahuje podrobnosti o nabídkách poskytnutých buď stávajícím, nebo také

potencionálním zákazníkům. Dále by v případě současných zákazníků databáze měla obsahovat informace o křížových prodejích, akvizičních nabídkách, speciálních nabídkách určeným pro udržení zákazníků a také je užitečné uchovávat bývalé adresy zákazníků (Rud 2002).

Datový sklad

Datový sklad jako zdroj dat je struktura, která spojuje alespoň dvě tyto databáze. Z těchto databází pak utváří jedno ucelené centrální úložiště, zde jsou data následně integrována, sumarizována, vyčištěna a následně distribuována do data martů. Data marty slouží k uchovávání podmnožin dat z centrálního úložiště, která jsou zde připravena pro koncové uživatele. Pro vytvoření kvalitního datového skladu je zapotřebí dostatek plánování a proškoleného personálu. To pak umožňuje snadný a efektivní přístup k interním datům (Rouse nedatováno, Rud 2002).

Externí zdroje

Mezi tyto zdroje lze zařadit prodejce a kompilátory seznamů. Prodejci seznamů jsou firmy, které ke své hlavní činnosti, kterou může být prodej prostřednictvím katalogů a časopisů. Prodej seznamů osob pak bývá jejich druhou činností. Tyto seznamy obsahují jména, adresy, kontaktní údaje a také demografická, behaviorální a psychografická data. Prodej těchto seznam je zprostředkován pomocí různých brokerů a kompilátorů.

Kompilátoři jsou firmy, jejichž seznamy buď vycházejí z jednoho seznamu, ten může vycházet například z telefonního seznamu. Dále pak skupují jiné seznamy a tyto seznamy se slučují a kombinují, čímž získají doplňkové informace. Dále pak vytvářejí vlastní výzkumy průzkumy, aby své seznamy zdokonalily a tím zdokonalily i své služby (Rud 2002).

3.1.6 Správa Big Data (Nástroje a technologie)

Takto velké množství dat je nemožné zpracovat ručně. Je tedy zapotřebí využít adekvátních nástrojů. Krom kvalitního hardware jak pro uložení, tak dostatečně rychlé zpracování dat je také zapotřebí i kvalitní software. Poskytovatelů takového software je celá řada. Patří mezi ně například Oracle nebo SAS Institute. V této práci jsou využity, nástroje od společnosti SAS Institute běžně označovaného jako pouze SAS.

SAS

„SAS řeší opravdové problémy, se kterými se každodenně setkáváte. Kromě klasických otázek typu „Jak upevnit vztah se zákazníkem a podpořit jeho loajalitu?“, které řeší většina odvětví, jsou to též otázky typické pro specifická odvětví, např.: řízení rizik v pojišťovnách, boj s podvody ve finančních institucích, hledání cross-sellových příležitostí obchodníků apod.“(SAS 2016).

Statistical Analysis Software Institute je americká mezinárodní organizace, která má sídlo v Cary v Severní Karolíně (North Carolina). SAS Institute vyvíjí analytické softwarové nástroje. Tyto nástroje řeší jak samotnou analýzu dat, tak jejich kvalitu, vizualizaci nebo dolování dat, tedy data mining. Pro tuto práci byly použity nástroje SAS Enterprise Guide pro analýzu dat a SAS Visual Analytics pro následnou vizualizaci dat.

SAS Enterprise guide

Jedná se o klienta pro přístup do databáze a zároveň vývojové prostředí pro skriptovací jazyk SAS Base běžícím na operačním systému Microsoft Windows. Tento klient umožňuje jak samostatné psaní skriptů pro importování, exportování, spojování, zpracování, filtrování či analyzování dat, tak takzvanou „klikací“ alternativu, kdy si uživatel může z nabídky vybrat, jakou metodu chce použít a naklikat si jednotlivá pole v SASu označená jako „variables“ tedy proměnné, která budou zpracována a Enterprise Guide již potřebný kód vygeneruje sám. Enterprise Guide umožňuje grafický výstup jednoduchých grafů, zobrazení tabulek nebo také log, kde se zobrazují poznámky (note) obarvené modře, provedené příkazy, varovné hlášky (warning) označené „…“ nebo chyby (Error) obarvené červeně, které většinou zapříčiní ukončení běhu skriptu. Dále jsou v logu zeleným písmem zapsány výsledky příkazů tedy například, že byla vytvořena tabulka s X záznamy (observation) a Y proměnnými (variables) a jak dlouho vytvoření trvalo(SAS 2016).

SAS Visual Analytics

SAS Visual Analytics (běžně zkracováno jako VA) je webové prostředí, které podporuje běh aplikací pro přípravu či vizualizaci dat nebo také přípravu reportů, které jsou okamžitě dostupné pro prohlížení z jakéhokoli webového prohlížeče nebo mobilního zařízení, ať už se jedná o chytrý telefon nebo tablet. SAS VA nabízí analytikům a reportérům mnoho možností pro tvorbu reportů, jak po stránce grafické, tak funkční. Lze

zde velmi jednoduše vytvářet forecasty (předpovědi), rozhodovací stromy, různé statistické metody nebo prosté grafy.

SAS VA neslouží k psaní programů ani tvorbě dat, pouze pracuje s daty ze zdroje, ke kterému jsou připojeny, a spojuje pokročilé analytické funkce s intuitivním ovládáním. Tyto data lze ještě dodatečně pomocí data builderu zpracovat, například propojit dvě a více tabulek, nebo vytvořit nové kalkulované hodnoty na základě již známých dat. Toto umožňuje využití v širokém okruhu business řešení, které mohou používat jak manažeři, tak analytici (Aanderud 2015, SAS 2016).

SAS DataFlux (Data Management Studio)

DataFlux Data Management Studio je nástroj, který kombinuje kontrolu datové kvality, datovou integraci a správu dat. DataFlux Data Management Server umožňuje klientovi Data Management Studia spouštět aplikace, spouštět joby a služby běžící v reálném čase ve vysoce výkonném prostředí. Joby se nahrávají ze studia na server, kde jsou následně spouštěny. Tyto joby pak mohou například propojovat zákaznická, produkční a další data. Mohou také integrovat nebo zajišťovat jejich kvalitu.

DataFlux Data Management Servery poskytují škálovatelné serverové prostředí pro jednotlivé joby, profily, služby běžící v reálném čase. Poté, co jsou nahrány na server, je mohou autorizovaní uživatelé spouštět. Studio se využívá pouze k přípravě jednotlivých jobů popřípadě spouštění menších a jednodušších jobů.

Jedním z možných rozšíření je DataFlux Web Studio, které je zcela volitelné. Jedná se o webovou aplikaci, která používá vlastní licencované moduly, které umožňují provádět jednotlivé úlohy z webového prohlížeče. Všechny tyto joby jsou spouštěny na DataFlux Web Studio Serveru, který podporuje veškeré moduly pro Web Studio (SAS @2016).

SAS Enterprise Miner

SAS Enterprise Miner je řešení pro vytváření prediktivních a deskriptivních modelů nad velkým objemem dat skrze větší množství zdrojů dat v dané organizaci. Tento nástroj nabízí mnoho funkcí a vlastností pro business analýzu a modelování dat. Jedná se o velmi efektivní nástroj k odhalování podvodů, minimalizaci rizik, optimalizaci potřebných zdrojů, vylepšení užití času aktiv, kampaní a redukce úbytku zákazníků (PredictiveAnalyticsToday 2017, SAS 2016).

3.2 Práce s daty

Pro práci s takto velkým zdrojem dat i po vybrání vzorku je potřeba mnoho věcí. Krom již zmiňovaných technologií skladování, zdrojů dat a software pro zpracování je potřeba si práci řádně naplánovat a zvolit i správnou metodu pro následnou analýzu dat. A samozřejmě nejdůležitějším z bodů je vybrat vhodná data pro analýzu.

3.2.1 Plánování

Na začátku plánování je za potřebí položit si několik zásadních otázek. „Co chci vytvořit?“ „Jaký je cíl mé práce?“ „Jak k cíli dojdu?“ „Jaké nástroje a zdroje budu používat? “ „Jak poznám, že jsem byl úspěšný?“

Výsledek každého projektu závisí na dobře a srozumitelně definovaných cílech a souvislostech spojených s projektem, ať už se jedná o obchodní cíle nebo obecné využití projektu ku prospěchu organizace. Jako cíl můžeme považovat získání informací o zvycích klientů, úspěšnost produktů u určitých zákazníků a jak dané produkty vylepšit, proč dochází ke ztrátě zákazníků a tento úbytek regulovat (Rud 2002).

3.2.2 Výběr Techniky

V dnešní době existuje velké množství nástrojů a postupů pro prediktivní i deskriptivní modely. Jedná se například o lineární regresi, logistickou regresi nebo dále také neuronové sítě, genetické algoritmy, klasifikační stromy a regresní stromy (Rud 2002).

Lineární regrese

Prostá lineární regresní analýza je metoda založená na kvantifikaci dvou proměnných. Z nichž jedna je závislou proměnnou, tedy tu kterou se snažíme predikovat, jejíž průběh je spojitý a proměnnou nezávislou, prediktivní, podle níž se snažíme predikci provádět.

„V průběhu lineární regrese je hledána taková přímka procházející jednotlivými body, pro niž platí, že součet druhých mocnin odchylek od každého bodu je minimální“ (Rud 2002,s.10).

Někdy ovšem může vztah mezi proměnnými být nelineární, a tak je tedy potřeba transformovat nezávislou proměnnou tak, aby umožňovala najít lepší proložení. Klíčovým měřítkem pro lineární regresi je tzv. „R-kvadrát“, který měří celkovou variabilitu dat vysvětlenou daným modelem (Rud 2002).

Logistická regrese

Logistická regrese je velmi podobná té lineární. Zásadním rozdílem však je, že závislá proměnná je diskrétní či kategorická na rozdíl od lineární regrese kde je závislá proměnná spojitá, jak již bylo dříve zmíněno. Díky této vlastnosti je tato regrese velmi užitečná zejména v marketingu kde díky tomu můžeme sledovat diskrétní akce jako například odezva na nabídku, nesplácení závazků. Logistická regrese se dá využít k predikci výsledků dvou či více úrovní. Nicméně při analýze cílených modelů dochází převážně k dvouúrovňovému výsledku, aby tedy bylo možné použít regrese, je zapotřebí napřed převést závislou proměnnou na spojitou hodnotu, která využívá funkce pravděpodobnosti výskytu události (Rud 2002).

Neuronové sítě

Tato metoda se od předchozích značně liší. Na rozdíl od regresí není postavena na statistickém rozdělení, ale na funkci lidského mozku. Tyto sítě jsou tvořeny vrstvami, které jsou skládány jednotlivými uzly. Toto uspořádání se může u jednotlivých sítí lišit v závislosti na typu a složitosti sítě. Postu procesu probíhá dle následujícího schématu:
Data se rozdělí na testovací a trénovací, jednotlivé uzly na první vrstvě dostanou váhy. Následně jednotlivé uzly obdrží data a zpracují je, výsledky porovnají se skutečnými hodnotami a upraví se váhy. Data se opět rozdistribuují a zpracují. Tyto iterace probíhají, dokud se nedosáhne požadované minimální chybovosti (Rud 2002, Siegel 2014).

Genetické Algoritmy

Tyto algoritmy stejně jako již zmiňované neuronové sítě nevyžadují statistické rozdělení. To je dáno tím, že vychází, jak název napovídá, z evolučního procesu „přežití nejpřizpůsobivějšího“. Pod tím si lze představit mnoho věcí. Základem tedy je porovnávání několika modelů a jejich postupná eliminace. Tyto modely jsou upravovány a následně eliminovány v sérii několika iterací dokud se nenajde nejvhodnější model pro danou úlohu.

Modely se upravují za pomoci mutací, kombinací, klonováním a porovnáním. U této metody je jako u všech ostatních zapotřebí napřed určit cíl modelu, následně je potřeba určit míru hodnocení vhodnosti daného modelu pro naše potřeby. Následně se vybere několik modelů, které se hodnotí například na základě jejich schopnosti předpovídat zůstatky. Posléze je každému modelu přiřazena hodnota reprezentující váhu poukazující na schopnost předpovídat zůstatky ve srovnání s konkurenčními modely. Dále kromě testování jsou modely obměňovány a upravovány, například zmiňovanými mutacemi nebo párováním. Takto se hledá optimální model po několik generací.

Genetické algoritmy jsou velmi zdlouhavý a náročný proces a je tedy zapotřebí využívat k jejich zpracování výkonné počítače. Dnešní stroje jsou již velmi výkonné a spolehlivé, a tak se tato metoda stává velmi populární (Rud 2002, Siegel 2014).

Klasifikační stromy

Klasifikační strom nebo také rozhodovací strom, funguje na principu dělení dat, kdy dochází k sekvenčnímu dělení dat, tak aby došlo k co největším rozdílům v závislé proměnné. Tedy účelem tohoto stromu je data roztřídit do odlišných skupin či větví za účelem vytvoření co největší separace hodnot závislé proměnné.

Tato metoda je ze jména vhodná k identifikaci segmentů, které se chovají požadovaným způsobem například při modelování odezvy nebo také pokud se snažíme porozumět chování trhu.

Klasifikační strom se vytváří v sekvenci několika kroků a pravidel, které poskytují značnou flexibilitu. Pravidla, podle kterých se data dělí, mohou být založena na dělení dle určitých kritérií například: pohlaví, příjem, věk, počet dětí, rodinný stav. (Rud 2002)

Faktorová analýza rozptylu

Další z metod statistické analýzy je faktorová analýza rozptylu. Tato metoda se zaměřuje na vytváření nových proměnných a na snížení rozsahu dat tak, aby nedocházelo ke ztrátám informace. U této metody je kladen důraz na vzájemné souvislosti vstupních proměnných (Sebera 2012).

Faktorová analýza předpokládá, že každou ze vstupních proměnných lze vyjádřit jako lineární kombinaci nevelkého počtu společných faktorů, které jsou skryté a jediného chybového faktoru. Snaha této analýzy je vyjádřit závislost proměnných (Meloun, 2017).

Jedno-faktorová analýza rozptylu

Tato metoda je známá také jako například analýza rozptylu jednoduchého třídění nebo také pod anglickým názvem one-way ANOVA(ANalysis Of VAriance). Na rozdíl od faktorové analýzy rozptylu je zde snaha zkoumat účinek pouze jednoho faktoru na závislou proměnnou. Jedná se o zobecněnou analogii zkoumání rozdílu průměru mezi dvěma nezávislými skupinami pomocí nepárového t-testu.

V případě jedno-faktorové analýzy rozptylu jde o rozlišování rozdílu průměru mezi více skupinami, reprezentující jednotlivé kategorie sledovaného faktoru, pomocí výpočtu testovacího kritéria F. A analýza tedy odhaluje, zda skupiny vytvořené na základě kvalifikačního faktoru si jsou podobné nebo jestli jejich jednotlivé průměry tvoří shluky, které by bylo možné identifikovat (Dallal 2012, VFU 2017).

3.2.3 Základní dělení dat

Po stanovení cíle a výběru metodiky je dalším krokem jakékoliv analýzy důležité, jaká data pro analýzu zvolíme. Vhodná data se dělí do tří základních kategorií na demografická, behaviorální a psychografická. Toto dělení s sebou přináší i jisté množství výhod i nevýhod (Rud 2002).

Demografická data: „Demografická data obecně popisují charakteristiky osob a domácností“ (Rud 2002, s. 19). To znamená, že do tohoto typu dat se řadí pohlaví, věk, rodinný stav, příjem, vlastnictví domu či bytu, druh a kvalita obydlí, vlastnictví dopravního prostředku, národnost, počet dětí a příbuzných, dosažená úroveň vzdělání a také zaměstnání. Charakteristiky dat jako jsou rodinný stav, vzdělání, a typ obydlí, patří po většinu času mezi velmi vhodná pro prediktivní modely, neboť se mění velmi zřídka. Tato data jsou také poskytována za podstatně nižší cenu než jiné typy. Jejich nevýhodou je ovšem obtížné získávání, neboť většina lidí považuje tato data za intimní a odmítá je tedy poskytovat, pokud k tomu nejsou motivováni například nabídkou určitého produktu a v případě, že jsou data poskytnuta, nemůžeme si být 100% jisti jejich pravdivostí, neboť lidé záměrně poskytují falešné či zkreslené údaje (Rud 2002, Poláková 2010).

Behaviorální data: Tato data obecně poskytují největší prediktivní sílu. V závislosti na odvětví do nich spadají informace jako množství a typ nákupu, datum a

výše platby, chování při krachu a další. Jedním z možných typů behaviorálních dat jsou aktivity na webových serverech, a to ať už se jedná o zaznamenání prodeje, jednotlivá kliknutí nebo pohyb po stránce jako takoví. Tato data je poměrně složité získat z vnějšího zdroje a je to i velmi nákladné. Nicméně nám umožňují mnohem lépe předpovídat budoucí vývoj než jiné druhy dat (Rud 2002).

Psychografická (Attitudální) data: Charakteristika těchto dat vychází z názorů, životního stylu nebo také osobních hodnot. Tato data jsou spojována s výzkumem trhu. Získávají se za pomocí různých průzkumů, marketingových kampaní a zájmových skupin. Také je lze odvodit z nákupního chování zákazníků. Pro zlepšení analýz a cíleného modelování se tento typ dat integruje do zákaznických databází

Dále psychografická data umožňují firmám, které již z demografických a behaviorálních dat nemohou získat žádné nové informace, nový pohled do života zákazníků. Dokážou podle nich odhadnout životní úroveň a stupeň současného, popřípadě potenciálního zákazníka a reagovat na jeho potřeby, připravovat cílené produkty a služby. To vše v reakci na nějakou životní událost, ať už se jedná o svatbu, narození dítěte, odchod do důchodu, ukončení nebo počátek studia.

Psychografická data mají ovšem velkou nevýhodu. Tou je, že vyjadřují jisté zamýšlené chování, to samozřejmě v nejlepším případě může silně korelovat se skutečným chováním, ale také může docházet pouze k okrajové korelaci. To je dáno tím, že se data získávají pomocí různých šetření a zájmových skupin, poznatky z tohoto šetření jsou dále aplikovány na širší skupiny lidí pomocí různých statistických metod, jako je například segmentace. A zde vzniká problém a je za potřebí otestovat zdali ke korelaci skutečně dochází (Rud 2002).

3.2.4 Výběr dat pro modelování

Výběr vhodných dat je pro úspěšnost a kvalitu modelu zcela zásadní. Nástroje jsou sice také velmi důležitou součástí, dat jsou zcela stěžejní, tedy model je jen tak kvalitní a relevantní jako jsou jeho zdrojová data.

Data pro získávání nových zákazníků

Nejlepší data pro modely se zaměřením na získávání nových zákazníků vycházejí z některé předešlé kampaně. V tomto případě nezáleží, jestli předešlá kampaň se shoduje se současným produktem či službou.

Nejsou-li data z předchozí kampaně k dispozici. Je možné využít externí data k sestrojení modelu se vybírají na základě podobnosti aktuálního a cizího produktu na, který jsou data orientována v externím zdroji. Tento model se pak používá k výpočtu tendence zákazníků k nákupu tohoto typu produktu.

Data pro modely rizika

Modely pro řízení rizika jsou kritickou součástí v mnoha sférách například bankovnictví nebo pojišťovnictví, kde například vzniká riziko nesplácení úvěru či hypotéky a u pojišťovnictví je to pak nárokování klienta na uhrazení škod.

Existují silné vztahy mezi finančním a pojistným rizikem. Proto používají pojišťovací instituce modely finančního rizika jako podporu pro svoje modely pojistného rizika. Toto například poukazuje na zajímavý fakt, že povaha splácení kreditů je prediktivní vůči nárokům z pojištění automobilů (Rud 2002).

Modelování takového typu je velmi náročné zejména proto, že data je potřeba podpořit historickými daty za určité období, a tak se špatně ověřují. Dále jsou tato data silně ovlivňována silou ekonomiky a populačními trendy. Tato data je pak ovšem náročné a mnohdy i velmi nákladné sehnat (Rud 2002).

Vzorkování dat

Pro správný model je vždy zapotřebí obsáhnout maximální množství možného rozsahu osob. To ovšem může v některých případech velice nákladné časově náročné a současně i náročné na úložný prostor a tak se využívá takzvaného vzorkování. Vzorek ovšem musí být dostatečně veliký, jak pro samotné vytvoření modelu, tak pro jeho validaci. Dalo by se namítat, že v dnešní době jsou počítače natolik výkonné, že s problémem velikosti úložiště či rychlosti zpracovaní, není nezbytně nutné. Vzorkování je ovšem stále relevantní neb proces urychlí a šetří finance. A ve výsledku produkuje v podstatě identické výsledky se zanedbatelnou odchylkou.

Jak velký by měl vzorek být? Na tuto otázku neexistuje jednoduchá odpověď. Vše se odvíjí od několika faktorů, jako je například očekávaná míra návratnosti u cílové

skupiny. Mezi tyto faktory můžou patřit závislost na riziku, schválené zakázky, nebo míra rizika, jako je například neschopnost splácet.

Aby byl vzorek dostačující je zapotřebí, aby měl dostatečnou prediktivní hodnotu. Je tedy důležité mít dostatečný počet pozorování neboli hodnot. A to jak pro respondenty tak nerespondenty. Žádné minimum neexistuje v tomto ohledu je minimální počet pozorování velmi relativní, existuje však jednoduché empirické pravidlo, které říká, že by vzorek měl obsahovat alespoň 25 pozorování. Dále také platí, že čím větší počet pozorování v dané buňce či na dané úrovní máme, tím větší je predikční schopnost. Pro zvolení velikosti vzorku je tedy velmi důležité znát tuto schopnost. Pokud je vzorek příliš malý, bude velmi náročné zjistit predikční schopnost nikoli však nemožné. Tímto způsobem vznikají robustnější modely.

Ve většině takových případů vyhoví náhodné vzorkování. Je-li účelem nového modelu získání chování nových zákazníků, které by doposavadní model za normálních okolností nepodchytil. Je nasnadě zvolit náhodnou skupinu jmen, mimo běžný výběr. Zde dochází k zásadnímu problému, model, který se pohybuje mimo cílovou skupinu, nebude mít tedy stejnou výkonnost, a tak musí zákonitě dojít ke ztrátám.

Pro cílovou skupinu se ve většině případů používá přibližně 10% populace, zde je možná zachovat celou původní skupinu náhodný vzorek se pak vybírá z necílové skupiny. Velikost takového vzorku by se měla pohybovat kolem 50-75 tisíc záznamů. Vzorek lze samozřejmě i dále oříznout to ovšem zapříčiní složitější tvorbu modelu pro případ s větším množstvím proměnných (PQ Systems 2016, Rouse 2014, Rud 2002, UTDallas 2017).

Klasifikace dat

Data se dělí do dvou základních tříd, a to na kvalitativní a kvantitativní. Kvalitativní se data v proměnných odlišují popisnými pojmy, jako je například označení pohlaví muž, je označen jako M (z anglického Male), žena potom jako F (z anglického Female). Stejný zápis by v kvantitativních datech mohl vypadat následovně: číslem 1 pro muže a číslem 2 pro ženy. Kvantitativní data totiž používají číselné označení. Tato data se dále dělí do několika kategorií (Rouse 2007, Rud 2002).

Nominální data

Jedná se o číselná data reprezentující atributy. Tato data mají zásadní vlastnost, a to takovou, že nezáleží na velikosti hodnoty. Tedy pokud by v předchozím případě byl muž označen jako 1 a žena jako 5 neznamená to, že by žena byla 5x lepší nebo naopak až na 5.

místě. Pouze to, že pro označení používáme tato čísla. Pro nominální hodnoty s pouze dvěma hodnotami by se měla používat čísla 0 a 1(Rud 2002).

Ordinální data

Tato data mají jistý relativní význam. Určují pořadí nebo důležitost. Například pro hodnocení finančního rizika označíme číslem 1 subjekt, který vždy splácí včas bez jediné výjimky, 2 se pak zpozdí výjimečně, 4 splácí zásadně pozdě a 5 splácí výjimečně. Význam je relativní, neboť tyto hodnoty můžeme otočit a hodnotit body tedy čím více bodů, tím lépe. Ale opět to neznamená, že subjekt s ohodnocením 1 je 5x lepší než subjekt s ohodnocením 5. Tento rozdíl může být markantnější, ale také ovšem podstatně menší vše záleží na tom, jak je hodnocení navrhnuto (Rud 2002).

Intervalová data

V tomto případě se jedná o data, která mají také relativní význam a nemají nulový bod. Ale už tentokrát se jedná o data, kde hodnota značí jistou míru a operace sčítání a odčítání získávají význam. Na rozdíl od předešlého případu, kde byla použita stupnice 1-5 je zde možné použít interval 300-1000 a následně určovat rozdíl rizikovosti na základě výpočtu (Rud 2002).

Spojitá data

Do těchto dat spadají data, jako jsou tržby, zůstatky na účtech, výše splátek a podobné. Lze tedy nad těmito daty provádět veškeré aritmetické operace, a tak získávat potřebné informace pro prediktivní modelování (Rud 2002).

3.2.5 Příprava dat

Příprava dat je dalším velmi nezbytným krokem pro úspěšné modelování. Je nepodstatné jestli se jedná a jednoduchou či velmi komplexní analýzu, jejíž výsledek bude vždy ovlivněn datovou kvalitou. Výběr dat je tedy stejně důležitý jako výběr užité metody. Zde je zapotřebí se důkladně seznámit se strukturou dat. Následně je zapotřebí v datech vyhledat chyby a tyto chyby ošetřit, ať už se jedná o chybné hodnoty, chybějící hodnoty nebo extrémní hodnoty vybočující z odůvodnitelných mezí popřípadě chybějící data po napojení více zdrojů (DataWatch , Rouse 2016, Rud 2002).

Čištění dat

Čištění dat je jedním z nejdůležitějších kroků, neboť na něm záleží kvalita finálního modelu. Čištění spojitých proměnných v datech se provede například pomocí nástrojů jako je SAS Enterprise Guide, kde použijeme proceduru PROC UNIVARIATE, ta nám následně odhalí informace o proměnné, jako například průměrnou hodnotu, medián, nejvyšší a nejnižší hodnotu. Jednoduchý příklad: pokud je medián 65 tisíc a maximální hodnota je pak 700 tisíc, může se jednat o chybu, protože se hodnota pochybuje o řád výš. Tato chyba mohla vzniknout přidáním 0 již při průzkumu, kdy ji zadával respondent, nebo když byla zadávána do systému. Po ověření lze chybu opravit, pokud se nepovede pravdivost informace ověřit je dobré celý záznam s touto chybou odstranit ze vzorku (Rouse 2016, Rud 2002).

Chyby dat a chybějící hodnoty

Nejlepším způsobem jak chyby u diskrétních, které mohou být způsobeny například překlepem nebo opomenutím, je zjištění četnosti jednotlivých hodnot. Většina hodnot pro diskrétní proměnné bude omezená například v případě pohlaví, kdy jsme omezeni dvěma hodnotami. Pro zjištění četnosti nebo také frekvence je možné použít mnoho způsobů v závislosti na softwaru, který je pro analýzu využit. V případě SASu je jednou z možností příkaz: „proc freq“.

```
proc freq data=example_1;
bbbntable answer /missing;
run;
```

Tabulka 1: Ukázka PROC FREQ

Tento příkaz nám odhalí seznam hodnot dané proměnné, jejich četnost a procentuální výskyt. Pokud máme například hodnoty viz tabulka, je možné odhalit, že hodnota D je nesprávná a tedy ji buď lze odstranit, nebo nahradit. Nahrazení je nejvhodnější provést za nejčetnější hodnotu tedy C. S chybějícími hodnotami lze na základě jejich četnosti naložit dvěma způsoby. Prvním v případě viz tabulka, kdy má chybějící hodnota vysokou četnost lze s ní nakládat jako s další kategorií. V případě, že by četnost byla nízká, lze s ní naložit jako s chybovou hodnotou D a buď ji do modelu nezahrnout, nebo ji nahradit jinou hodnotu (Microsoft 2017, Rud 2002).

Odpověď	Četnost	Procento	Součet četnosti	Součet Procent
	2821	11.28	2821	11.28
A	12999	52.00	15820	63.28
B	4044	16.18	19864	79.46
C	5135	20.54	24999	100.00
D	1	0.00	25000	100.00

Tabulka 2: Kategorická chyba v datech

(Rud 2002)

3.3 Analýza rizika

K nejčastějšímu využití analýzy rizika dochází v bankách například při hodnocení úvěru tzv. credit scoring. Vždy když jde klient zažádat o úvěr, je mu předloženo několik otázek jako například: „Bydlíte v domě nebo bytě?“ „Ve vlastním nebo v pronájmu?“ „Jak dlouho žijete na současné adrese?“ „Jste v zaměstnaneckém poměru, podnikáte nebo nezaměstnaný?“ „Jste ženatý/vdaná?“ „Máte děti?“ a tak dále. Odpovědi na tyto otázky jsou využity k následnému vypočtení hodnocení úvěru. Každá odpověď je bodově ohodnocena, hodnoty všech odpovědí jsou následně sečteny a převedeny na celkové hodnocení úvěru (Rud 2002).

Takovýto způsob hodnocení se objevil poprvé v šedesátých letech. Kdy firma Fair, Isaac and company vyvinula první hodnotící algoritmus založený na několika klíčových faktorech. Z počátku byly ostatní společnosti skeptické a používali doposavadní metodu, tedy o udělení úvěru rozhodoval pracovník banky, u kterého bylo o úvěr zažádáno. S postupem času se prokázalo, že výsledky algoritmu přibližně odpovídají skutečnosti a došlo k přechodu (Rud 2002).

S příchodem nových výkonnějších technologií s lepšími výpočetními schopnostmi se tyto algoritmy zdokonalovaly a komplikovaly. To vedlo k lepšímu hodnocení rizik a lepším výsledkům. Hodnotící algoritmus, který se stal standardem je stále ve vlastnictví firmy Fair, Isaac and company a tato společnost si jej přísně střeží. Nicméně zveřejnila některé z faktorů:

Historie minulých plateb

- Informace o splácení účtů určitých typů (např. kreditních karet).

- Údaje o nepříznivých veřejných záznamech. (např. bankrot, soudních procesech) inkasech a případné nesplacené dluhy.
- Závažnost případných neplacených dluhů.
- Zůstatek těchto nesplacených dluhů a inkas.
- Doba, po kterou nejsou dluhy a inkasa spláceny; Doba, která uplynula od zápisu podobně nepříznivých záznamů do úvěrové historie.
- Počet nesplacených dluhů.
- Počet dluhů splacených dle dlouhodobých podmínek.

Zůstatky úvěrů

- Výše zůstatků na úvěrových účtech.
- Výše zůstatků na určitých typech účtů.
- V některých případech nedostatek hotovosti na některých účtech.
- Počet účtů se zůstatky.
- Využití úvěrových limit (poměr výše zůstatků k celkové výši úvěrových limitů u určitých typů obnovujících se účtů).
- Využití účtů splátkového prodeje (poměr výše zůstatků k původní výši úvěru u určitých typů účtů splátkového prodeje).

Délka úvěrové historie

- Doba od otevření účtů.
- Doba od otevření určitých typů účtů.
- Celková doba, po kterou je zaznamenána nějaká aktivita na libovolných účtech.

Hledání a nabytí nového úvěru

- Počet účtů, otevřených v daném časovém okamžiku a jejich typy.
- Počet žádostí o úvěr, podaných v daném časovém okamžiku.
- Doba, která uplynula od otevření aktivních účtů a jejich typy.
- Doba, která uplynula od podání žádosti o úvěr.
- Znovuobnovení pozitivní úvěrové historie po předcházejících problémech.

Typy otevřených úvěrů

- Počet různých typů účtů (úvěrových karet, účtů splátkového prodeje, hypoték apod.)

Firmy Fair, Isaac and company není jedinou firmou, která se zabývá vývojem algoritmu hodnocení rizika. Existuje mnoho společností, které se touto problematikou zabývají. Některé vyvíjejí tyto algoritmy čistě pro svou potřebu, jiné pak pro jejich prodej (Rud 2002).

3.3.1 **BASEL**

„Basel III představuje pravidla regulace bankovnictví vydaná Basilejským výborem pro bankovní dohled“ (Management mania, 2016).

Tato pravidla vztahující se ke kapitálové přiměřenosti neboli schopnosti absorbovat riziko mají zajistit stabilitu bankovního sektoru. V případě, že banku postihne dočasné nepříznivé období, je nutné, aby byla dostatečně kapitálově vybavena. Pokud by došlo ke ztrátě je zapotřebí, aby banka měla dostatek kapitálu k absorpci ztráty a zamezení krachu. Úprava kapitálové přiměřenosti podle BASEL II požaduje minimální míru 8% kapitálu vzhledem k objemu aktiv a riziku banky. Banky by měly postupně přecházet na model BASEL III, který pravidla dále upravuje.

Pro určení míry rizika se jako jedna z metod používá pravděpodobnost selhání z anglického, probability of default (PD). Jedná se o stupeň pravděpodobnosti, že dluh nebude splacen včas (Financial Times, 2017).

PD je jedním z rizikových prvků, který se používá u takzvaného IRB přístupu. Tento přístup je založen na interním ratingovém systému. Banka si tento systém vypracuje sama na základě zásad, které jí stanoví regulátor. IRB přístup třídí expozici do šesti kategorií a to následovně:

- Podniky
- Suveréni (státy, jejich centrální banka, veřejnoprávní instituce a multilaterální banky)
- Banky a ostatní finanční instituce
- Drobná klientela
- Specializované úvěrové expozice
- Akcie a majetkové účasti

Mezi další rizikové prvky, se kterými IRB přístup pracuje, patří:

- Expozice při selhání (Exposure at Default, EAD) - celkové množství aktiv, které jsou vystaveny riziku v případě, že dlužník nedostojí svým závazkům.
- Míra ztráty při selhání (Loss Given Default, LGD) - podíl aktiv ztracených v případě, že nastane selhání, vyjadřuje se v %. Tedy 1 - výtěžnost. Výtěžnost (recovery rate) představuje podíl navrácené částky z celkové expozice, pokud nastane případ, kdy dlužník neplní či přestal plnit závazek.
- Doba splatnosti (Maturity, M) - zpravidla nominální doba splatnosti, která se měří v letech. IRB přístup je (na rozdíl od standardního) dvou dimenzionální, bere v úvahu jak dlužníka, tak transakci. Zatímco pravděpodobnost selhání se vztahuje k dlužníkovi, ostatní rizikové prvky se vztahují k příslušné transakci.

IRB přístup je tedy založen na podstatě přidělení interního ratingu ve chvíli, kdy dojde k posouzení charakteristik, myšleno dlužníka i transakce, a zařazení do příslušné kategorie. Následně je také odhadnuta PD s dalšími rizikovými faktory. PD je tedy důležitým prvkem funkce rizikové váhy neboli zmiňovaného minimálního kapitálového požadavku (Investopedia 2017, Kadlčáková 2002, Segoviano 2002, BIS 2001).

Zavedení IRB přístupu má několik předpokladů a zároveň požadavků. Banka musí splnit řadu požadavků, aby bylo možné zavést IRB, zde je kategorický výčet minimálních požadavků, které je nutné splnit a demonstrovat před implementací regulátorovi. Požadavků je mnohem víc, výčet tedy není kompletní (Investopedia 2017, Kadlčáková 2002, Segoviano 2002, BIS 2001).

Interní rating musí být organizovaný a řízený. Tyto funkce plní představenstvo a vyšší management dané banky, který systém musí schválit. Systém je navržen, zaveden a provozován nezávislým útvarem řízení úvěrového rizika. Následně představenstvo a vyšší management dostává měsíční reporty o interním ratingu, tedy například rizikový profil jednotlivých stupňů nebo porovnání poměru skutečných a očekávaných selhání, dále případné ztráty migrace mezi stupni (Investopedia 2017, Kadlčáková 2002, Segoviano 2002, BIS 2001).

Dalším požadavkem je přidělování ratingových stupňů. Minimální počet ratingových stupňů by se měl pohybovat mezi šesti až devíti stupni pro splácené úvěry a dva stupně pro ty nesplácené. Žádný ze stupňů by přesahovat limit 30% dlužníků a nemělo by tedy docházet k jejich kumulaci na daném stupni. Dlužník vždy obdrží rating, který musí být

přezkoumaný nezávislým útvarem, dříve než mu je poskytnut úvěr. Aktualizace ratingu by měla probíhat v intervalech, a to buď po třiceti dnech u slabších klientů, kde by mělo docházet k aktualizacím téměř okamžitě po získání zásadních informací, 60 dnech nebo 90 dnech u běžných a silných klientů. Pro selhání musí existovat jednotná definice, Basel zahrnuje nejen případy, kdy ke splácení nedochází, ale i případy, kdy je pravděpodobné, že jistina či úrok dlužníkem splacena nebude (Investopedia 2017, Kadlčáková 2002, Segoviano 2002, BIS 2001).

Data smějí pocházet ze tří zdrojů. Prvním zdrojem jsou zdroje interní, tedy ty, které si instituce sama vytvoří. Druhým jsou pak data externí, to znamená například data od externích agentur, kde jsou přijata s příslušnou škálou a následně jí jsou přeškolovány na interní stupně a PD. Posledním zdrojem jsou pak data sdružená mezi institucemi. Data musí také umožňovat následující operace, zařazení dlužníků do stupňů, odhady ztrát spojené s každým stupněm, odhady migrace dlužníků mezi stupni v průběhu času, retrospektivní realokaci. Data se musí pohybovat v časovém řádu alespoň pěti let a každý stupeň má přiřazen odhad jednoletého PD. Dlužníkova historie stupňů PD se musí uchovávat (Investopedia 2017, Kadlčáková 2002, Segoviano 2002, BIS 2001).

Systém musí být řádně validován a otestován. K tomu slouží stresové neboli zátěžové testování, které musí být prováděno minimálně každých šest měsíců. Princip tohoto testování spočívá ve vyhodnocení dopadu větší migrace dlužníků do nižších ratingových stupňů. Dále se musí minimálně jednou ročně provést validace používaného modelu, kdy se porovnává skutečná míra selhání s očekáváním podle PD.

Po zavedení IRB stává se komplexní součástí řízení úvěrového rizika a využívá se ke stanovení cen úvěrových produktů, přiřazení interních limitů úvěrové angažovanosti a úvěrové pravomoci, tvorbu rezerv a opravných položek, analýzu rentability banky, tedy jako nástroj strategické alokace zdrojů (Investopedia 2017, Kadlčáková 2002, Segoviano 2002, BIS 2001).

Posledním minimálním požadavkem je kolaterál, to znamená finanční zajištění. V tomto případě je možné použít nejen finančních prostředků jako u standardních metod, ale také určitý fyzický kolaterál v podobě komerčních a rezidenčních nemovitostí. Řízením tohoto zajištění musí být pověřen samostatný útvar. Při rozšířené jinak též zdokonalené metodě je krom PD potřeba stanovit i vlastní odhady LGD, EAD a M a musí splnit doplňkové požadavky. Znamená to zejména, že krom stupňů PD musí mít ještě explicitní

škálu LGD. Pro určení LGD stupně je třeba uvážit další faktory, které se vztahují nejen k dlužníkovi, ale i k transakci včetně jistiny. Minimální časový interval pozorování pro odhad LGD a EAD musí být alespoň sedm let, měl by tedy pokrýt úplný ekonomický cyklus. Pro ručitele jsou přidělovány stejně přísné minimální požadavky (Investopedia 2017, Kadlčáková 2002, Segoviano 2002, BIS 2001).

3.4 Merchant Category Code (MCC)

Merchant Category Code tedy MCC někdy značený jako Merchant Category Classification je čtyřmístné číslo označující obchodní a finanční služby podle ISO 18245. MCC jsou používány ke klasifikaci obchodů a poskytovatelů služeb podle typu nabízeného zboží či služeb. MCC jsou přiřazovány podle typu obchodu nebo názvu viz níže. Obchodník obdrží MCC od kreditní společnosti ve chvíli, kdy začne přijímat kreditní karty jako formu platby. Tento kód pak odráží kategorii, v jaké se obchod nachází. Například veterinární klinika obdrží MCC 0742 – Veterinární služby (Veterinary Services) železářství by dostalo MCC 5072 – Železářské vybavení a zásoby (Hardware equipment and supplies) MCC je následně využito kreditními společnostmi pro poskytování cashback odměn svým zákazníkům za nákupy v určitých kategoriích. Dále je v USA možnost tyto kódy používat pro vyplňování daňového přiznání (CitiBank 2014, Dwyer nedatováno, Visa 2017).

V této práci jsou MCC použity pro agregaci dat a zkoumání rizikovosti klienta na základě jeho transakcí. MCC se dělí do několika kategorií podle toho, jaké služby obchodník poskytuje. Každá s kategorií MCC má svůj stanovený interval, ve kterém jsou další podkategorie. A tedy každý obchodník nebo poskytovatel služeb spadající do této kategorie dostane příslušné číslo. Například zmiňovaná veterinární klinika spadá do intervalu 0001-1499 zemědělské služby kde mají hodnotu 0742 právě služby veterinární. U železářství je to pak interval 5000-5599 maloobchody s podkategorií 5072 Železářské vybavení a zásoby (CitiBank 2014, Dwyer 2004, Visa 2017).

Pokud by se poskytovatel služeb nebo obchodník pohyboval ve více kategoriích, například by obchodoval s počítači a elektronickými součástkami, tedy pod kategoriemi 5045 a 5065 byla by mu přidělena pouze jedna kategorie, a to na základě toho, která z kategorií má v jeho obchodě větší podíl (CitiBank 2014, Dwyer 2004, Visa 2017).

3.4.1 **Obecné MCC**

V případě obecných MCC je každé společnosti, organizaci či obchodníkovi poskytnuto kategorické MCC, toto MCC může sdílet s jinými společnostmi a organizacemi ve stejné kategorii. Nejedná se tedy o ekvivalent unikátního ID. Například řetězec Billa, Lidl a Tesco budou mít MCC stejné a to 5411 – Potraviny a supermarkety.

- 0001-1499 -> Zemědělské služby

  Do této kategorie spadají služby zabývající se zemědělstvím a příbuznými obory například zmiňované veterinářské služby. Dále sem mohou patřit zemědělská družstva nebo zahradnické a terénní služby.

- 1500-2999 -> Sjednané služby

  Do této Kategorie spadají například betonářské, elektrikářské a truhlářské služby. A další služby spjaté se stavbou a údržbou.

- 4000-4799 -> Doprava

  Tato kategorie zahrnuje většinu dopravních služeb, ať už se jedná o vlaky, autobusy, taxíky či letiště. Také sem spadají dopravní poplatky jako například mýtné. Co ovšem do této kategorie nepatří, jsou jednotlivé aerolinky, které mají kategorii vlastní.

- 4800-4999 -> Údržbářské služby

  Do této kategorie patří veškeré služby spjaté s údržbou nebo inženýrskými sítěmi, tedy například plyn, elektřina, voda, kabelová nebo satelitní televize, internet, telekomunikační služby a další.

- 5000-5599 -> Maloobchody

  V této kategorii se nachází téměř všechny obchody. Od obchodů prodávajících specifický druh zboží například obchod s elektrotechnikou, elektronikou, kancelářské vybavení, instalatérství, železářství nebo například obchody s drahými kameny po obchody s potravinami a supermarkety.

- 5600-5699 -> Obchody s oblečením

  Obchody s oblečením a módními doplňky nespadají do předchozí kategorie, ale do této samostatné kategorie.

- 5700-7299 -> Obchody se smíšeným zbožím

V této kategorii se nachází lékárny, knihkupectví, zastavárny, obchody se šperky, restaurace, bary, rychlé občerstvení, obchody se softwarem nebo s digitálním zbožím.

- 7300-7999 -> Obchodní služby

Do této kategorie spadají služby prováděné na zakázku či jiné obchodní služby, jako například konzultační, marketingové, úklidové služby. Také zde lze nalézt pracovní agentury a placená parkoviště či garáže.

- 8000-8999->Profesionální služby a členské organizace

Do této kategorie spadají služby spjaté s konkrétní profesí například lékařské, nemocniční, právnické či pečovatelské. Dále členské organizace a kluby například politická sdružení, strany a kluby. Dále jsou v této kategorii zařazeny vzdělávací instituty.

- 9000-9999 -> Státní služby

V poslední z obecných kategorií jsou služby a poplatky pod správou státu. Patří sem například platba daní, poštovného, pokut, soudní výdaje nebo také platba kauce.

(CitiBank 2014)

3.4.2 **Cestovní a Zábavní MCC**

Do cestovních a zábavních MCC spadají poskytovatelé konkrétních služeb a to Aerolinky, půjčovny vozidel a ubytovací zařízení. Na rozdíl od obecných MCC kdy je přiřazena jedna kategorie několika různým společnostem viz **Obecné MCC**, má zde každá společnost svůj vlastní MCC.

- 3000-3299 -> Aerolinky

Jak název napovídá, v této kategorii se nacházejí MCC jednotlivých Aerolinek z celého světa. Například 3217 ČSA nebo 3008 Lufthansa.

- 3300-3499 -> Půjčovny aut

Tento interval je rezervován pro jednotlivé půjčovny osobních automobilů.

- 3500-3999 -> ubytování

Poslední kategorií je ubytování, sem spadají hotely, ubytovny, country kluby, rezorty a lázně. Například Holiday Inn má MCC 3501 nebo Hilton, který má MCC 3504.

(CitiBank 2014)

4 Vlastní práce

4.1 Příprava analýzy

Jedná se o analýzu podobnosti MCC, které byly popsány v teoretické části práce, a následné analýzy rizikovosti klientů na základě MCC. Rizikovost klienta se posuzuje dle metody Basel, kdy se využívá probability of default (PD) neboli pravděpodobnost selhání rovněž popsáno v teoretické části. Prvním krokem přípravy bylo získání vzorku a jeho následné importování do SASu.

4.2 Vzorek

Vzorek dat pochází z transakční databáze ČSOB. Data ve vzorku jsou anonymizována a zašuměna, slouží tedy pouze k analýze a některá pole tedy neodpovídají skutečnosti například uid_party_id pouze reprezentuje klientské id a neodpovídá její skutečné hodnotě. Jedná se o historická data v intervalu od 1. 10. 2015 do 31. 12. 2015.

4.2.1 Analýza vzorku

Vzorek obsahuje dva miliony záznamů, konkrétně 2,077,142 záznamů. V těchto záznamech se nachází sto osmdesát tisíc záznamů (přesně 180,009) unikátních klientských ID, tedy sto osmdesát tisíc klientů. Z toho je tři tisíce (3009) klientů v defaultu, tedy selhali při splácení úvěru. To činí 1,62% vzorku. V tomto vzorku je zastoupeno všech 13 hlavních kategorií MCC.

Data ve vzorku byla agregována podle uid_party_id a spočteny hodnoty počtu transakcí a součtu objemu všech transakcí nad jednotlivými id.

Z tabulek je možné vyčíst, že pět nejméně aktivních klientů provedlo v daném intervalu pouze tři transakce kreditní kartou, zatímco pět nejaktivnějších provedlo 97, 100, 130, 131 a 159 transakcí. Průměrnou hodnotou počtu transakcí je přibližně 11,17. Střední hodnota pro počet transakcí je 10, nejčastější hodnota pak 7. Hodnota rozptylu je 45,09.

U objemu transakcí je průměrná hodnota 13112,25. Modus je roven 15000, medián 8765,46. Pět klientů, kteří měli v součtu nejmenší hodnoty objemu, se pohybují v intervalu od 54 korun do 140 korun, pět klientů s nejvyšším součtem se pohybuje v intervalu 554,404 do 2,871,325.

Moment			
N	186009	Součet vah	186009
Průměr	11.167336	Součet záznamů	2077142
Směrodatná odchylka	6.71502439	Rozptyl	45.0915525
Šikmost	1.88631821	Špičatost	8.34548723
Variační koeficient	60.1309425	Střední chyba průměru	0.01556971

Tabulka 3: Moment - Počet transakcí
(vlastní zpracování)

Základní statistická měřítka			
Poloha		Variabilita	
Průměr	11.16734	Směrodatná odchylka	6.71502
Medián	10.00000	Rozptyl	45.09155
Modus	7.00000	Rozpětí	156.00000
		Mezikvartilové rozpětí	8.00000

Tabulka 4: Počet transakcí
(vlastní zpracování)

Extrémní záznamy			
Nejnižší		Nejvyšší	
Hodnota	záznam	Hodnota	záznam
3	186004	97	78514
3	185957	100	124153
3	185953	130	171182
3	185947	131	166027
3	185935	159	166193

Tabulka 5: Extrémy - Počet transakcí
(vlastní zpracování)

Moment			
N	186009	Součet vah	186009
Průměr	13112.2515	Součet záznamů	2077142
Směrodatná odchylka	19429.0782	Rozptyl	377489080
Šikmost	25.267926	Špičatost	2692.78711
Variační koeficient	148.174997	Střední chyba průměru	45.0490065

Tabulka 6: Moment - Objem transakcí
(vlastní zpracování)

Základní statistická měřítka			
Poloha		Variabilita	
Průměr	13112.25	Směrodatná odchylka	19429
Medián	8765.46	Rozptyl	377489080

Základní statistická měřítka			
Poloha		Variabilita	
Modus	15000.00	Rozpětí	2871271
		Mezikvartilové rozpětí	9973

Tabulka 7: Objem transakcí
(vlastní zpracování)

Extrémní záznamy			
Nejnižší		Nejvyšší	
Hodnota	záznam	Hodnota	záznam
54.00	64512	554404	157811
92.77	50738	729410	46851
94.30	129446	764163	128697
139.00	166615	1156517	112558
140.00	163774	2871325	166193

Tabulka 8: Extrémy - Objem transakcí
(vlastní zpracování)

4.3 Analýza podobnosti MCC

MCC se pohybují v intervalu 0001-9999, všechna čísla však nejsou zastoupena, skutečný počet MCC je tedy v řádech stovek. To je ovšem pro běžnou analýzu příliš mnoho a je potřeba toto množství regulovat. První částí analýzy tedy bylo rozdělit jednotlivé MCC do podskupin dle intervalu, ve kterém se nacházely. V tomto případě je porovnávána kategorická proměnná s proměnnou spojitou.

```
data mcc_prep;
set bakalarka_output;
if mcc >= 1 and mcc<=1499 then Master_mcc=1;
if mcc >= 1500 and mcc<=2999 then Master_mcc=2;
if mcc >= 4000 and mcc<=4799 then Master_mcc=3;
if mcc >= 4800 and mcc<=4999 then Master_mcc=4;
if mcc >= 5000 and mcc<=5599 then Master_mcc=5;
if mcc >= 5600 and mcc<=5699 then Master_mcc=6;
if mcc >= 5700 and mcc<=7299 then Master_mcc=7;
if mcc >= 7300 and mcc<=7999 then Master_mcc=8;
if mcc >= 8000 and mcc<=8999 then Master_mcc=9;
if mcc >= 9000 and mcc<=9999 then Master_mcc=10;

if mcc >= 3000 and mcc<=3299 then Master_mcc=11;
if mcc >= 3300 and mcc<=3499 then Master_mcc=12;
if mcc >= 3500 and mcc<=3999 then Master_mcc=13;
run;
```

Tabulka 9: Rozdělení MCC
(vlastní zpracování)

4.3.1 **Parametrická analýza**

Poté byla nad výslednou tabulkou zavolána procedura one-way ANOVA, kde byla nastavena závislá proměnná txn_bc_am a deskriptivní proměnná MCC s požadovaným, Bartlettův a Scheffeho testem. ANOVA tedy porovnávala rozdíly mezi jednotlivými kategoriemi na základě objemu transakcí. Následující tabulky jsou jen výběrem z kompletního výpisu proc ANOVA, tabulka Scheffeho Test byla také vzhledem k počtu 156 záznamů redukována pouze na jednostranná porovnání kategorií, které lze považovat za shodné, tedy pokud se kategorie 13 a 12 shodují, je v tabulce uveden pouze vztah 13-11 a 11-13 je opomenut. Jedinou výjimkou je vztah 7-9, který není shodný, ale v tabulce se nachází, objasnění této výjimky je níže.

Z tabulky Scheffeho Testu lze vyčíst, že ke shodám mezi kategoriemi dochází a lze je tedy sjednotit do jedné podkategorie. Lze si povšimnout, že se v tabulce nenachází žádný vztah s kategorií 5, je tedy od všech ostatních kategorií významně odlišná. Všechny ostatní kategorie, lze v tabulce nalézt. Na první pohled by se tedy zdálo, že kategorie 1, 2, 4, 6, 7, 8, 9, 10 lze sjednotit. Tyto kategorie mají ovšem pouze jediný společný bod a tím je kategorie 2. Mezi kategoriemi ostatními kategoriemi však dochází k vzájemným rozdílům a nemůžou být tedy sloučeny. Kategoriím je přiřazeno vždy číslo, podle kategorie s nejnižší numerickou hodnotou. Nabízí se tedy tyto kombinace:

- 1,3,10 - Toto je jedna z ordinálně prvních kombinací, která se nám nabízí, s kategoriemi 1 a 10 dochází k podobnosti i u jiných kategorií, například 2 a 4, ale jsou to jediné dvě kategorie, které jsou současně dostatečně podobné kategorii 3 a sobě navzájem.
- 2,4,6,8 – Podíváme-li se do tabulky všechny tyto kategorie, jsou si podobné. Nicméně podíváme-li se na rozdíly průměrů mezi jednotlivými kategoriemi u kategorií 4,6 a 8 se pohybujeme řádově v desítkách a u rozdílu s kategorií 2 jsme vždy ve stovkách. I přesto ponechme tedy kombinaci 2,4,6,8.
- 11,12,13 - Tato kombinace je jasnou volbou, neboť jednotlivé kategorie jsou navzájem shodné.

Z tabulky ANOVA je však patrné, že se nulová hypotéze zamítá a přijímá se hypotéza alternativní, tedy, že kategorie jsou rozdílné. Tuto alternativní hypotézu také potvrzuje Bartlettův test homogenity, ze kterého lze vyčíst, že rozptyl vzorku není homogenní.

Výsledky Anovy a Scheffeho testu jsou tedy nespolehlivé a jednotlivé kategorie se musí posuzovat samostatně.

Zdroj	DF	Anova SS	Průměr^2	F	Pr > F
Master_mcc	12	957528228950	79794019079	9101.53	<.0001

Tabulka 10: ANOVA
(vlastní zpracování)

Bartlettův test homogenity rozptilu txn_bc_am			
Zdroj	DF	Chi^2	Pr > Chi^2
Master_mcc	12	1782650	<.0001

Tabulka 11: Bartlettův test homogenity
(vlastní zpracování)

Porovnání významnosti 0.05 úrovně jsou indikovány ***.				
Master_mcc porovnání	Rozdíl mezi průměry	Současný 95% limit jistoty		
13 - 11	491.430	-280.719	1263.578	
13 - 12	566.943	-922.817	2056.702	
11 - 12	75.513	-1323.134	1474.159	
7 - 9	574.505	447.207	701.804	***
2 – 7	-504.948	-1047.809	37.914	
2 - 9	69.557	-487.620	626.735	
2 - 6	430.459	-113.772	974.691	
2 - 4	454.344	-99.767	1008.455	
2 - 8	471.252	-74.729	1017.234	
2 - 10	540.456	-85.729	1166.640	
2 – 1	552.138	-55.448	1159.725	
6 – 4	23.885	-95.643	143.413	
6 - 8	40.793	-32.305	113.891	
6 - 10	109.996	-205.210	425.202	
6 – 1	121.679	-154.743	398.101	
4 - 8	16.908	-110.349	144.166	
4 - 10	86.111	-245.861	418.084	
4 – 1	97.794	-197.605	393.193	
8 - 10	69.203	-249.014	387.420	
8 – 1	80.886	-198.965	360.737	
10 - 1	11.683	-403.441	426.807	
10 – 3	86.217	-233.572	406.007	
1 – 3	74.535	-207.104	356.173	

Tabulka 12: Scheffeho Test
(vlastní zpracování)

4.3.2 Neparametrická analýza

Neparametrická analýza vychází z Kruskal-Wallisova testu a následného porovnání Wilcoxnovým ohodnocením. Tento test ukazuje, že testová statistická hodnota je rovna

442915,5687 jak je možno vyčíst z tabulky Kruskal-Wallisův test na řádce Chi^2. Stupeň volnosti pro tuto statistiku je 12. Řádek PR>Chi^2 značí, že se nulová hypotéza zamítá a existuje tedy významný rozdíl mezi všemi skupinami. Zamítnutí či přijetí nulové hypotézy pro jednotlivé porovnání skupin mezi sebou každá s každou bude provedeno pomoci Wilcoxonova ohodnocení a Dunnovi metody mnohonásobného porovnání.

Kruskal-Wallisův Test	
Chi-kvadrát	182138.6875
DF	12
Pr > Chi-kvadrát	<.0001

Tabulka 13: Kruskal-Wallisův Test
(vlastní zpracování)

Tabulka Wilcoxonova ohodnocení je základem pro další metodu, konkrétně Dunnovu metodu mnohonásobného srovnání. Tato metoda spočívá ve srovnání rozdílu mezi součty hodnocení jednotlivých skupin. Skupiny se porovnávají všechny se všemi a tak vznikne sedmdesát osm porovnání.

Wilcoxonovo ohodnocení (stupně sum) pro objem transakcí klasifikováno skupinami MCC					
Skupiny MCC	N	Součet hodnocení	Očekávání Pod H0	Směrodatná odchylka pod H0	Průměr hodnocení
5	1038788	9.08551E11	1.07886E12	432077694	874625.64
6	107302	1.2614E11	1.11441E11	191269203	1175563.34
7	807556	1.00198E12	8.38705E11	421253043	1240752.29
9	11538	1.29338E10	1.1983E10	64226530.9	1120978.08
8	50847	5.00199E10	5.28082E10	133539476	983732.75
3	39824	3.36841E10	4.13601E10	118502501	845823.28
1	2468	2717365407	2563194462	29769612	1101039.47
4	14666	1.57988E10	1.52317E10	72356234.4	1077238.90
10	1888	1870015498	1960822992	26041277.7	990474.31
2	626	742978258	650145759	14999627.4	1186866.23
13	429	696462919	445547174	12417739.7	1623456.69
11	1107	1951740409	1149698651	19944208.7	1763089.80
12	103	177494730	106972865	6085084.35	1723249.81

Tabulka 14: Wilcoxonovo ohodnocení
(vlastní zpracování)

Před ověřením či zamítnutím nulové hypotézy, je nejprve použita Bonferroniho korekce pro výpočet kritické hodnoty, která je užita v Dunnově metodě. Tato korekce se provádí tím způsobem, že se neporovnává přímo, hodnota alfa, ale upravená hodnotu alfa.

Úprava se provede vydělením hodnoty α=0,05 počtem testování v našem případě hodnotou třináct. Hodnota alfa je tedy rovna přibližně 0,0038.

Dunnova metoda mnohonásobného porovnání či srovnání je zde použita, neboť rozsahy jednotlivých výběrových souborů nejsou stejné. Pro zjištění, zda se výběry od sebe významně liší nebo jsou rozdíly nevýznamné, je tedy zapotřebí použít právě tuto metodu.

Následující tabulka reprezentuje srovnání všech kategorií s první kategorií z Wilcoxonova ohodnocení, tedy kategorií 5. Porovnáním absolutní hodnoty rozdílu Ti -> Součet hodnocení i-té v tomto případě páté kategorie a Ti -> součet hodnocení j-té kategorie, tedy současné komparované kategorie. A výsledku vzorce Dunnovy metody. Jestliže tedy platí |Ti-Tj| > Dunn viz tabulka. Hypotéza se zamítá. Z tabulky lze tedy vyčíst, že všechny hypotézy se zamítají a jednotlivé kategorie nelze sloučit.

Skupiny MCC	N	Součet hodnocení	kmpr. skupina	Ti-Tj	ABS Ti-Tj	Dunn	rozhodnutí
5	1038788	9,09E+11	6	7,82E+11	7,82411E+11	6171,017	Zamítá
5	1038788	9,09E+11	7	-9,3E+10	93429000000	2855,1	Zamítá
5	1038788	9,09E+11	9	8,96E+11	8,95617E+11	18015,57	Zamítá
5	1038788	9,09E+11	8	8,59E+11	8,58531E+11	8740,961	Zamítá
5	1038788	9,09E+11	3	8,75E+11	8,74867E+11	9826,786	Zamítá
5	1038788	9,09E+11	1	9,09E+11	9,08524E+11	38784,44	Zamítá
5	1038788	9,09E+11	4	8,93E+11	8,92752E+11	16003,07	Zamítá
5	1038788	9,09E+11	10	9,09E+11	9,08532E+11	44331,06	Zamítá
5	1038788	9,09E+11	2	9,09E+11	9,08544E+11	76941,11	Zamítá
5	1038788	9,09E+11	13	9,09E+11	9,08544E+11	92934,25	Zamítá
5	1038788	9,09E+11	11	9,09E+11	9,08531E+11	57872,5	Zamítá
5	1038788	9,09E+11	12	9,09E+11	9,08549E+11	189634,7	Zamítá

Tabulka 15: Dunnova metoda
(Vlastní zpracování)

5 Výsledky a diskuse

5.1 Podobnost skupin MCC

Prvním cílem bylo prověřit, zda jsou si jednotlivé skupiny MCC kódů dostatečně podobné a lze je tedy sjednotit do menšího množství podskupin.

5.1.1 Jedno faktorová analýza

Pro porovnání podobnosti hlavních skupin MCC byla použita. Jedno faktorová analýza rozptylu neboli one-way anova. Touto metodou bylo vzájemně porovnáno 13 hlavních kategorií (viz teoretická část) se součtem objemu transakcí v daných kategoriích. Scheffeho test poukazoval na jistou podobnost mezi některými z kategorií, ale Bartlettův test homogenity prokázal, že rozptyl vzorku není homogenní čili výsledky Scheffeho testu jsou nespolehlivé. Nulová hypotéza byla zamítnuta na základě F-testu a bylo tedy potvrzeno, že jednotlivé skupiny kategorií si nejsou podobny.

5.1.2 Neparametrická analýza

Neparametrická analýza pomocí Kruskal-Wallisova testu s Wilcoxonovým ohodnocením. Které následovala Dunnova metoda mnohonásobného porovnání prokázala, že opravdu mezi kategoriemi ke shodám nedochází. Tedy, že veškeré kategorie jsou významně rozdílné.

6 Závěr

Porovnání hlavních MCC kategorií tedy prokázalo, že rozdíly mezi všemi kategoriemi jsou významné a nelze je sloučit. To tedy neřeší problém analýzy rizikovosti kategorií a veškeré kategorie se musí hodnotit samostatně. Nelze tedy vytvořit sloučením podskupiny pro zjednodušení analýzy rizika a je tedy zapotřebí hodnotit všechny samostatně. Výsledky tedy mohou sloužit, jako podklad pro finální rizikovou analýzu klientů na základě kategorií, ve kterých provádějí transakce kreditní kartou. Doporučením pro zlepšení výsledků této analýzy by krom použití objemu transakcí nad kategorií, by byl pro srovnávání skupin použit i počet transakcí, které nad nimi byly klientem provedeny.

7 Seznam použitých zdrojů

Aanderud, T., 2015. What is the difference in SAS Visual Analytics and SAS Enterprise Guide? [online]. Bi-notes.com. [cit. 2017-01-23]. Dostupné z: http://bi-notes.com/2015/09/difference-sas-visual-analytics-enterprise-guide/

ABBOTT, D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. USA, NJ, Somerset: Wiley, 2014. ISBN 978-1-118-72793-5.

Adshead A., 2013. Big data storage: Defining big data and the type of storage it needs. [online]. Computerweekly.com. [cit. 2016-09-20]. Dostupné z: http://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs

Arthur L., 2013. What is Big Data? [online]. Forbes.com. [cit. 2016-09-15]. Dostupné z: https://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#7469b405c85b

BIS: Basel Committee on Banking Supervision, 2001 [online]. Bundesbank.de. [cit. 2017-02-02]. Dostupné z: https://www.bundesbank.de/Redaktion/EN/Downloads/Tasks/Banking_supervision/the_internal_ratings_based_approach.pdf?__blob=publicationFile

BIS: Working Papers, 2001 [online]. Eprints.lse.ac.uk. [cit. 2017-02-02]. Dostupné z: http://eprints.lse.ac.uk/24948/1/dp428.pdf

Citiban: Merchant category codes, 2014. [online]. Citibank.com. [cit. 2017-02-22]. Dostupné z: https://www.citibank.com/tts/card_solutions/commercial_cards/site/docs/dod/mcc_codes_0220.pdf

Dallal, J., 2001. Single Factor Analysis of Variance [online]. Jerrydallal.com. [cit. 2017-01-05]. Dostupné z: http://www.jerrydallal.com/lhsp/anova1.htm

Data mining solutions: Datové sklady OLAP, 2002 [online]. datamining.xf.cz . [cit. 2016-08-22]. Dostupné z: http://datamining.xf.cz/view.php?cisloclanku=2002102808

DataWatch: What is Data Preparation?, 2017 [online]. Datawatch.com . [cit. 2017-01-17]. Dostupné z: http://www.datawatch.com/what-is-data-preparation/

DOLÁK, Ondřej, 2011. Big data. [online]. Systemonline.cz. [cit. 2016-06-23]. Dostupné z: http://www.systemonline.cz/clanky/big-data.htm

Dwyer B., 2004. Merchant Category Code: Reporting & Rates. [online]. Cardfellow.com [cit. 2017-02-20]. Dostupné z: https://www.cardfellow.com/merchant-category-code-mcc/

Financial Times: Definition of probability of default, 2017 [online]. Lexicon.ft.com . [cit. 2017-01-30]. Dostupné z: http://lexicon.ft.com/Term?term=probability-of-default

Getbase: CREATING AND MAINTAINING A CUSTOMER DATABASE, 2016 [online]. Getbase.com [cit. 2016-12-05]. Dostupné Z: https://getbase.com/learn/customer-data-base/

Hylbak L., 2014. 4 Sources of Internal Data That Should Inform Your Marketing Strategy [online]. bizible.com [cit. 2016-11-18]. Dostupné z: http://www.bizible.com/blog/4-sources-of-internal-data-that-should-inform-your-marketing-strategy

IBM: What is big data?, 2017. [online] Ibm.com. [cit. 2016-04-20]. Dostupné z: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

IBM: What is MapReduce, 2017 [online]. Ibm.com. [cit. 2016-08-26]. Dostupné z: https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/

Investopedia: Advanced Internal Rating-Based - AIRB, 2017 [online]. Investopedia.com. [cit. 2017-02-20]. Doručeno z: http://www.investopedia.com/terms/a/airb.asp

Investopedia: Default Probability, 2017 [online]. Investopedia.com. [cit. 2017-02-20]. Doručeno z: http://www.investopedia.com/terms/d/defaultprobability.asp

IRS: Internal Revenue Bulletin, 2004 [online]. IRS.gov. [cit. 2017-02-25]. Dostupné z: https://www.irs.gov/irb/2004-31_IRB/ar17.html

Kadlčáková N., Sůvová H., 2002. Regulatorní a modelový přístup k úvěrovému riziku v bance. [online]. Cnb.cz. [cit. 2017-02-12]. Doručeno z: https://www.cnb.cz/cs/verejnost/pro_media/clanky_rozhovory/media_2002/cl_02_020321b.html

Management Mania: BASEL III, 2016 [online]. Managementmania.com. [cit. 2017-02-02]. Dostupné z: https://managementmania.com/cs/basel-iii

Marketing Teacher: Customer Database, 2017 [online]. Marketingteacher.com. [cit. 2017-01-03]. Dostupné z: http://www.marketingteacher.com/customer-database/

Marr B., 2016. Big Data: 33 Brilliant And Free Data Sources For 2016 [online]. forbes.com . [cit. 2016-11-02]. Dostupné z: https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#695f02d7b54d

Mbaskool: Customer Database, 2017 [online]. Mbaskool.com. [cit. 2017-01-03]. Dostupné Z: http://www.mbaskool.com/business-concepts/marketing-and-strategy-terms/1806-customer-database.html

McAfeE, A., Brynjolfsson, E., 2012. Big Data: The Management Revolution, Harvard Business Review [online]. hbr.org . [cit. 2016-08-22]. Dostupné z: https://hbr.org/2012/10/big-data-the-management-revolution

McKinsey&Company, 2016. How companies are using big data and analyitcs [online]. mckinsey.com. [cit. 2016-08-15]. Dostupné z: http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/how-companies-are-using-big-data-and-analytics

Meloun, M., 2017. Faktorová Analýza [online]. Meloun.upce.cz. [cit. 2017-01-05]. Dostupné z: https://meloun.upce.cz/docs/research/chemometrics/methodology/4dmetody.pdf

Microsoft: Zpracování chyb v datech, 2017 [online]. Microsoft.com. [cit. 2017-02-03]. Dostupné z: https://technet.microsoft.com/cs-cz/library/ms141679(v=sql.100).aspx

MINER, G., NISBER, R., ELDER IV, J.: Handbook of Statistical Analysis and Data Mining Applications. Academic Press, 2009. ISBN 978-0123747655.

Poláková Z., Klufová R. Demografické metody a analýzy: Demografie české a slovenské populace, 2010. ISBN 978-80-7357-546-5.

PQSystems: Sampling, 2016 [online]. Pqszstems.com. [cit. 2017-01-12]. Dostupné z: http://www.pqsystems.com/qualityadvisor/DataCollectionTools/sampling.php

Rouse M., 2004. Big Data [online]. Techtarget.com. [cit. 2016-09-20]. Dostupné z: http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

Rouse M., 2007. Data classification [online]. Techtarget.com. [cit. 2017-01-18]. Dostupné z: http://searchdatamanagement.techtarget.com/definition/data-classification

Rouse M., 2016. Data Preparation. [online]. Techtarget.com. [cit. 2017-01-20]. Dostupné z: http://searchbusinessanalytics.techtarget.com/definition/data-preparation

Rouse M., 2015. Data warehouses. [online]. Techtarget.com. [cit. 2016-09-20]. Dostupné z: http://searchsqlserver.techtarget.com/definition/data-warehouse

RUD, P., O.: Data Mining. Praha: Computer Press, 2002. ISBN 8072265776.

SAS: Big data: What it is and why it matters, 2016 [online]. SAS.com. [cit. 2016-06-25]. Dostupné z: http://www.sas.com/en_th/insights/big-data/what-is-big-data.html

SAS: DataFlux Data Management Studio, 2016 [online]. Support.SAS.com. [cit. 2017-01-28]. Dostupné z: http://support.sas.com/documentation/onlinedoc/dfdmstudio/2.7/dmpdmsug/dfUnity.html

SAS: SAS Enterprise Guide, 2016 [online]. SAS.com. [cit. 2017-01-28]. Dostupné z: https://www.sas.com/en_us/software/enterprise-guide.html

SAS: SAS Visual Analytics, 2016 [online]. SAS.com. [cit. 2017-01-28]. Dostupné z: https://www.sas.com/en_us/software/business-intelligence/visual-analytics.html

SAS: SAS PRODUCTS & SOLUTIONS, 2016 [online]. Support.SAS.com. [cit. 2017-01-28]. Dostupné z: http://support.sas.com/software/products/guide/

SAS: Společnost SAS, 2016 [online]. SAS.com. [cit. 2016-06-25]. Dostupné z: https://www.sas.com/cs_cz/company-information.html

Sebera M., 2012. Analýza hlavních komponent a Faktorová analýza [online]. Fsps.muni.cz. [2017-01-05]. Dostupné z: http://www.fsps.muni.cz/~sebera/vicerozmerna_statistika/pca.html

Shrimphood: Neuronové sítě – vývoj a testování, 2017 [online]. Shrimphood.net. [cit. 2017-01-05]. Dostupné z: http://www.shrimphood.net/neuronove-site-vyvoj-a-testovani.html

Schiller, M., 2003. Co se skrývá pod zkratkou ETL [online]. Systemonline.cz. [cit. 2016-08-22]. Dostupné z: https://www.systemonline.cz/clanky/co-se-skryva-pod-zkratkou-etl.htm

SIEGEL, E. Predictive Analytics. Hoboken: John Wiley & Sons, 2013. ISBN 978-1-118-35685-2.

STANČÍK, M., 2013. Big Data bez správné analýzy nejsou k ničemu [online]. Computerworld.cz. [cit. 2016-08-03]. Dostupné z: http://computerworld.cz/technologie/big-data-bez-spravne-analyzy-nejsou-knicemu-50285

Techtarget: Data sampling, 2014 [online]. Techtarget.com. [cit. 2017-01-15]. Dostupné z: http://searchbusinessanalytics.techtarget.com/definition/data-sampling

UT Dallas: Data collection and sampling, 2017 [online]. Utdallas.edu. [cit. 2017-01-12]. Dostupné z: https://www.utdallas.edu/~scniu/OPRE-6301/documents/Data_Collection_and_Sampling.pdf

Veterinární a farmaceutická univerzita: Analýza rozptylu (ANOVA) , 2017 [online]. Cti.vfu.cz. [cit. 2017-01-05]. Dostupné z: http://cit.vfu.cz/statpotr/potr/teorie/pred,n3/anova.htm

Villars, R. L, Eastwood, M., Olofson, C. W., 2011. Big Data What it is and Why you should Care [online]. Tracemyflows.com. [cit. 2016-08-26]. Dostupné z: http://www.tracemyflows.com/uploads/big_data/idc_amd_big_data_whitepaper.pdf

Visa: Visa Commercial Solutions, 2004 [online]. Web.archive.org. [cit. 2017-02-22]. Dostupné z: https://web.archive.org/web/20070710202209/http://usa.visa.com/download/corporate/resources/mcc_booklet.pdf

White, T. Hadoop: The Definitive Guide, 3rd Edition, 2012. ISBN: 978-1-4493-1152-0.

8 Přílohy

Příloha A: Výstupy proc Univariate pro počet transakcí

Tabulka 1A: Momenty – Počet transakcí

Tabulka 2A: Základní statistická měřítka – Počet transakcí

Tabulka 3A: Testy polohy – Počet transakcí

Tabulka 4A: Kvantily – Počet transakcí

Tabulka 5A: Extrémní záznamy – Počet transakcí

Příloha B: Výstup proc univariate pro objem transakcí

Tabulka 1B: Momenty – Objem transakcí

Tabulka 2B: Základní statistická měřítka – Objem transakcí

Tabulka 3B: Testy polohy – Objem transakcí

Tabulka 4B: Kvantily – Objem transakcí

Tabulka 5B: Extrémní záznamy – Objem transakcí

Příloha C: Výstup proc ANOVA

Tabulka 1C: Klasifikační úrovně

Tabulka 2C: ANOVA 1

Tabulka 3C: ANOVA 2

Tabulka 4C: ANOVA 3

Tabulka 5C: Bartlettův test homogenity

Tabulka 6C: Analýza skupin

Tabulka 7C: Srovnání skupin MCC

Příloha D: Výstup proc NPAR1WAY

Tabulka 1D: Wilcoxonovo ohodnocení

Tabulka 2D: Kruskal-Wallisův Test

Příloha E: Výsledná tabulka Dunnovy metody

Tabulka 1E: Donnova metoda mnohonásobného porovnání

Příloha A: Výstupy proc Univariate pro počet transakcí

Tabulka 1A: Momenty – Počet transakcí

Momenty			
N	186009	Součet Vah	186009
Průměr	11.167336	Součet záznamů	2077142
Směrodatná Odchylka	6.71502439	Rozptyl	45.0915525
Šikmost	1.88631821	Špičatost	8.34548723
neopravené SS	31584459	opravené SS	8387389.5
Koef. Variability	60.1309425	Směrodatná chyba průměru	0.01556971

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 2A: Základní statistická měřítka – Počet transakcí

Základní statistická měřítka			
Poloha		Variabilita	
Průměr	11.16734	Směrodatná odchylka	6.71502
Medián	10.00000	Rozptyl	45.09155
Modus	7.00000	Rozpětí	156.00000
		Mezikvartilové rozpětí	8.00000

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 3A: Testy polohy – Počet transakcí

Testy plohy: Mu0=0				
Test	Statistika		p Hodnota	
Studentovo t	t	717.2474	Pr > |t|	<.0001
Znak	M	93004.5	Pr >= |M|	<.0001
Značená hodnost	S	8.6499E9	Pr >= |S|	<.0001

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 4A: Kvantily – Počet transakcí

Kvantily (Definice 5)	
Úroveň	Kvantil
100% Max	159
99%	34
95%	24
90%	20
75% Q3	14

Kvantily (Definice 5)	
Úroveň	Kvantil
50% Median	10
25% Q1	6
10%	4
5%	4
1%	3
0% Min	3

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 5A: Extrémní záznamy – Počet transakcí

Extrémní záznamy			
Nejnižší		Nejvyšší	
Hodnota	Záznam	Hodnota	Záznam
3	186004	97	78514
3	185957	100	124153
3	185953	130	171182
3	185947	131	166027
3	185935	159	166193

Zdroj: ČSOB, SAS, vlastní zpracování

Příloha B: Výstup proc univariate pro objem transakcí

Tabulka 1B: Momenty – Objem transakcí

Momenty			
N	186009	Součet Vah	186009
Průměr	13112.2515	Součet záznamů	2438996784
Směrodatná Odchylka	19429.0782	Rozptyl	377489080
Šikmost	25.267926	Špičatost	2692.78711
neopravené SS	1.02197E14	opravené SS	7.0216E13
Koef. Variability	148.174997	Směrodatná chyba průměru	45.0490065

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 2B: Základní statistická měřítka – Objem transakcí

Základní statistická měřítka			
Poloha		Variabilita	
Průměr	13112.25	Směrodatná odchylka	19429
Medián	8765.46	Rozptyl	377489080

Základní statistická měřítka			
Poloha		Variabilita	
Modus	15000.00	Rozpětí	2871271
		Mezikvartilové rozpětí	9973

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 3B: Testy polohy – Objem transakcí

Testy plohy: Mu0=0				
Test	Statistika		p Hodnota	
Studentovo t	t	291.0664	Pr > |t|	<.0001
Znak	M	93004.5	Pr >= |M|	<.0001
Značená hodnost	S	8.6499E9	Pr >= |S|	<.0001

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 4B: Kvantily – Objem transakcí

Kvantily (Definice 5)	
Úroveň	Kvantil
100% Max	2871325.31
99%	82634.38
95%	35727.49
90%	25000.00
75% Q3	15000.00
50% Median	8765.46
25% Q1	5027.08
10%	2940.13
5%	2094.13
1%	1068.98
0% Min	54.00

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 5B: Extrémní záznamy – Objem transakcí

Extrémní záznamy			
Nejnižší		Nejnižší	
Hodnota	Hodnota	Hodnota	Hodnota
54.00	64512	554404	157811
92.77	50738	729410	46851
94.30	129446	764163	128697
139.00	166615	1156517	112558
140.00	163774	2871325	166193

Zdroj: ČSOB, SAS, vlastní zpracování

Příloha C: Výstup proc ANOVA

Tabulka 1C: Klasifikační úrovně

Klasifikační úrovně		
Clasifikace	Úrovně	Hodnoty
Master_mcc	13	1 2 3 4 5 6 7 8 9 10 11 12 13

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 2C: ANOVA 1

Zdroj	Stupně volnosti	Součet Čtverců	Průměr Čtverců	F Hodnota	Pr > F
Model	12	957528228950	79794019079	9101.53	<.0001
Chyba	2.08E6	1.8210404E13	8767102.9417		
Upravený celek	2.08E6	1.9167932E13			

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 3C: ANOVA 2

R-kvadrát	Koef. Var	RMSE	Průměr
0.049955	252.1640	2960.929	1174.208

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 4C: ANOVA 3

Zdroj	Stupně volnosti	Anova SS	Průměr^2	F Value	Pr > F
Master_mcc	12	957528228950	79794019079	9101.53	<.0001

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 5C: Bartlettův test homogenity

Bartlettův Test homogenity rozptylu			
Zdroj	Stupně volnosti	Chi-kvadrát	Pr > Chi^2
Master_mcc	12	1782650	<.0001

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 6C: Analýza skupin

Úrovně skupiny mcc	N	Objem transakcí	
		Průměr	Směrodatná odchylka
1	2468	928.31033	1665.1704
2	626	1480.44866	3758.4068
3	39824	853.77575	2483.1963
4	14666	1026.10441	2032.5537
5	1038788	569.49315	979.8572
6	107302	1049.98940	1847.7018
7	807556	1985.39628	4443.7091
8	50847	1009.19619	2540.1162
9	11538	1410.89118	3092.2113
10	1888	939.99316	1821.7662
11	1107	5851.63610	8114.2896
12	103	5776.12330	6509.6743
13	429	6343.06597	11508.8597

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 7C: Srovnání skupin MCC

Srovnání významně odlišná jsou označena ***.				
Srovnání Skupin mcc	Rozdíl průměrů	Současný 95% limit jistoty		
13 - 11	491.430	-280.719	1263.578	
13 - 12	566.943	-922.817	2056.702	
13 - 7	4357.670	3701.986	5013.353	***
13 - 2	4862.617	4011.640	5713.595	***
13 - 9	4932.175	4264.590	5599.760	***
13 - 6	5293.077	4636.258	5949.895	***
13 - 4	5316.962	4651.934	5981.989	***
13 - 8	5333.870	4675.601	5992.139	***
13 - 10	5403.073	4676.898	6129.247	***
13 - 1	5414.756	4704.556	6124.956	***
13 - 3	5489.290	4830.259	6148.321	***
13 - 5	5773.573	5117.928	6429.218	***
11 - 13	-491.430	-1263.578	280.719	
11 - 12	75.513	-1323.134	1474.159	
11 - 7	3866.240	3457.891	4274.589	***
11 - 2	4371.187	3692.225	5050.150	***
11 - 9	4440.745	4013.548	4867.942	***
11 - 6	4801.647	4391.478	5211.816	***
11 - 4	4825.532	4402.342	5248.721	***
11 - 8	4842.440	4429.952	5254.927	***
11 - 10	4911.643	4397.681	5425.605	***
11 - 1	4923.326	4432.193	5414.459	***

Srovnání významně odlišná jsou označena ***.				
Srovnání Skupin mcc	Rozdíl průměrů	Současný 95% limit jistoty		
11 - 3	4997.860	4584.158	5411.562	***
11 - 5	5282.143	4873.856	5690.430	***
12 - 13	-566.943	-2056.702	922.817	
12 - 11	-75.513	-1474.159	1323.134	
12 - 7	3790.727	2452.848	5128.606	***
12 - 2	4295.675	2852.012	5739.337	***
12 - 9	4365.232	3021.481	5708.984	***
12 - 6	4726.134	3387.699	6064.569	***
12 - 4	4750.019	3407.536	6092.502	***
12 - 8	4766.927	3427.779	6106.075	***
12 - 10	4836.130	3462.330	6209.931	***
12 - 1	4847.813	3482.389	6213.237	***
12 - 3	4922.348	3582.825	6261.870	***
12 - 5	5206.630	3868.770	6544.490	***
7 - 13	-4357.670	-5013.353	-3701.986	***
7 - 11	-3866.240	-4274.589	-3457.891	***
7 - 12	-3790.727	-5128.606	-2452.848	***
7 - 2	504.948	-37.914	1047.809	
7 - 9	574.505	447.207	701.804	***
7 - 6	935.407	891.291	979.523	***
7 - 4	959.292	846.166	1072.417	***
7 - 8	976.200	914.123	1038.278	***
7 - 10	1045.403	732.569	1358.237	***
7 - 1	1057.086	783.372	1330.800	***
7 - 3	1131.621	1061.928	1201.313	***
7 - 5	1415.903	1395.761	1436.046	***
2 - 13	-4862.617	-5713.595	-4011.640	***
2 - 11	-4371.187	-5050.150	-3692.225	***
2 - 12	-4295.675	-5739.337	-2852.012	***
2 - 7	-504.948	-1047.809	37.914	
2 - 9	69.557	-487.620	626.735	
2 - 6	430.459	-113.772	974.691	
2 - 4	454.344	-99.767	1008.455	
2 - 8	471.252	-74.729	1017.234	
2 - 10	540.456	-85.729	1166.640	
2 - 1	552.138	-55.448	1159.725	
2 - 3	626.673	79.774	1173.572	***
2 - 5	910.956	368.141	1453.770	***
9 - 13	-4932.175	-5599.760	-4264.590	***
9 - 11	-4440.745	-4867.942	-4013.548	***
9 - 12	-4365.232	-5708.984	-3021.481	***
9 - 7	-574.505	-701.804	-447.207	***
9 - 2	-69.557	-626.735	487.620	
9 - 6	360.902	227.881	493.923	***
9 - 4	384.787	215.832	553.742	***
9 - 8	401.695	261.688	541.702	***

Srovnání významně odlišná jsou označena ***.				
Srovnání Skupin mcc	Rozdíl průměrů	Současný 95% limit jistoty		
9 - 10	470.898	133.832	807.964	***
9 - 1	482.581	181.470	783.692	***
9 - 3	557.115	413.569	700.661	***
9 - 5	841.398	714.299	968.497	***
6 - 13	-5293.077	-5949.895	-4636.258	***
6 - 11	-4801.647	-5211.816	-4391.478	***
6 - 12	-4726.134	-6064.569	-3387.699	***
6 - 7	-935.407	-979.523	-891.291	***
6 - 2	-430.459	-974.691	113.772	
6 - 9	-360.902	-493.923	-227.881	***
6 - 4	23.885	-95.643	143.413	
6 - 8	40.793	-32.305	113.891	
6 - 10	109.996	-205.210	425.202	
6 - 1	121.679	-154.743	398.101	
6 - 3	196.214	116.547	275.880	***
6 - 5	480.496	436.960	524.032	***
4 - 13	-5316.962	-5981.989	-4651.934	***
4 - 11	-4825.532	-5248.721	-4402.342	***
4 - 12	-4750.019	-6092.502	-3407.536	***
4 - 7	-959.292	-1072.417	-846.166	***
4 - 2	-454.344	-1008.455	99.767	
4 - 9	-384.787	-553.742	-215.832	***
4 - 6	-23.885	-143.413	95.643	
4 - 8	16.908	-110.349	144.166	
4 - 10	86.111	-245.861	418.084	
4 - 1	97.794	-197.605	393.193	
4 - 3	172.329	41.188	303.469	***
4 - 5	456.611	343.711	569.512	***
8 - 13	-5333.870	-5992.139	-4675.601	***
8 - 11	-4842.440	-5254.927	-4429.952	***
8 - 12	-4766.927	-6106.075	-3427.779	***
8 - 7	-976.200	-1038.278	-914.123	***
8 - 2	-471.252	-1017.234	74.729	
8 - 9	-401.695	-541.702	-261.688	***
8 - 6	-40.793	-113.891	32.305	
8 - 4	-16.908	-144.166	110.349	
8 - 10	69.203	-249.014	387.420	
8 - 1	80.886	-198.965	360.737	
8 - 3	155.420	64.568	246.273	***
8 - 5	439.703	378.036	501.370	***
10 - 13	-5403.073	-6129.247	-4676.898	***
10 - 11	-4911.643	-5425.605	-4397.681	***
10 - 12	-4836.130	-6209.931	-3462.330	***
10 - 7	-1045.403	-1358.237	-732.569	***
10 - 2	-540.456	-1166.640	85.729	
10 - 9	-470.898	-807.964	-133.832	***

Srovnání významně odlišná jsou označena ***.				
Srovnání Skupin mcc	Rozdíl průměrů	Současný 95% limit jistoty		
10 - 6	-109.996	-425.202	205.210	
10 - 4	-86.111	-418.084	245.861	
10 - 8	-69.203	-387.420	249.014	
10 - 1	11.683	-403.441	426.807	
10 - 3	86.217	-233.572	406.007	
10 - 5	370.500	57.747	683.253	***
1 - 13	-5414.756	-6124.956	-4704.556	***
1 - 11	-4923.326	-5414.459	-4432.193	***
1 - 12	-4847.813	-6213.237	-3482.389	***
1 - 7	-1057.086	-1330.800	-783.372	***
1 - 2	-552.138	-1159.725	55.448	
1 - 9	-482.581	-783.692	-181.470	***
1 - 6	-121.679	-398.101	154.743	
1 - 4	-97.794	-393.193	197.605	
1 - 8	-80.886	-360.737	198.965	
1 - 10	-11.683	-426.807	403.441	
1 - 3	74.535	-207.104	356.173	
1 - 5	358.817	85.196	632.439	***
3 - 13	-5489.290	-6148.321	-4830.259	***
3 - 11	-4997.860	-5411.562	-4584.158	***
3 - 12	-4922.348	-6261.870	-3582.825	***
3 - 7	-1131.621	-1201.313	-1061.928	***
3 - 2	-626.673	-1173.572	-79.774	***
3 - 9	-557.115	-700.661	-413.569	***
3 - 6	-196.214	-275.880	-116.547	***
3 - 4	-172.329	-303.469	-41.188	***
3 - 8	-155.420	-246.273	-64.568	***
3 - 10	-86.217	-406.007	233.572	
3 - 1	-74.535	-356.173	207.104	
3 - 5	284.283	214.955	353.610	***
5 - 13	-5773.573	-6429.218	-5117.928	***
5 - 11	-5282.143	-5690.430	-4873.856	***
5 - 12	-5206.630	-6544.490	-3868.770	***
5 - 7	-1415.903	-1436.046	-1395.761	***
5 - 2	-910.956	-1453.770	-368.141	***
5 - 9	-841.398	-968.497	-714.299	***
5 - 6	-480.496	-524.032	-436.960	***
5 - 4	-456.611	-569.512	-343.711	***
5 - 8	-439.703	-501.370	-378.036	***
5 - 10	-370.500	-683.253	-57.747	***
5 - 1	-358.817	-632.439	-85.196	***
5 - 3	-284.283	-353.610	-214.955	***

Zdroj: ČSOB, SAS, vlastní zpracování

Příloha D: Výstup proc NPAR1WAY

Tabulka 1D: Wilcoxonovo ohodnocení

Wilcoxonovo ohodnocení (stupně sum) pro objem transakcí klasifikováno skupinami MCC					
Skupiny MCC	N	Součet hodnocení	Očekávání Pod H0	Směrodatná odchylka pod H0	Průměr hodnocení
5	1038788	9.08551E11	1.07886E12	432077694	874625.64
6	107302	1.2614E11	1.11441E11	19126920	3 1175563.34
7	807556	1.00198E12	8.38705E11	42125304	3 1240752.29
9	11538	1.29338E10	1.1983E10	64226530.9	1120978.08
8	50847	5.00199E10	5.28082E10	133539476	983732.75
3	39824	3.36841E10	4.13601E10	118502501	845823.28
1	2468	2717365407	2563194462	2976961	2 1101039.47
4	14666	1.57988E10	1.52317E10	72356234.4	1077238.90
10	1888	1870015498	1960822992	26041277.7	990474.31
2	626	742978258	650145759	14999627.4	1186866.23
13	429	696462919	445547174	12417739.7	1623456.69
11	1107	1951740409	1149698651	19944208.7	1763089.80
12	103	177494730	106972865	6085084.35	1723249.81
Average scores were used for ties.					

Zdroj: ČSOB, SAS, vlastní zpracování

Tabulka 2D: Kruskal-Wallisův Test

Kruskal-Wallisův Test	
Chi-kvadrát	182138.6875
DF	12
Pr > Chi-kvadrát	<.0001

Zdroj: ČSOB, SAS, vlastní zpracování

Příloha E: Výsledná tabulka Dunnovy metody

Tabulka 1E: Donnova metoda mnohonásobného porovnání

Skupiny MCC	N	Součet hodnocení	cpr. skupina	Ti-Tj	ABS Ti-Tj	Dunn	rozhodnutí
5	1038788	9,09E+11	6	7,82E+11	7,82411E+11	6171,017	Zamítá
5	1038788	9,09E+11	7	-9,3E+10	93429000000	2855,1	Zamítá
5	1038788	9,09E+11	9	8,96E+11	8,95617E+11	18015,57	Zamítá
5	1038788	9,09E+11	8	8,59E+11	8,58531E+11	8740,961	Zamítá
5	1038788	9,09E+11	3	8,75E+11	8,74867E+11	9826,786	Zamítá
5	1038788	9,09E+11	1	9,09E+11	9,08524E+11	38784,44	Zamítá

Skupiny MCC	N	Součet hodnocení	cpr. skupina	Ti-Tj	ABS Ti-Tj	Dunn	rozhodnutí
5	1038788	9,09E+11	4	8,93E+11	8,92752E+11	16003,07	Zamítá
5	1038788	9,09E+11	10	9,09E+11	9,08532E+11	44331,06	Zamítá
5	1038788	9,09E+11	2	9,09E+11	9,08544E+11	76941,11	Zamítá
5	1038788	9,09E+11	13	9,09E+11	9,08544E+11	92934,25	Zamítá
5	1038788	9,09E+11	11	9,09E+11	9,08531E+11	57872,5	Zamítá
5	1038788	9,09E+11	12	9,09E+11	9,08549E+11	189634,7	Zamítá
6	107302	1,26E+11	7	-8,76E+11	8,7584E+11	6253,187	Zamítá
6	107302	1,26E+11	9	1,13E+11	1,13206E+11	18855,02	Zamítá
6	107302	1,26E+11	8	7,61E+10	76120100000	10361,23	Zamítá
6	107302	1,26E+11	3	9,25E+10	92455900000	11292,32	Zamítá
6	107302	1,26E+11	1	1,26E+11	1,26113E+11	39181,42	Zamítá
6	107302	1,26E+11	4	1,10E+11	1,10341E+11	16942,52	Zamítá
6	107302	1,26E+11	10	1,26E+11	1,26121E+11	44678,78	Zamítá
6	107302	1,26E+11	2	1,26E+11	1,26133E+11	77141,98	Zamítá
6	107302	1,26E+11	13	1,26E+11	1,26133E+11	93100,62	Zamítá
6	107302	1,26E+11	11	1,26E+11	1,2612E+11	58139,29	Zamítá
6	107302	1,26E+11	12	1,26E+11	1,26138E+11	189716,2	Zamítá
7	807556	1,00E+12	9	9,89E+11	9,89046E+11	18043,89	Zamítá
7	807556	1,00E+12	8	9,52E+11	9,5196E+11	8799,164	Zamítá
7	807556	1,00E+12	3	9,68E+11	9,68296E+11	9878,594	Zamítá
7	807556	1,00E+12	1	1,00E+12	1E+12	38797,6	Zamítá
7	807556	1,00E+12	4	9,86E+11	9,86181E+11	16034,93	Zamítá
7	807556	1,00E+12	10	1,00E+12	1E+12	44342,57	Zamítá
7	807556	1,00E+12	2	1,00E+12	1E+12	76947,74	Zamítá
7	807556	1,00E+12	13	1,00E+12	1E+12	92939,74	Zamítá
7	807556	1,00E+12	11	1,00E+12	1E+12	57881,32	Zamítá
7	807556	1,00E+12	12	1,00E+12	1E+12	189637,3	Zamítá
9	11538	1,29E+10	8	-3,7E+10	37086100000	19845,27	Zamítá
9	11538	1,29E+10	3	-2,1E+10	20750300000	20346,89	Zamítá
9	11538	1,29E+10	1	1,29E+10	12906626346	42680,94	Zamítá
9	11538	1,29E+10	4	-2,9E+09	2865000000	23948,45	Zamítá
9	11538	1,29E+10	10	1,29E+10	12915099845	47777,33	Zamítá
9	11538	1,29E+10	2	1,29E+10	12926370217	78976,99	Zamítá
9	11538	1,29E+10	13	1,29E+10	12926835371	94626,66	Zamítá
9	11538	1,29E+10	11	1,29E+10	12914282596	60552,92	Zamítá
9	11538	1,29E+10	12	1,29E+10	12932025053	190469,8	Zamítá
8	50847	5,00E+10	3	1,63E+10	16335800000	12877,87	Zamítá
8	50847	5,00E+10	1	5E+10	49992726346	39667,45	Zamítá
8	50847	5,00E+10	4	3,42E+10	34221100000	18038,07	Zamítá

Skupiny MCC	N	Součet hodnocení	cpr. skupina	Ti-Tj	ABS Ti-Tj	Dunn	rozhodnutí
8	50847	5,00E+10	10	5E+10	50001199845	45105,61	Zamítá
8	50847	5,00E+10	2	5E+10	50012470217	77389,97	Zamítá
8	50847	5,00E+10	13	5E+10	50012935371	93306,21	Zamítá
8	50847	5,00E+10	11	5E+10	50000382596	58467,94	Zamítá
8	50847	5,00E+10	12	5E+10	50018125053	189817,2	Zamítá
3	39824	3,37E+10	1	3,37E+10	33656926346	39920,77	Zamítá
3	39824	3,37E+10	4	1,79E+10	17885300000	18588,52	Zamítá
3	39824	3,37E+10	10	3,37E+10	33665399845	45328,55	Zamítá
3	39824	3,37E+10	2	3,37E+10	33676670217	77520,12	Zamítá
3	39824	3,37E+10	13	3,37E+10	33677135371	93414,18	Zamítá
3	39824	3,37E+10	11	3,37E+10	33664582596	58640,1	Zamítá
3	39824	3,37E+10	12	3,37E+10	33682325053	189870,3	Zamítá
1	2468	27173654	4	-1,6E+10	15771626346	41871,23	Zamítá
1	2468	27173654	10	8473499	8473499,09	58841,69	Zamítá
1	2468	27173654	2	19743871	19743871,49	86122,22	Zamítá
1	2468	27173654	13	20209025	20209024,88	100667,2	Zamítá
1	2468	27173654	11	7656250	7656249,98	69615,58	Zamítá
1	2468	27173654	12	25398707	25398706,77	193541,7	Zamítá
4	14666	1,58E+10	10	1,58E+10	15780099845	47055,4	Zamítá
4	14666	1,58E+10	2	1,58E+10	15791370217	78542,36	Zamítá
4	14666	1,58E+10	13	1,58E+10	15791835371	94264,22	Zamítá
4	14666	1,58E+10	11	1,58E+10	15779282596	59984,95	Zamítá
4	14666	1,58E+10	12	1,58E+10	15797025053	190290	Zamítá
10	1888	18700155	2	11270372	11270372,4	88758,36	Zamítá
10	1888	18700155	13	11735526	11735525,79	102931,5	Zamítá
10	1888	18700155	11	-817249	817249,11	72851,48	Zamítá
10	1888	18700155	12	16925208	16925207,68	194729,1	Zamítá
2	626	7429783	13	465153,4	465153,39	120621,6	Zamítá
2	626	7429783	11	-1,2E+07	12087621,51	96239,44	Zamítá
2	626	7429783	12	5654835	5654835,28	204631,6	Zamítá
13	429	6964629	11	-1,3E+07	12552774,9	109448	Zamítá
13	429	6964629	12	5189682	5189681,89	211165,7	Zamítá
11	1107	19517404	12	17742457	17742456,79	198250,8	Zamítá

Zdroj: ČSOB, SAS, vlastní zpracování