Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra statistiky

Diplomová práce

Prediktivní modelování chování zákazníků vybrané databáze

Bc. Matěj Šůcha

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Bc. Matěj Šůcha

Provoz a ekonomika

Název práce

Prediktivní modelování chování zákazníků vybrané databáze

Název anglicky

Predictive modeling of customers behavior of a selected database

---

Cíle práce

Cílem této práce je pomocí prediktivního modelování identifikovat faktory, které ovlivňují chování zákazníků vybrané databáze.

Metodika

Práce bude postavena na využití statistických postupů data mining. Dle povahy dat se počítá se zapojením klasifikačních metod (např. shluková analýza), dále pak rozhodovacích stromů a především regresních modelů (klasických i logistických).

Doporučený rozsah práce

60 – 80 stran

Klíčová slova

Big data, prediktivní modelování, chování zákazníka, kupní rozhodování

Doporučené zdroje informací

ABBOTT, D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.

BERRY, M J A. – LINOFF, G. Data mining techniques : for marketing, sales, and customer relationship management. Indianapolis: Wiley, 2011. ISBN 978-0-470-65093-6.

HEBÁK, P. *Statistické myšlení a nástroje analýzy dat.* Praha: Informatorium, 2015. ISBN 978-80-7333-118-4.

MACHEK, M. – KELLER, K L. – JUPPA, T. – KOTLER, P. *Marketing management.* Praha: Grada, 2013. ISBN 978-80-247-4150-5.

NISBET, R. – MINER, G. – ELDER, J. Handbook of statistical analysis and data mining applications. Amsterdam: Amsterdam, 2009. ISBN 978-0-12-374765-5.

RUD, O P. Data mining : praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computer Press, 2001. ISBN 80-7226-577-6.

Předběžný termín obhajoby

2017/18 LS – PEF

Vedoucí práce

Ing. Tomáš Hlavsa, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 19. 2. 2018

prof. Ing. Libuše Svatošová, CSc.

Vedoucí katedry

Elektronicky schváleno dne 20. 2. 2018

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 07. 03. 2018

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Prediktivní modelování chování zákazníků vybrané databáze" jsem vypracoval samostatně pod vedením vedoucího diplomové 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 diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 19.3.2018 _____

Poděkování

Rád bych touto cestou poděkoval Ing. Tomáši Hlavsovi, Ph.D. za jeho odborný dozor, rady, pomoc, a všeobecnou ochotu při zpracování této práce.

Prediktivní modelování chování zákazníků vybrané databáze

Abstrakt

Diplomová práce je zaměřena na chování zákazníka, kdy jsou za dodržení data miningového procesu pomocí prediktivních modelů určeny faktory, které mají největší vliv na odchod zákazníka od dané společnosti. V první části práce jsou zkoumány teoretická východiska, kdy je pozornost věnována fenoménu Big Data, následně statistickým metodám pro práci s daty, prediktivnímu modelování, data miningu a problematice chování zákazníka. Ve druhé části práce pak probíhá vlastní výzkum. K analýze byly vybrány data o zákaznících telekomunikační společnosti, na jejichž základě pak byly sestaveny čtyři prediktivní modely. Cílem těchto modelů je předpovídat klasifikaci zákazníků podle toho, jestli u společnosti zůstanou, nebo odejdou. Použity byly logistická regrese, rozhodovací strom, náhodný les a neuronová síť. Jako nejpřesnější model pro daný soubor se ukázala logistická regrese, kdy se jako významné faktory ovlivňující zákazníkovo setrvání či odchod ukázaly Internet přes optický kabel, Smlouva na 2 a 1 rok, Výše celkových plateb a Délka smlouvy v měsících. Na základě těchto výsledků pak byla vytvořena charakteristika odcházejícího zákazníka a navrhnuta opatření.

Klíčová slova: Big data, data mining, prediktivní modelování, logistická regrese, chování zákazníka

Predictive modeling of customer behavior of a selected database

Abstract

This diploma thesis is focused on customer behavior, where by folowing the data mining process the factors with the highest influence on customer churn are being observed with the use of predicitve models. The first part of this thesis discusses the theoretical background, with the main focus on the Big Data phenomenon, statistical methods used when working with data, predictive modeling, data mining, and customer behavior. The experiment is then carried out in the second part of the thesis. Four predictive models with the goal of predicting customer churn were built based on a data set containing information on a telecommunications company customers. Logistic regression, decision tree, random forest, and neural network were built and the most accurate model for this particular dataset turned out to be the logistic regression. As the most influential factors in customer churn, this model showed Having the internet over fiber optic, 2 or 1 year contract, Total payments and Tenure. Based on these results, a characteristic of a churner was created as well as a list of recommendations for the company.

Keywords: Big Data, data mining, predictive modeling, logistic regression, customer behavior

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Big Data je pojem, který se v současnosti užívá v čím dál tím více odvětvích. Už delší dobu to není pojem, který by byl spojován pouze s informačními technologiemi a příslušnými odděleními společností. Množství dat, která jsou denně vyprodukována po celém světě, jsou nejen ohromná, ale také velmi různorodá. V podstatě každý člověk, který využívá internet, je chodícím generátorem dat. Data se však sbírají i jinde, jako například v obchodních řetězcích, v bankovnictví, v leteckém či automobilovém průmyslu atd. Bylo jen otázkou času, kdy lidem dojde potenciál, který v sobě tato data ukrývají. Před několika lety se začaly používat databázové prostředky a přístupy, které sloužily k práci s nasbíranými daty. Z dnešního pohledu však umožňovaly jen základní operace. Tyto klasické přístupy byly dostačující, s rostoucími objemy dat se však stávaly zastaralými a nedostačujícími, a tak bylo třeba vyvinout nové přístupy. Jako reakce na tuto potřebu bylo vyvinuto mnoho programů a přístupů, které umožňují práci s takto velkými objemy dat. Souhrnně se data, která jsou tak velká, že na práci s nimi běžná výpočetní technika nestačí, nazývají Big Data. Tyto nové programy a přístupy jsou zaměřeny právě na práci s nimi. Je tedy zřejmé, že analyzovat takto velké soubory vyžaduje velmi výkonnou výpočetní techniku, což už ale v dnešní době cloudových technologií není žádným problémem. Pro firmy tak není nutnost mít supervýkonné počítače přímo na pracovišti, ale většina analýz se již provádí právě přes cloud neboli na vzdáleném výkonném počítači. S nárůstem významnosti tohoto odvětví také narůstá poptávka nejen po výkonnějších technologiích, ale také po lidech, kteří se na práci s daty specializují, takzvaných datových specialistech. Big Data tedy nejen umožňují pozorovat předem neviděné souvislosti, ale zároveň tvoří nárok na vývoj na poli technologií a poptávku po specialistech v tomto oboru.

Jedním z polí, kde se analýza dat používá nejvíce, je marketing a celkově aktivity zaměřené na zákazníka. Pomocí Big Data analýz lze predikovat velké množství informací, které mohou společnostem poskytnout významnou konkurenční výhodu. Správně provedené analýzy mohou společnostem pomoci například k určení zákazníkových potřeb a chtíčů, stejně tak jako mohou odhalit, co zákazníky naopak odrazuje. V současné době je souboj společností o zákazníky na nejvyšší úrovni v historii lidstva, a právě z tohoto důvodu je znalost zákazníků pro společnosti nutností. Je všeobecně známo, že udržení stávajícího zákazníka je pro společnost méně nákladné než získání nového. Problém nastává u telekomunikačních společností, kde dochází k největšímu přesunu zákazníků mezi

konkurenty (Hassouna et al., 2015). Tyto společnosti kvůli tomu tratí ohromné sumy peněz. Pro telekomunikační společnosti je tak analýza zákaznických dat naprosto zásadní. Díky ní může daná společnost určit faktory, které nejvíce přispívají k odlivu jejich zákazníků, a na základě výsledků přijmout opatření vedoucí ke snížení tohoto odlivu. Vzhledem k velkému množství zákaznických dat, které mají tyto společnosti k dispozici, lze právě pomocí datových analýz (převážně prediktivního modelování) poznat, co zákazníky u dané firmy těší, a co naopak odrazuje. Na základě toho pak mohou přizpůsobit nejen nabízené služby a jejich kvalitu, ale i celou prezentaci společnosti navenek. Existuje široká škála prediktivních modelů, které jsou schopny tyto faktory určit. I když Hassouna et al. (2015) považují za nejvhodnější model pro predikování odchodu zákazníka rozhodovací strom, vhodnost modelu se odvíjí od skladby dostupných dat. Je tedy možné, že pro jednu společnost bude tím nejvhodnějším modelem již zmiňovaný rozhodovací strom, pro jiné to bude neuronová síť, a pro jiné to může být logistická regrese. Tato práce bude za využití čtyř modelů analyzovat zákaznická data a zkoumat vliv jednotlivých faktorů na odchod zákazníků.

2 Cíl práce a metodika

2.1 Cíl práce

Cílem této práce je určit faktory, které výraznou měrou přispívají k odchodu zákazníků od dané společnosti. Jinými slovy, cílem je za využití prediktivního modelování určit faktory, které nejvíce ovlivňují chování zákazníků vybrané databáze, a na základě těchto výsledků společnosti navrhnout opatření. Prvním z dílčích cílů práce je sestavení čtyř prediktivních modelů, které budou s co největší přesností zákazníky řadit buď jako odcházející, či zůstávající. Dalším dílčím cílem je výběr toho nejlepšího modelu. Posledním dílčím cílem je sestavit charakteristiku odcházejícího zákazníka.

2.2 Metodika

První část práce zabývající se teoretickými východisky byla zpracována studiem literatury. Vlastní práce pak byla založena na analýze datového souboru obsahujícího informace o 7043 zákaznících telekomunikační společnosti. Pro účely analýzy byla využita jedna z metod data miningu, která se nazývá CRISP-DM, a jejíž kroky jsou detailněji popsány v části zabývající se teoretickými východisky. V první části bylo třeba získat data a následně jim porozumět. Porozumění dat obsahovalo seznámení se s jednotlivými proměnnými a jejich jednotkami a také popis dat obsažených v souboru. Popis dat byl proveden procentuálními vyjádřeními a také za využití průměru. Následovalo sestavení jednotlivých modelů, kterými byly logistická regrese, rozhodovací strom, náhodný les a neuronová síť. Jednotlivé modely byly sestaveny v softwaru Rapidminer. Pro vyhodnocení přesnosti jednotlivých modelů byly využity tři metody. Tou první bylo jednoduché měření výkonnosti modelu, kdy byl model aplikován na stejná data, která byla použita pro jeho sestavení. Druhou metodou byla tzv. cross-validation, kdy byl základní soubor rozdělen na deset podsouborů, následně byl model sestaven na devíti z nich a testován na desátém podsouboru. Tento krok vedl ke snížení celkové přesnosti u všech modelů, neboť byl model testován na předem neviděných datech. Měření přesnosti modelu bylo provedeno jednoduchým výpočtem, kdy byly sečteny počty přesně zařazených zákazníků a vyděleny celkovým počtem zákazníků.

$$Přesnost\ klasifikace = \frac{Skutečná\ pozitiva + Skutečná\ negativa}{Celkový\ počet\ zákazníků}$$

Rovnice 1 Přesnost klasifikace. Zdroj: Hassouna et al., 2015

Poslední metodou k vyhodnocení výkonnosti jednotlivých modelů bylo sestavení ROC křivek, které zobrazují přesnost modelu v grafické podobě. Při sestavování této křivky je na osu y zanášen poměr skutečných pozitiv, který se spočítá následovně:

$$Poměr\ skutečných\ pozitiv = \frac{Skutečná\ pozitiva}{Celková\ pozitiva}$$

Rovnice 2 Poměr skutečných pozitiv. Zdroj: Hassouna et al., 2015

a na osu x je zanášen poměr falešných pozitiv, který se spočítá následovně:

$$Poměr\ falešných\ pozitiv = \frac{Falešná\ negativa}{Celková\ negativa}$$

Rovnice 3 Poměr falešných negativ. Zdroj: Hassouna et al., 2015

Výsledkem je pak křivka, která v ideálním případě prochází bodem (0,1), a znázorňuje přesnost klasifikace. U logistické regrese, rozhodovacího stromu a náhodného lesa byly také určeny váhy jednotlivých proměnných, a to opět pomocí nástroje v softwaru Rapidminer. U logistické regrese byly navíc tyto váhy vneseny go grafu jednak pro lepší názornost, jednak pro určení stěžejních faktorů. Následně byl určen model, který byl na daný soubor dat nejpřesnější. Určen byl dvěma způsoby, a to jak analýzou předchozích výsledků, tak zobrazením ROC křivek všech modelů v jednom grafu. Nejlepší model byl ten, který měl nejvyšší procentuální přesnost klasifikace, a jehož ROC křivka se nejvíce přibližovala bodu (0,1). Model pak byl aplikován na soubor zákaznických dat. Model, který byl na základě analýzy určen jako nejvhodnější, byl použit pro finální modelování chování zákazníka. Výsledkem byl soubor, který u každého zákazníka navíc uváděl, jestli byl modelem označen jako odcházející, či zůstávající. Na základě těchto výsledků byla sestavena charakteristika odcházejícího zákazníka. V poslední řadě byla na základě všech předchozích analýz sestavena doporučená opatření ke snížení počtu odcházejících zákazníků.

3 Teoretická východiska

3.1 Úvod do Big Data

Big Data se stávají nedílnou součástí všedního života. A to už nejen u velkých firem, které zpracovávají terrabyty dat každou vteřinu, ale i u menších podniků, které si pomalu začínají uvědomovat potenciál, který leží právě v datech, které jsou nyní běžně dostupné i menším firmám. Ač je pojem Big Data relativně nový a zdaleka ne všichni jsou s ním obeznámeni, je to pojem, který bude udávat tempo vývoje ať už na poli medicíny, výzkumu, či marketingu. V dnešní době existuje naprostý přebytek dat. Může se zdát, že objemy dat, které jsou denně produkovány všemi lidmi a přístroji na světě, jsou naprosto zbytečné. V historii se však našli lidé, kteří v těchto datech viděli velký potenciál. Tušili, že tolik dat bude možné využít na víc věcí než k tomu, k čemu byla primárně určena. Právě díky těmto lidem vznikl obor na zpracovávání velkých objemů dat, ze kterých lze vyčíst mnohdy až neuvěřitelné informace. Jak píše Gordon (2014), díky využití Big Data byla jedna americká telefonní společnost schopna snížit dobu na vyřízení zakázek o celých 92 % a další firma zlepšila efektivitu umisťování generátorů elektrické energie o 99 %. Podle Gordona (2014) bude trh s Big Data technologiemi a službami růst v průměru přes 30 % za rok, což je zhruba 7krát více, než trh s informačními a komunikačními technologiemi. Využívání dat bude tedy pro podniky naprosto zásadní aktivitou často rozhodující o úspěchu a neúspěchu.

Big Data také přinášejí změnu do organizační hierarchie. Ať si to lidé uvědomují či ne, spolupráce mezi manažery a datovými specialisty bude v éře Big Data naprosto nepostradatelná. Datoví specialisté budou ti, kdo budou umět data zpracovat, kdežto manažeři budou ti, kdo budou vědět, co s danými výstupy dělat. Jejich kooperace bude pro podniky klíčová, protože když si budou tato dvě oddělení rozumět, firma bude díky jejich vzájemné práci s Big Data získávat cennou konkurenční výhodu. Naproti tomu, když tyto subjekty nebudou spolupracovat, jeden se bude snažit udělat všechno bez druhého, firma bude na Big Data analýzách pouze tratit.

3.1.1 Historie zpracování dat

Data existovala daleko dříve, než byly k dispozici počítače schopné s nimi pracovat. Vyvstává tedy otázka, proč se daty nezabývali lidé již před čtyřiceti lety? Odpověď je velmi jednoduchá – data nebylo tak lehké získat (čili nebyly v elektronické podobě) (Simon, 2013).

První tabulkový software byl vyvinut až v polovině 70. let Danem Bricklinem (Simon, 2013). V letech 80. se začaly rozvíjet aplikace jako MRP (manufacture-resource planning) a ERP (enterprise-resource planning), které měly za úkol zautomatizovat standardní procesy. Tyto původní databáze uměly pracovat pouze s uspořádanými daty, bylo tedy náročnější je zadávat do systému (Simon, 2013). Poté se začaly vyvíjet relační datové modely, ve kterých jsou data zapsána do tabulek, které jsou propojené alespoň s jednou další tabulkou. Například tak bylo umožněno zaznamenávat více nákupů u jednoho zákazníka. Každá tabulka pak představuje entitu, jejichž vztah lze vyjádřit pomocí grafu vztahů mezi entitami (Simon, 2013). Toto vylepšení přineslo zásadní změnu, protože umožnilo mnoho nových operací s daty a tabulkami jako celky. Tento model zaznamenal největší rozvoj na přelomu milénia, ale také se hodil pouze na práci s uspořádanými daty. Vzhledem k popularitě tohoto modelu si dodnes většina lidí myslí, že před prací s daty je nutné je řádně strukturovat, nicméně doba již pokročila a už tomu tak není (Simon, 2013).

Velká změna v práci s daty přišla kolem roku 2005, kdy lidstvo vstoupilo do éry Web 2.0 neboli éry sociálních sítí. Do roku 2005 firmy pracovaly převážně s daty, které si samy vytvořily uvnitř společnosti. S nástupem sociálních sítí však narostl objem, rozmanitost a rychlost vzniku dat, která byla pro firmy externí. Do doby před nástupem sociálních sítí byla většina dat strukturovaných, uspořádaných a konzistentních (Simon, 2013), nicméně s webem 2.0 přišla velká změna a většina dat je nyní nestrukturovaných, chaotických a není jednoduché je uchovávat v tradičních tabulkách. Podle Simona (2013) tyto data představují zhruba 80 % všech dat v podnicích.

3.1.2 **Typy dat**

Datový management pracuje s několika typy dat. Strukturovaná a nestrukturovaná data byla již zmíněna, existují však další dva typy dat, jež přišla do povědomí s nástupem webu 2.0 (to však neznamená, že dříve neexistovaly). Jsou to semi – strukturovaná data a metadata. Semi-strukturovaná data obsahují charakteristiky jak strukturovaných dat, tak těch nestrukturovaných. Patří sem například značkovací jazyky, jako je XML, e-maily a elektronická výměna dat (EDI – electronic data interchange) (Simon, 2013). Druhým typem dat jsou metadata, což jsou data o datech. Lidé je často používají, aniž by si to uvědomovali. Jsou to například značky, kterými označují fotky na sociálních sítích a pomocí těchto označení je pak mohou hledat cizí lidé. Metadata slouží k detailnějšímu popisu dat –

například fotografie obsahují mnoho metadat, ať už datum a čas pořízení, GPS lokaci, velikost, či uživatelem vytvořené značky (Simon, 2013). Bez metadat by bylo velmi těžké hledat mezi miliardami datových záznamů volně dostupných lidem, ať už na sociálních sítích, či ve veřejných databázích.

Jak již bylo zmíněno dříve, nestrukturovaná data existovala již dávno, aniž by si to lidé uvědomovali. Co je však nového je to, že velká většina těchto dat je k dispozici téměř okamžitě, protože jsou v digitální podobě. A ne však proto, že by se našel nějaký dobrovolník, který by dennodenně skenoval a přepisoval záznamy do elektronické podoby, ale jednoduše proto, že valná většina dat už vzniká v digitální podobě. I přes to, že je většina dat, i když těch nestrukturovaných, firmám k dispozici relativně kdykoliv, zatím jen pouze velmi málo z nich začalo využívat jejich potenciál (Simon, 2013). Tento jev má v mnoha případech jasnou příčinu – firmy nejsou schopny správně spravovat ani svoje interní strukturovaná data, natož pak aby se zabývaly těmi nestrukturovanými mimo jejich organizaci (Simon, 2013). Další příčinou bývá to, že firmy svá data využívají pouze na to, na co byly původně určeny a nevidí v datech potenciál pro celý podnik. Jednotlivá oddělení pracují se svými daty pouze pro svoje účely. Pro správné využívání dat, které přinese benefity celému podniku, je třeba, aby mezi sebou oddělení spolupracovala co se týče datového managementu (Simon, 2013).

3.1.3 **Co jsou Big Data**

Základní myšlenkou za pojmem Big Data je to, že vše, co lidé dělají, ze sebou nechává digitální stopu (Marr, 2015). Tato data pak lze využít k analýze. Tažnými silami tohoto nového světa jsou stále rostoucí objemy dat a stále se zvyšující technologické možnosti, jak z dat dostat poznatky vhodné pro využití v podnicích. Je důležité říci, že to nejsou objemy dat, které jsou tak důležité, ale schopnost lidí tyto objemy zanalyzovat. Tato schopnost se neustále zvyšuje a rozšiřuje na více a více lidí, a to jak díky rozvoji cloud computingu společně se zvyšujícími se rychlostmi připojení k internetu, tak i díky novým kreativním způsobům zpracování dat (Marr, 2015). Všechno tohle znamená, že k práci s velkými objemy dat již není třeba stavět superpočítače, ale stačí mít ve firmě lidi, kteří vědí, jak s daty naložit a cloudové počítače jsou schopny udělat veškerou tvrdou práci za ně. Oblastí, kde leží asi největší hodnota Big Data, je schopnost analyzovat neuspořádaná data. Neuspořádaná data jsou taková data, která není možno skladovat a indexovat klasickými

postupy nebo v klasických databázích (Marr, 2015). Patří sem například emailové konverzace, příspěvky na sociálních sítích, obsah videí, fotky, zvuky atd. (Marr, 2015).

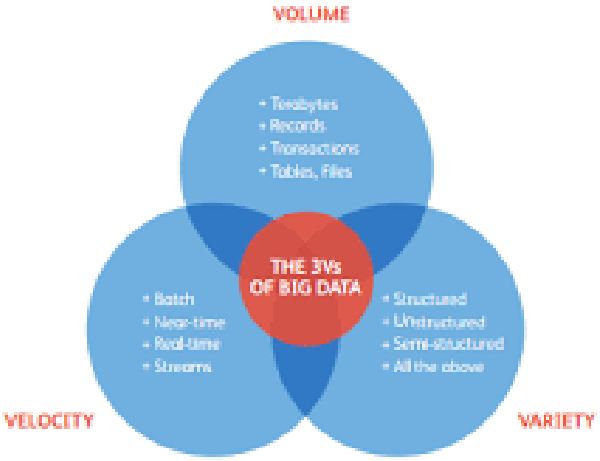
Definic, co přesně se skrývá pod pojmem Big Data, je hodně. Skoro všechny však obsahují zmínku o třech „V“ Big dat – Volume, Variety a Velocity. V poslední době se stále častěji však hovoří o dalším „V“, kterým je Veracity. Zmiňuje se o něm například firma IBM, která je na poli Big Data považována za inovátora a velkého hráče. Naproti tomu firma SAS zmiňuje ty tři základní „V“ a přidává k nim pojmy Variability a Complexity.

Volume neboli objem znázorňuje neustále rostoucí objem dat. Jak píše McAfee (2012), v roce 2012 bylo každý den vytvořeno zhruba 2,5 exabytů dat a toto číslo se každých 40 měsíců zdvojnásobuje. Podle tohoto pravidla by se mělo v roce 2018 vytvářet kolem 10 exabytů dat denně. Podle společnosti IBM je denně vyprodukováno 2,5 bilionu gigabytů dat, která pocházejí z mnoha různých zdrojů. Tyto data nejsou nijak uspořádána – jednoduše se zaznamená to, co se stalo, a důležitost daného záznamu se ukáže až při případných analýzách, což vede ke druhému „V“.

Variety neboli rozmanitost poukazuje na to, že data jsou mnoha typů a pocházejí z mnoha zdrojů. Obrovská množství dat pocházejí ze sociálních sítí a chytrých telefonů, což jsou věci, které jsou stále relativně nové a produkují takové typy dat, na které nejsou staré databázové systémy stavěny a neví si s nimi rady (McAfee, 2012). IBM hezky popisuje tuto rozmanitost dat – v roce 2011 bylo k dispozici 150 exabytů zdravotních záznamů, každý měsíc se na Facebooku sdílí přes 30 miliard částí obsahu, na YouTube je měsíčně shlédnuto přes čtyři miliardy hodin videí a na Twitteru je denně posláno přes 400 milionů tweetů. Díky rozmanitosti dat je také možno doplnit chybějící záznamy, a to pomocí fúze dat.

Velocity neboli rychlost ukazuje jednak na fakt, že data jsou zpracovávána skoro v reálném čase, což umožňuje podnikům okamžité reakce, a jednak na to, s jakou rychlostí data přibývají. Například moderní auta mají kolem 100 senzorů a data z nich jsou v reálném čase zaznamenávány a vyhodnocovány. Newyorská burza během každého nakupovacího dne zaznamená jeden terrabyte dat, které je nutné zpracovávat okamžitě, jelikož na burze každá vteřina rozhoduje (IBM).

Posledním „V“ podle IBM je Veracity neboli věrohodnost. S tím, kolik dat se každou vteřinu nasbírá, přichází problém věrohodnosti dat. Nelze očekávat, že všechna zaznamenaná data budou pravdivá a věrohodná – občas jsou záměrně poskytnuta matoucí data, občas lidé tvoří nevěrohodná data nevědomky.



Obrázek 1 3 V Big Dat. Zdroj: Gerasimou, 2016.

SAS přidává další pojmy, kterými jsou Variability a Complexity. Variabilita se může zdát skoro stejná jako rozmanitost, ale přeci jen se najdou malé rozdíly. Variabilita poukazuje na skutečnost, že data proudí v takových rychlostech a variacích, že je těžké zvládat denní či sezónní výkyvy v proudění dat. Navíc, díky neuspořádanosti dat, je práce s nimi o to složitější (SAS). Komplexnost odkazuje na spousty zdrojů dat, které produkují data v různých formátech, což také ztěžuje práci s nimi (SAS).

Model 3 „V“ je pak využit k formulování definice Big Data. V roce 2001 formulovala výzkumná skupina META (nyní Gartner) Big Data takto: Big Data jsou informační zdroje o velkém objemu, velké rychlosti a vysoké rozmanitosti, které ke zlepšenému rozhodování, objevování nových souvislostí a k optimalizaci procesů vyžadují nové formy zpracování (Gartner, 2011). Tato definice se používá širokou veřejností a vycházejí z ní i další definice, jako například i ta od De Maura et al. (2016), která definuje Big Data jako informační zdroj charakterizovaný tak velkým objemem, rychlostí a rozmanitostí, že k tomu, aby se tento zdroj dal převést na něco hodnotného, je třeba specifických technologických a analytických metod. Z obou těchto definic je patrných několik poznatků – rozvoj technologií v Big Data hraje nemalou roli, k tomu, aby Big Data měla nějakou hodnotu, je třeba aplikovat mnohdy složité analytické a technologické metody,

a fakt, že 3 „V“ (4 „V“) jsou takovým středobodem celého oboru Big Data. V neposlední řadě je důležitá zmínka o tom, že se data převádějí na něco hodnotného. To je celou podstatou Big Data – převést nepřeberné množství různých dat na něco, co bude mít pro uživatele určitou hodnotu.

3.1.4 **Možnosti využití**

Big Data mají využití v mnoha sférách dnešní společnosti. Podle společnosti Intel se Big Data používají hlavně k objevování vzorů v chování a k předpovídání trendů. Datové analýzy jsou dnes schopné pomáhat v mnoha odvětvích, počínaje zdravotnictvím, přes ochranu životního prostředí až k marketingu. Níže jsou popsány příklady použití Big Data.

Polem, ke kterému Big Data neodmyslitelně patří, je marketing. Od počátku konzumerismu, kdy lidé začali masivně nakupovat vše, co mohli, bylo cílem prodejců vytvářet reklamy tak, aby fungovaly na co největší počet lidí. Tyto reklamy však nemohly fungovat na všechny zákazníky, protože lidé, ač si jsou v mnoha ohledech nákupního chování velmi podobní, jsou odlišní, a každý má trochu jiné potřeby. Jak tedy zacílit na co největší počet zákazníků? Zde přicházejí na řadu Big Data. Společnost Amazon, která začala jako internetový obchod s knihami, byla jednou z prvních firem, která využila individuálního targetingu. Díky datům, které postupně sbírala z uskutečněných nákupů, byla časem schopna odhadovat zákazníkovy oblíbené tituly a žánry a navrhovat tak tituly, které by zákazník sám třeba ani neobjevil, ale přesně zapadaly do jeho stylu (Big Data on AWS). Nyní je Amazon jednou z největších firem na světě, která prodává nepřeberné množství druhů zboží, díky čemuž je tato analýza zákaznického chování ještě důležitější. Když zákazník nakupuje na Amazonu, neustále se mu zobrazují produkty, které by se mu mohly líbit. Jsou to výrobky, které Amazon vybral buď podle předešlých nákupů tohoto zákazníka, nebo podle toho, co si kupují ostatní zákazníci. Tento systém pomohl Amazonu stát se jednou z největších firem světa. Podobných technologií využívá i Google a další firmy, které mají obrovské zdroje dat. Další věcí, kterou lze pomocí analýzy dat odhalit, je to, co mění zákazníkovo chování, a co je pro něj tím rozhodujícím kritériem nákupu. Při odhalení těchto kritérií se práce prodejců značně zjednodušuje, jelikož vědí, jak přesně na zákazníky cílit, jaké vlastnosti produkt nesmí, a naopak musí mít a spoustu dalšího. Big data mají v marketingu nespočet využití, a to jak ve velkých nadnárodních firmách, tak již i v menších

lokálních firmách, protože v sobě skrývají velký potenciál ke zvýšení tržeb a snížení nákladů.

Big Data lze například použít i v politice. Vládní organizace je mohou používat na zlepšení účinnosti co se týče produktivity, cen a inovací. Big data však mohou být využita i ve prospěch jedince. Před prezidentskými volbami v USA v roce 2013 začala demokratická strana využívat data ke svému prospěchu. Pomocí big data analýz byli datoví technici schopni celému volebnímu týmu pomoci identifikovat potenciální voliče a také byli schopni přijít na to, jak přesvědčit větší počet lidí zúčastnit se voleb (Executive Office of the President, 2016). Právě při volbách v roce 2013 byla tato data považována za největší výhodu týmu kolem Baracka Obamy proti týmu Mitta Romneyho (Executive Office of the President, 2016).

Dalším neodmyslitelným polem, kde se big data sice již využívají, ale zdaleka ne naplno, je zdravotnictví. Podle Stroma (2013) se analýza dat ve zdravotnictví obecně využívá ke zlepšení bezpečnosti, účinnosti a efektivnosti. Podle Institutu Medicíny (Strome, 2013) ročně v amerických nemocnicích zemře mezi 44 a 98 tisíci pacientů na chyby zdravotníků. Kromě toho, že kvůli těmto chybám umírají ročně desetitisíce lidí, tak tyto odstranitelné chyby také stojí peníze. Podle Stroma (2013) vyjdou ročně v USA na 17 až 29 miliard amerických dolarů, což je částka, kterou pokud by se podařilo ušetřit, tak by společně se zachráněnými životy mohla přinést společnosti mnoho benefitů. Tyto a i jiné, méně závažné chyby Strome (2013) věří, že lze správným využitím dat eliminovat. Je třeba dodat, že většina chyb není způsobena nezkušeností, zbrklostí, či zanedbáním, ale jsou způsobeny kombinací mnoha faktorů a nedostatečnou bezpečností zdravotního systému jako celku (Strome, 2013). Data vygenerovaná ve zdravotnictví mohou pomoci získat mnohem hlubší pochopení nejen velkého počtu zákroků a procedur, ale i například toho, jak snížit pravděpodobnosti návratu pacienta do nemocnice 30 dní po jeho propuštění (Strome, 2013). Velký problém zde leží v tom, že stále existuje mnoho doktorů, kteří nemají elektronické záznamy. Tito doktoři, ať jich je už sice jen zlomek, brání většímu rozkvětu big dat ve zdravotnictví (Simon, 2013). I když je například v dané zemi již většina záznamů v elektronické podobě, neznamená to však, že jsou připraveny k použití. Mnoho zdravotnických záznamů je těžké najít nebo je těžké k nim získat přístup, což nepomáhá většímu využití datových analýz v tomto oboru (Strome, 2013). Dalším problémem, který ve spojení zdravotnictví a dat vzniká, je bezpečnost informací. Celkově je však možných

výhod plynoucích z využití dat ve zdravotnictví mnoho. Například nositelné technologie přinášejí nové možnosti do oblasti monitorování stavu pacientů a seniorů. K maximalizaci využití dat na poli medicíny je však zapotřebí nejen znalosti fungování lidského organismu, ale i znalostí v oboru statistiky a informačních technologií. Ke správnému fungování týmů složených z doktorů a datových specialistů bude zapotřebí dostatečná komunikace a zastřešení vědomostních rozdílů (Strome, 2013).

3.1.5 **Problémy s Big Data**

Big data, ač přináší nepřeberné množství výhod (samozřejmě pokud s nimi je správně naloženo), přináší i spoustu problémů. Neustálá vylepšení na poli technologií i na poli big dat udržují všechny velmi bdělé, nicméně je velmi náročné s tím vším udržet krok (Marr, 2017). Podle Marra (2017) souvisí s tímto oborem hlavně tři sporné oblasti, kterými jsou soukromí dat, jejich bezpečnost a datová diskriminace.

Již v roce 1791 se lidé v Americe zajímali o soukromí, když bylo čtvrtým pozměňovacím návrhem americké Konstituce lidem dopřáno rozumné očekávání soukromí (Marr, 2017). Lidé, kteří tento návrh sestavili a podepsali, však těžko mohli tušit, jaké komplikace v tomto směru přinesou moderní technologie 21. století. Nemohli tušit, že v dnešní době za sebou každý uživatel internetu nechává takovou stopu, že v podstatě z jeho života zbývá velmi málo soukromých věcí. Kvůli technologiím na rozpoznávání obličeje by se člověk neschoval ani kdyby přestal využívat jakékoliv technologie (Marr, 2017), banka má přehled o jakékoliv transakci platební kartou či na internetu, díky lokačním službám se ze smartphonu vlastně stalo sledovací zařízení apod. – jinými slovy, pokud si chce člověk v moderním světě uhlídat svoje soukromí, má to velmi těžké. Většina dat je sice používána na velmi triviální věci, nicméně z takového objemu dat, co je dnes k dispozici, se dá získat i spoustu informací lehce použitelných k páchání zla (Marr, 2017). Vlády všude po světě si pomalu začínají uvědomovat potenciál skrytý v datech a snaží se tak regulovat pravidla soukromí na internetu. V USA je toto téma stále předmětem diskuzí, kdy se Kongres společně s Federální komunikační komisí snaží udělat povinnost pro poskytovatele internetu, kteří by měli zákazníkům sdělovat, jaká data jsou ukládána a na co budou použita (Marr, 2017). Naproti tomu EU už je v tomto směru dále, jelikož již v roce 2018 má vstoupit v platnost nové nařízení pod názvem Obecné nařízení o ochraně osobních údajů (anglická zkratka GDPR – General Data Protection Regulation). Cílem tohoto nařízení je dát občanům

kontrolu nad jejich osobními daty zpět do jejich rukou. Nařízení se bude týkat všech firem držících informace o občanech EU. Podle Marra (2017) je „transparentní a etické používání dat vitální, a to nejen proto, že to je to správné jednání, ale i proto, že se blíží tvrdší regulace". Firmy by se také měly snažit být co nejvíce transparentní a dávat svým zákazníkům vědět, jaká data a z jakého důvodu jsou uchovávána, díky čemuž si firma může u svých zákazníků vypěstovat důležitou důvěru.

Dalším problémovou oblastí je bezpečnost dat. To, že zákazník souhlasí se zpracováním jeho osobních dat, neznamená, že se s těmito daty nemůže stát něco, k čemu vůbec určeny nebyly. Tím, jak čím dál více roste objem dat a síť propojených zařízení, roste také riziko narušení bezpečnosti (Marr, 2017). Ne všichni mají přístup k velkým datovým souborům obsahujícím informace, které mohou přinést zásadní objev či pouze trendy zákazníkova chování. Čím více dat společnost má, tím lépe musí mít vyřešenou bezpečnost těchto dat. V dnešní době jsou hackerské útoky skoro na denním pořádku a dovednosti hackerů se neustále zlepšují. Z toho důvodu je také nutné neustále zlepšovat datovou bezpečnost, neboť únik dat by pro firmy mohl mít katastrofální následky. Podle Chena et al. (2012) velké firmy v roce 2012 utratily za počítačovou bezpečnost 38,2 miliardy USD a malé a střední firmy za bezpečnost utratily více než za jakýkoliv jiný aspekt IT. Únik samotných dat by v tuto chvíli byl asi tím nejmenším, protože legislativní dopady takovéhoto úniku by mohly společnost poslat do velkých finančních potíží. Problémem je, že v dnešní době stále není dostatek profesionálů s potřebnými dovednostmi, kteří by byli schopni podniku zaručit bezpečnost jejich dat (Marr, 2017). Podle článku Chena et al. z roku 2012 bude koncem roku 2018 v USA chybět 140 až 190 tisíc lidí s dobrými analytickými znalostmi. Řešení tohoto problému se může zdát lehce paradoxní, neboť se nachází právě uvnitř dat samotných, respektive v jejich analýze. Tou správnou analýzou by se daly hrozby detekovat a následně i ochránit (Marr, 2017).

Datová diskriminace je posledním z velkých problémů spjatých s big data podle Marra (2017). Nejedná se o diskriminaci dat, nýbrž o diskriminaci na základě zjištění získaného z dané analýzy datového souboru. V USA se již v dnešní době používá kreditové skóre k rozhodnutí o tom, kdo si může půjčit peníze, nemluvě o pojišťovnictví, ve kterém se na základě datových analýz určuje výše pojistného. Je tedy možné, že v budoucnu, pouze na základě analýz, nebudou určití lidé například přijímáni do škol. Takovýchto příkladů by bylo možné udat mnoho, nicméně je jasné, že to je budoucnost, která nepočítá s lidským

charakterem, je to budoucnost, ve které jsou všichni lidé zaškatulkovaní a jejich život je předpovězen analýzami. Podle Marra (2017) je důležité, aby firmy u svých dat zajistily to, že soubor dat je pouze reprezentativním vzorkem zákazníků, algoritmy upřednostňují spravedlivost, jsou si vědomy předsudků přítomných v datech a srovnávají výsledky svých analýz s tradičními statistickými praktikami.

3.2 Statistické metody ke zpracování dat

3.2.1 Rozdělení dat

Ve statistice existují dva způsoby, jak rozdělit data. Tím prvním je rozdělení dat na kvalitativní a kvantitativní. Kvalitativní data slouží k slovnímu a kategorickému popisu kvality věcí a jevů. Různé kategorie mohou a nemusí mít vnitřní strukturu. Pokud se neobjevuje žádná struktura, jedná se o nominální kategorie (například rasa, pohlaví, atd.). Pokud je v dané kategorii struktura, jedná se o ordinální kategorii. Mezi tyto patří například ukazatele velikosti (malý, střední, velký) nebo ukazatele souhlasu (rozhodně nesouhlasím až po rozhodně souhlasím). Kvantitativní data jsou jednoduše taková data, která jsou vyjádřena čísly. Kvantitativní data jsou pokaždé asociována s nějakou škálou či měřítkem.

Druhým způsobem, jak lze rozdělit data, je na primární a sekundární data. Primární data jsou taková data, která byla shromážděna pro jeden přesně daný cíl. Jinými slovy někdo tato data sesbíral přímo od zdroje těchto dat. Data jsou shromažďována lidmi, kteří se občas přímo podílí na daném výzkumu a jsou tak motivováni, aby byl výzkum úspěšný a mohou tak manipulovat výsledky. Příkladem, jak získat primární data může být například dotazník. Druhým typem dat jsou data sekundární. Toto jsou data, která byla sesbírána k jinému účelu. V praxi to znamená, že primární data jednoho výzkumu jsou sekundárními daty výzkumu druhého. Zdrojem sekundárních dat mohou být například knihy, ale i dotazníky, které byly použity už někde jinde.

3.2.2 Souhrnné statistiky

Souhrnné statistiky jsou základními statistickými operacemi, které lze s daným souborem dat provádět. Ač jsou základní, často toho však jsou schopny o daném souboru říci nejvíce. Do souhrnných statistik patří měření centrální tendence a měření rozptylu. Patří sem měření centrální tendence pomocí průměru, mediánu, modusu a dalších, a měření rozptylu pomocí variačního rozpětí, rozptylu a směrodatné odchylky.

3.2.3 **Rozdělení**

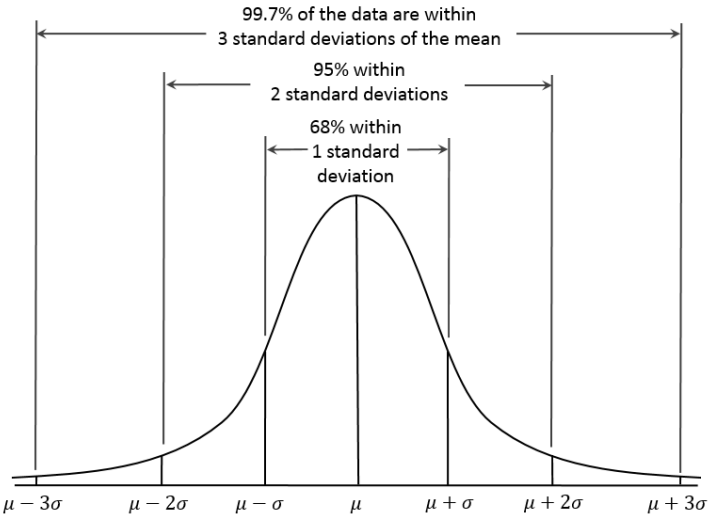
Rozdělení, nebo také distribuce, pomáhají znázornit to, jaké hodnoty se v daném vzorku nebo populaci vyskytují nejčastěji. Toto grafické znázornění zobrazuje všechny body, které se v daném vzorku či populaci vyskytují. Je to tedy velmi užitečný nástroj, který je schopen rychle a jednoduše poskytnout informace o tom, jak jsou prvky rozděleny. Existují dva základní typy rozdělení, které pak mají mnoho poddruhů. Těmi dvěma základními jsou diskrétní rozdělení a spojité (kontinuální) rozdělení (Lehman, 2005).

Diskrétní rozdělení se používá pro zobrazení diskrétních dat, jako jsou například odpovědi ano/ne či různé škály. Pro zajímavost, mezi diskrétní data se neřadí světová populace, i když by se na první pohled mohlo zdát, že to je přesně dané číslo. Je to z toho důvodu, že se populace neustále mění, a tak je považována za kontinuální veličinu, která se pohybuje v určitém intervalu (Hendl, 2015). Mezi diskrétní rozdělení patří například Bernoulliho rozdělení, které se zabývá jevy, které mohou mít pouze jeden ze dvou možných výsledků, při čemž každý má danou pravděpodobnost výskytu. Jako příklad lze použít hod mincí. Padne orel? Množinou možných výskytů tohoto jevu je buď ano nebo ne, oba jevy mají přitom stejnou pravděpodobnost výskytu 50 %. Na Bernoulliho rozdělení navazuje binomiální distribuce, která se na rozdíl od předešlého rozdělení zabývá otázkou kolikrát z určitého počtu pokusů bude jev úspěšný. Jako příklad lze opět využít hod mincí – kolikrát z 10 hodů padne orel. Dalším zástupcem diskrétních rozdělení je Poissonova distribuce, která navazuje na binomiální. Podobně jako v předchozím typu rozdělení, předmětem zkoumání je to, kolikrát se daný jev vyskytne. Oproti binomiální, která počítá s omezeným časem, Poissonova distribuce počítá, kolikrát se daný jev objeví během předem nespecifikovaného časového či prostorového rámce. Jinými slovy, není pevně dáno n. V předchozích dvou typech rozdělení se k počítání používala pravděpodobnost, v tomto případě se používá veličina λ, která představuje průměrný nebo očekávaný počet výskytů jevu během daného experimentu (Hendl, 2015).

Kontinuální rozdělení popisuje pravděpodobnosti možných hodnot kontinuální náhodné proměnné, což je proměnná, jejíž množina možných hodnot je nekonečná a nepočitatelná. Z tohoto důvodu se pravděpodobnosti těchto proměnných počítají jako obsah plochy pod křivkou, z čehož plyne, že je nutné využívat rozmezí, ve kterém se daná proměnná nachází. Pokud by se počítala pravděpodobnost jednoho čísla, byla by to vždy nula. Jedním z příkladů kontinuálního rozdělení je uniformní distribuce. Jak už název

napovídá, jedná se o distribuci, jejíž pravděpodobnosti jsou totožné na každém bodu intervalu. Mezi tento druh distribucí se také řadí normální (Gaussovo) rozdělení, které je tím vůbec nejvyužívanějším rozdělením. Slovo normální zde pak neznamená běžné či obyčejné, nýbrž se vztahuje ke staršímu významu tohoto slova – řídící se zákonem, předpisem, nebo modelem (Hendl, 2015). Hlavním předpokladem normálního rozdělení je to, že pozorování jsou seskupeny kolem průměru a čím více se hodnoty vzdalují od průměru, tím jich je méně. To, jak moc jsou hodnoty vzdáleny od průměru, se počítá pomocí rozptylu, který je značen σ2.

Za pomoci rozptylu je pak možné spočítat směrodatnou odchylku (σ), která je druhou odmocninou rozptylu. Pomocí směrodatné odchylky pak lze určit intervaly, které rozdělují všechny prvky množiny podle vzdálenosti od průměru. 68,27 % všech prvků množiny se nachází jednu směrodatnou odchylku od průměru, 95,45 % dvě směrodatné odchylky a 99,70 % všech prvků se nachází ve vzdálenosti do tří směrodatných odchylek od průměru.



Obrázek 2 Křivka normálního rozdělení. Zdroj: Kerlner, 2014.

Křivka, která tvoří graf normálního rozdělení, se nazývá Bellova křivka, nicméně tato křivka je grafem pro více rozdělení než jen normální.

3.2.4 **Testování hypotéz**

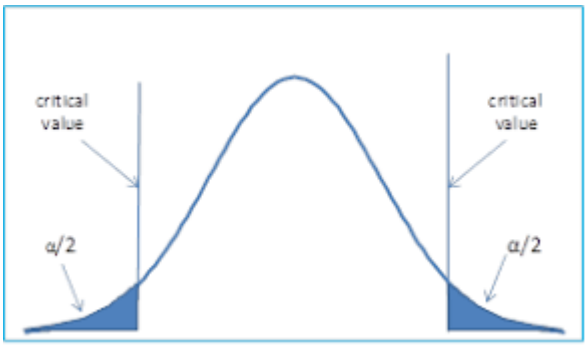
Práce s daty nezahrnuje pouze počítání základních statistik a zkoumání toho, jak jsou data v souboru rozdělena. Testování hypotéz je to, o co ve statistice jde – dokázat nebo vyvrátit nějakou hypotézu o daném souboru dat. K tomu slouží testování statistických hypotéz. Testů existuje velká řada a je proto důležité vědět, kdy jaký použít. V první řadě je důležité, jakého typu data jsou (kvalitativní či kvantitativní). Kvantitativní data mají tři základní charakteristiky – centrální tendence, rozpětí (spread) a tvar, což pomůže určit druh potřebného testu (Lehman, 2005).

Obecným důvodem pro testování je určení toho, zdali je daná hypotéza extrémně nepravděpodobná na základě vypozorovaných dat. Existují dva přístupy k testování, a to testování významnosti a testování hypotézy. První zmíněný přístup, který byl prosazován Ronaldem Fischerem, se snaží kvantifikovat důkazy proti tomu, že je daná hypotéza pravdivá, zatímco druhý přístup, prosazovaný pány Jerzy Neymanem a Egon Pearsonem, se soustřeďuje na důkazy toho, že daná hypotéza je pravdivá (Penn State). V podstatě se jedná o to samé, kdy je cílem najít důkazy pro a proti dané hypotéze.

Testování lze rozdělit do několika kroků. V prvním kroku je zapotřebí zvolit nulovou hypotézu. Nulová hypotéza je obecným tvrzením, které říká, že mezi dvěma pozorovanými jevy či mezi dvěma skupinami neexistuje žádný vztah nebo mezi více skupinami není rozdíl. Z tohoto vyplývá, že odmítnutí nulové hypotézy a přijmutí alternativní hypotézy je cílem celého statistického testování. Nulová hypotéza se obecně považuje za pravdivou, dokud důkazy neurčí jinak. Současně s určením nulové hypotézy je také zapotřebí určit tu alternativní. Alternativní hypotéza je vlastně všechno, co není nulová. Alternativní hypotéza je formulována jako neplatnost nulové hypotézy. Správné formulování hypotéz je pro výsledek celého testování kritické, poněvadž i při správných propočtech se bez správně formulovaných hypotéz ke správnému výsledku není možno dopracovat. Při určování hypotéz je také nutné brát v potaz, že ne vždy je možné dělat testování na celé populaci, tudíž se nejčastěji pracuje s náhodnými vzorky (Lehman, 2005). Při vyhodnocování hypotéz může dojít také ke dvěma druhům chyb, jelikož neexistuje žádná jistota, že chyba nebude udělána. Chyba prvního druhu, anglicky Type I Error, nastává v situaci, kdy je daná hypotéza pravdivá, ale je špatně označena za chybnou. Chyba druhého druhu, Type II Error, nastává při přesném opaku neboli když je hypotéza nepravdivá, nicméně je označena jako

pravdivá. Frekvence těchto chyb je pro statistické vyhodnocování velmi důležitá a mělo by být snahou vyhodnocovatelů udržet ji co nejnižší (Dixon, 1957).

Rozhodnutí, jestli danou hypotézu přijmout nebo vyvrátit, se zakládá na informacích zjištěných během pozorování a na míře rizika, že je rozhodnutí špatné. Této míře rizika se říká hladina významnosti, která se značí řeckým α. Tato hladina udává vzácnost výskytu dané hodnoty, která, za předpokladu že je daná hypotéza pravdivá, by se vyskytla velmi zřídka. Tuto hodnotu si určuje ten, kdo danou analýzu provádí. Většinou se volí tím menší, čím důležitější je analýza, neboť špatné vyhodnocení (type I error) by mohlo mít špatné následky. Jinými slovy, hladina významnosti α udává pravděpodobnost výskytu type I error. Hladina významnosti β pak udává pravděpodobnost výskytu druhého typu chyby neboli přijetí hypotézy, která je nepravdivá (Dixon, 1957).



Obrázek 3 Kritické hodnoty. Zdroj: Zaiontz, 2013.

Dále je nutné určit, které hodnoty dané statistiky, nazývané kritická oblast, způsobí to, že daná hypotéza bude přijata, a které hodnoty způsobí to, že hypotéze bude zamítnuta. Následně je zapotřebí spočítat hodnotu dané statistiky, a to z vypozorovaných hodnot. V posledním kroku nastává rozhodnutí, zda hypotézu přijmout či odmítnout, které je založeno na tom, zdali se získaná hodnota statistiky nachází uvnitř či vně oblasti zvolené v předešlých krocích (Dixon, 1957).

K otestování nulové hypotézy se používá testová statistika, kterou je také možno označit jako testovací kritérium (Lehman, 2005). To, jaké kritérium je použito, závisí na typu testu a dalších faktorech. Testů se dá použít celá řada, přičemž každý se hodí na něco trochu jiného. Dalo by se říci, že těmi základními typy testů, které se pak využívají společně s dalšími druhy testů, jsou tyto tři – one sample test, two sample test a párový test (Lehman, 2005). První zmíněný test se používá pro porovnání vzorku s populací z hypotézy. Charakteristiky populace jsou v tomto případě známy z teorie nebo vypočteny z populace.

Druhý test se používá pro porovnání dvou vzorků, například při porovnávání experimentálního a kontrolního vzorku. Párové testy se také používají při srovnávání dvou vzorků, ale tam, kde není možné kontrolovat důležité proměnné. V tomto případě se však nesrovnávají dva vzorky, ale prvky jednoho vzorku jsou spárovány s prvky druhého a rozdíl mezi nimi vytvoří nový vzorek. Průměr těchto rozdílů se pak porovnává s nulou. Dále se využívá také z-test, který se používá při porovnávání průměrů s přísnými podmínkami co se týče normality a daného normálního rozdělení. Naproti tomu t-test se používá při porovnávání průměrů s volnějšími podmínkami (Lehman, 2005). Dalším druhem testu používaným k vypočtení testové statistiky je Chí kvadrát test, který má tři možnosti využití. V první řadě jej lze využít pro určení, zdali má daná populace přesně daný rozptyl. V tomto případě nulová hypotéza říká, že má. Druhá možnost využití Chí kvadrátového testu je k ověření nezávislosti mezi dvěma proměnnými, které zpravidla bývají spíše kategorické než číselné. Nulovou hypotézou je v tomto případě nezávislost mezi proměnnými. Posledním využitím tohoto testu je, jestli křivka adekvátně reprezentuje data neboli test dobré shody (nulová hypotéza je, že ano) (Lehman, 2005). F – test je také velmi rozšířeným testem u statistických analýz. Používá se pro analýzu rozptylu, často se také označuje ANOVA, a určuje, zdali je seskupování dat do kategorií smysluplné (Lehman, 2005). Jinými slovy, F-test určuje, jestli je rozdíl mezi rozptyly dvou skupin uvnitř populace či vzorku. Nulová hypotéza je v tomto případě shoda rozptylů.

Jakmile je zvolen vhodný test, lze spočítat testovací kritérium z vypozorovaných dat. Následně je nutné najít kritickou hodnotu pro dané testovací kritérium při zvolené hladině významnosti. Tuto hodnotu lze nalézt ve statistických tabulkách a udává minimální hodnotu testovacího kritéria k zamítnutí nulové hypotézy (Penn State).

Jakmile jsou známy obě hodnoty testovacího kritéria, spočítané a tabulkové, je možno vyhodnotit celé testování. Pokud je hodnota spočítaného kritéria větší než hodnota tabulkového, nachází se daná hodnota v oblasti zamítnutí, a je tedy možno zamítnout nulovou hypotézu s hladinou významnosti 1-α (Penn State). Moderní statistika celý proces zjednodušuje zavedením p-hodnoty, což je maximální hodnota pravděpodobnosti chyby prvního druhu. Pro vyhodnocování analýzy pomocí p-hodnoty platí následující pravidla: pokud je p-hodnota menší než hladina významnosti α, lze zamítnout nulovou hypotézu a přijmout alternativní (Penn State).

3.2.5 **Regresní analýza**

Podle Montgomeryho et al. (2011) je regresní analýza jednou z nejrozšířenějších technik k analýze více faktorových dat. Použitelnost této techniky pramení převážně z logického procesu, ve kterém je využito rovnic k vyjádření vztahu mezi zvolenou proměnnou a množinou souvisejících prediktorových proměnných (Montgomery, 2011). Jinými slovy, regresní analýza se používá ke zjištění toho, jak se změní závislá proměnná při změně jedné nezávislé proměnné, ceteris paribus. Regrese se často používá k předpovídání vývoje a je hojně využívána ve strojovém učení, což je v poslední době mohutně rozvíjející se obor. V dnešní době už tento druh analýzy zvládají i ty nejzákladnější softwarové programy, jako například Microsoft Excel, nicméně je dobré vědět, jak jednotlivé metody analýzy fungují. Existuje mnoho druhů regresních modelů, jako například lineární, logistické, dále se pak také často využívá rozhodovacích stromů a neuronových sítí.

Jednoduchý lineární regresní model je tím relativně nejzákladnějším druhem regresních modelů. V tomto modelu je předpokladem, že závislá proměnná je lineární kombinací parametrů. Zápis této nejjednodušší formy regresního modelu vypadá následovně:

$$y=\beta_0 +\beta_1 x_1 +c,$$

Rovnice 4 Rovnice regresního modelu. Zdroj: Vlastní zpracování.

kde $\beta_0 +\beta_1 x_1$ je deterministická část modelu a c je stochastická část (Freund et al., 2006). Často se také parametrům β říká regresní koeficienty a stochastické části náhodná složka. U náhodné složky se počítá s jistými předpoklady, a to že její průměr je roven nule, rozptyl je konstantní a je normálně rozdělena (Freund, 2006). Cílem celé regresní analýzy je určit to, jak se změní y v závislosti na x neboli udělat závěry o závislé proměnné při využití informací z nezávislých proměnných (Freund, 2006). Ne vždy je však regresní analýza dostatečně průkazná a je třeba udělat korelační model, který popisuje závislost mezi proměnnými. Korelační model slouží k určení toho, zdali mají dvě proměnné dvourozměrné normální rozdělení. Toto rozdělení je definováno pěti parametry: průměry x a y, rozptyly x a y a korelačním koeficientem r, který měří sílu lineárního vztahu mezi danými proměnnými (Freund, 2006). Korelační koeficient nabývá hodnot mezi -1 a 1 včetně, kde negativní hodnota koeficientu značí nepřímou úměru a pozitivní hodnota přímou (Montgomery, 2011), a čím více se hodnota koeficientu blíží svému maximu, tím je lineární vztah silnější. Naopak čím blíže je nule, tím slabší je pozorovaný vztah mezi proměnnými (Freund, 2006). Lineární

regresní model však může obsahovat i více nezávislých proměnných a v takovém případě se nazývá vícenásobným. Jeho podstata je stejná jako u jednoduchého, nicméně výpočty jsou u tohoto modelu náročnější a často se vyskytují i vztahy mezi nezávislými proměnnými, které vznikají jako výsledek multikolinearity (Freund, 2006). K odhadování parametrů β se často využívá běžné metody nejmenších čtverců.

Ne vždy je možno použít lineární regresní model a je nutno sáhnout po jiném, a to například při práci s kvalitativními proměnnými, ať už di – či poly – chotomickými (Freund, 2006). V takovémto případě je tedy nutno využít logistický regresní model, což je model křivočarý (Freund, 2006). Logistická regrese je model, který je založen na matematicky zaměřeném přístupu, který je určen k analýze toho, jak na sebe vzájemně působí proměnné (Hassouna et al, 2015). Predikce je pak založena na sestavení rovnic, které spojují vstupní hodnoty s výstupním polem. Hassouna et al. (2015) ve svém článku využívají logistickou regresi k předpovězení zákazníkova chování, kde vstupními hodnotami jsou faktory ovlivňující zákazníkovo chování a výstupem je pak pravděpodobnost zvoleného chování. V tomto případě logistická regrese předpovídá pravděpodobnost odchodu zákazníka formulováním rovnic, určením vstupních hodnot, určením faktorů ovlivňujících toto chování a určením výstupního pole, kterým je pravděpodobnost odchodu (Hassouna et al, 2015). Pro logistickou regresi udávají Hassouna et al. (2015) tři rovnice:

$$p\,(y = 1\,|x1, \ldots, xn) = f(y)\ (1)$$

$$f(y) = 1\ (1+e{-}y)\ (2)$$

$$y = \beta 0 + \beta 1\ x1 + \beta 2\ x2 + \cdots + \beta n\ xn\ (3),$$

Rovnice 5 Rovnice logistické regrese. Zdroj: Hassouna et al, 2015.

kde:

- y je závislou proměnnou, v tomto případě binární,
- β_0 je konstantou,
- β_i je váha přidělená specifické proměnné x_i
- x_1, …., x_n jsou nezávislé proměnné.

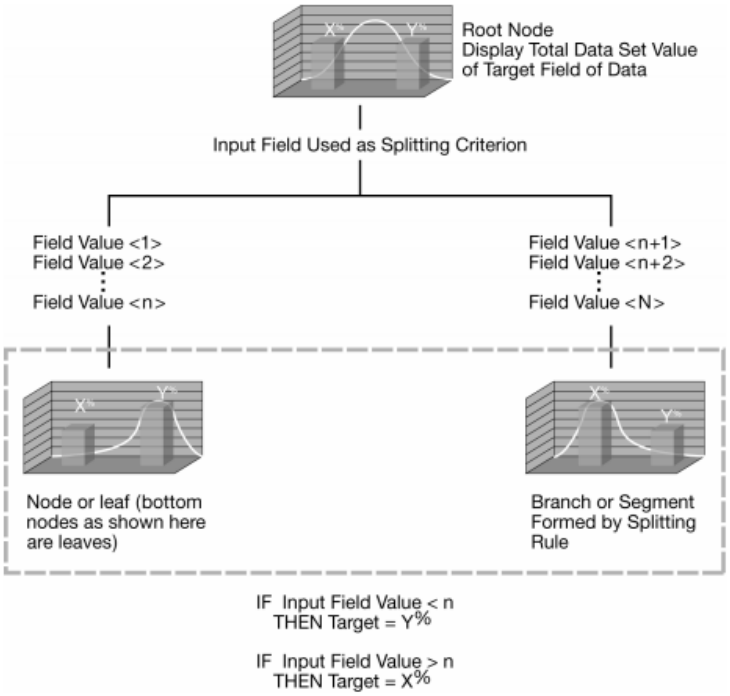
Poté nastává analýza dat, která vede k vytvoření regresních rovnic, na jejichž základě je třeba vykonat hodnotící proces (Hassouna et al, 2015). V tomto případě, kdy je cílová proměnná binární, se hodnotí, jestli je p-hodnota u jednotlivých záznamů větší než předdefinovaná hodnota. Při práci s logistickou regresí je třeba brát v potaz multikolinearitu,

která může vést k nesprávným závěrům ohledně vztahů mezi závislými a nezávislými proměnnými (Hassouna et al, 2015). Interpretace koeficientů je v případě logistické regrese složitější, než tomu je například u regrese lineární. Vzhledem k nelineárnímu tvaru funkce nelze totiž brát velikost koeficientu jako konečný směr a velikost působení nezávislé proměnné na cílovou proměnnou (Řeháková, 2000). Podle Řehákové (2000) koeficienty β určují změnu logitu spojenou s nárůstem proměnné spojené s tímto koeficientem o jednu jednotku, ceteris paribus. Z tohoto důvodu je nutné počítat šance jevu namísto jeho pravděpodobnosti. To, jak daná nezávislá proměnná působí na závislou proměnnou, lze zjistit pomocí matematické operace exp (β). Exp (β) pak udává velikost násobku, o který se změní šance, jestliže se hodnota nezávislé proměnné změní o jednu jednotku, ceteris paribus (Řeháková, 2000). Pokud je $\beta > 0$, šance se zvýší, pokud je $\beta < 0$, šance se sníží (Řeháková, 2000).

3.2.6 **Rozhodovací stromy**

Dalším často využívaným postupem při práci s daty jsou takzvané rozhodovací stromy. Je to vlastně jednoduchá forma multivariátní analýzy, která má jedinečnou schopnost doplnit či nahradit tradiční statistické analýzy, jako je například vícenásobná lineární regrese, či širokou škálu data-miningových nástrojů a technik, jako jsou například neuronové sítě (SAS, 2008). Podle Hassouny et al. (2015) jsou rozhodovací stromy tou nejpopulárnější metodou pro prediktivní modelování a jsou to vlastně grafy, které prezentují vztahy mezi proměnnými. Tyto rozhodovací stromy jsou většinou výsledkem algoritmu, který identifikuje různé možnosti, jak rozdělit soubor dat do segmentů podobných větvím stromu (SAS, 2008). Nejčastěji používanými algoritmy jsou CART, C5.0 a CHAID (Hassouna et al., 2015). Jednotlivé segmenty určené algoritmem pak vytvoří obrácený strom, kde kořenový uzel je na vrcholu stromu a tento hlavní uzel v sobě obsahuje objekt analýzy (SAS, 2008). Která proměnná bude obsažena v hlavním uzlu se určí podle toho, jaký je poměr jejího informačního zisku (Hassouna et al., 2015). Ta proměnná, jejímž modelováním se získá nejvíce informací, bude náležet do tohoto uzlu (Hassouna et al., 2015). Získání rozhodovacího pravidla, pomocí kterého je možno dále tvořit segmenty a větve pod hlavním uzlem, je založeno na metodě, která extrahuje vztah mezi objektem analýzy, který slouží jako cílové pole v datech, a jedním či více poli, která slouží jako vstupní pole k vytvoření segmentů a větví (SAS, 2008). Hodnoty ve vstupním poli jsou pak použity k odhadu

pravděpodobné hodnoty výsledku či odezvy (SAS, 2008). Jakmile je získán vztah, je možné odvodit jedno či více pravidel, která popisují vztah mezi vstupy a cílovými hodnotami (SAS, 2008). Tyto pravidla pak lze použít k zobrazení rozhodovacího stromu, který pomáhá vizuálně analyzovat síť vztahů, které charakterizují vstupní a cílové hodnoty. Stromy jsou reprezentovány a hodnoceny shora dolů a rozhodnutí se provádí posunováním z hlavního uzlu ke všem listům stromu (Hassouna et al., 2015). Rozhodovací strom lze tedy využít k předpovědi hodnot nových nebo předem neviděných pozorování, která obsahují hodnoty pro vstupy, ale nemusí obsahovat hodnoty pro cíle (SAS, 2008).

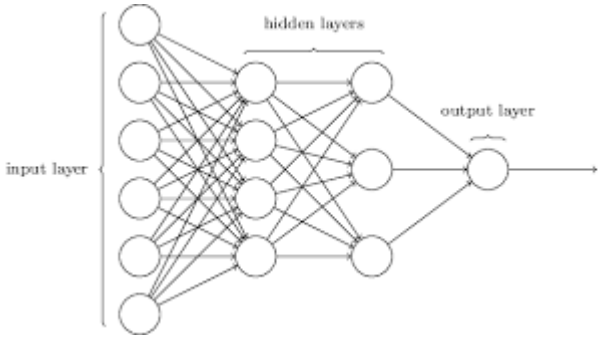


Obrázek 4 Příklad rozhodovacího stromu. Zdroj: SAS, 2008

Oproti logistické regresi mají rozhodovací stromy několik výhod. Stromy je jednoduché na vizualizaci a jsou snadné k porozumění, o datech není třeba mít nějaké předchozí předpoklady, jelikož je využit neparametrický přístup, a mohou pracovat jak s numerickými, tak s kategorickými daty (Hassouna et al., 2015). Hassouna et al. (2015) ve svém článku však zmiňují i nevýhody rozhodovacích stromů, mezi které patří absence robustnosti modelu, fakt, že výkon modelu je velmi ovlivněn komplexními interakcemi mezi proměnnými a atributy a v neposlední řadě také složitá vizualizace a porozumění při práci s komplexními stromy.

3.2.7 **Neuronové sítě**

Neuronové sítě se také řadí do metod pro zpracování dat. Podle Manyiky (2011) jsou to početní metody, které byly inspirovány skutečnou strukturou a fungováním biologických neurálních sítí, které hledají vzory v datech. Hodí se zejména k hledání nelineárních vzorů a lze je například použít k identifikaci důležitých zákazníků, u kterých je riziko, že přejdou ke konkurenci (Manyika, 2011). Pro svoji podobnost s lidskou nervovou soustavou se neuronové sítě často používají v umělé inteligenci. Například Apple začal v modelech iPhonů 8 a X používat bionický čip, který je právě díky neuronovým sítím schopen se sám učit z návyků svého majitele a lépe tak reagovat. Základem neuronové sítě jsou umělé neurony, které jsou mezi sebou spojeny synapsemi schopnými přenosem signálu, stejně jako v lidské nervové soustavě (SAS, 2008). Síť pak funguje tak, že synaptickým signálem je reálné číslo a výstup z každého neuronu je spočítán pomocí nelineární funkce (SAS, 2008). Jednotlivé neurony a synapse také mohou mít různou váhu, která se mění v závislosti na pokroku strojového učení, což pak může zvýšit nebo snížit sílu signálu pouštěného dále (SAS, 2008). Vstupů může být několik, výstup je však z jednoho neuronu pouze jeden (SAS, 2008).



Obrázek 5 Neuronová síť. Zdroj: Nielsen, 2017.

3.2.8 **Data Mining**

Data mining (dále DM), jakožto další metoda pro práci s big data, je podle Nisbeta et al. (2009) procesem extrahování implicitních, předem neznámých a potenciálně užitečných informací z dat. Je nutné rozlišovat mezi statistickou analýzou a data miningem. Statistická analýza používá model k charakterizování určité struktury (pattern) v datech,

zatímco DM využívá struktury v datech k sestavení modelu (Nisbet et al., 2009). Dalším rozdílem mezi těmito dvěma technikami je to, jakým způsobem se pracuje s daty. Tradiční statistická analýza využívá minulých informací k určení budoucího stavu; naproti tomu DM využívá minulé informace k tvorbě vzorců založených nejen na vstupních datech, ale také na logických důsledcích těchto dat (Nisbet et al., 2009). Nisbet et al. (2009) zformulovali definici DM takto: „Data mining je použití algoritmů strojového učení k nalezení slabých struktur vztahů mezi prvky dat ve velkých, hlučných a neuspořádaných souborech dat, které mohou vést ke zvýšení prospěchu v nějaké formě“. Podle stejného zdroje má DM dvě základní funkce, a to poskytování ucelenějšího porozumění dat pomocí hledání doposud neviděných vztahů a tvorbu prediktivních modelů, které umožňují lepší rozhodování. DM lze použít v mnoha oblastech, jako jsou například předpovědi prodeje, vědecké objevy, CRM, sport, získávání zákazníků a mnoho dalších.

Mezi hlavní aktivity ukrývající se pod pojmem data mining patří podle Nisbeta et al. (2009) těchto pět kategorií:

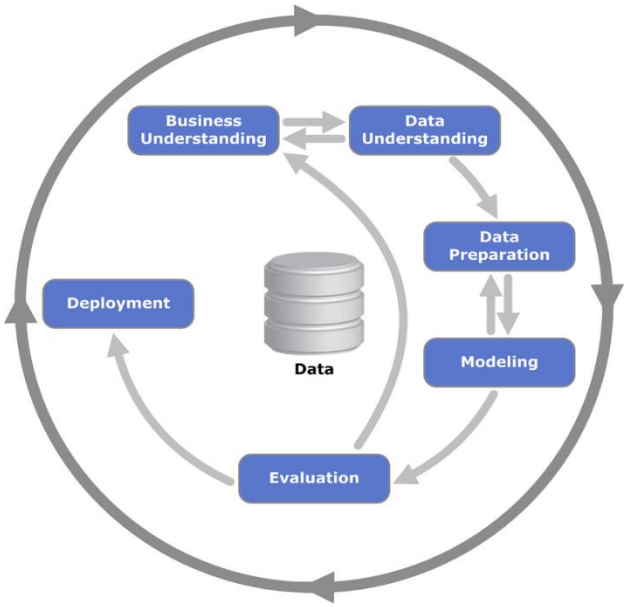
- Analýza průzkumných dat, což je pohled na datový soubor skrze souhrnné statistické parametry, ktwrý umožňuje hrubé odhalení trendů a vzorů v datech;
- Deskriptivní modelování, což už je složitější pohled na data, který může obsahovat určování celkové pravděpodobnosti rozdělení dat, modely popisující vztahy mezi proměnnými, shlukovou analýzu nebo segmentaci;
- Prediktivní modelování, které obsahuje klasifikaci, která se používá u kategorických proměnných, a regresi, která se používá u kontinuálních proměnných. Cílem je pak sestavit model, kde hodnota jedné proměnné může být předpovězena hodnotami jiných proměnných;
- Objevování vzorů a pravidel, což může být použito na vytvoření asociačních pravidel,
- Vyhledávání podle obsahu, což je proces, který začíná s již známým vzorem a snaží se najít podobné vzory v jiném souboru dat. Nejčastěji se tato metoda používá u textových materiálů nebo u obrazových souborů dat.

Pro samotný proces data miningu se nejčastěji používá CRISP-DM, což je zkratka pro Cross Industry Standard Process for Data Mining. CRISP-DM je nejkompletnější formát pro vyjádření DM procesu a celkem obsahuje šest kroků (Nisbet et al., 2009). Celý proces je velmi komplexní a vyžaduje porozumění mnoha oblastí.

Jednotlivé kroky procesu jsou popsány níže a informace pocházejí od Nisbeta et al. (2009).

- Prvním krokem v celém procesu DM je porozumění byznysové stránce procesu. V tomto kroku je důležité určit, co se bude dělat, a co je považováno za úspěch. Tento první krok má tři části. V té první je třeba nadefinovat obchodní cíle modelu, určit, co vytvořilo potřebu pro tvorbu modelu a stanovit kritéria úspěchu. V druhém kroku se posoudí okolí celého procesu (inventura zdrojů, identifikace risků a posouzení okolí, kde bude model působit). Poslední krok obsahuje formulaci cílů a úkolů DM. Je důležité si uvědomit, že primárním cílem není jen vytvořit dobrý prediktivní model, nýbrž nasadit dobrý prediktivní model, který splňuje obchodní cíle. Mezi cíle lze zařadit vytvoření vhodné databáze, ze které lze jednoduše extrahovat data pro modelování, či vytvoření modelu, který generuje významnou přidanou hodnotu. Mezi úkoly se řadí například získání vhodných dat, vytvoření listu predikujících proměnných, vytvoření modelu s přijatelnou přesností, spuštění modelu a jeho následná kontrola a aktualizace.
- Druhým krokem je porozumění dat. Tento krok je velmi důležitý a dá se rozdělit do čtyř částí V té první je pozornost soustředěna na akvizici dat, kde je nutné identifikovat dostupné zdroje dat. Ve druhé části nastává integrace dat neboli uvědomění si, že data jsou v jiném formátu, jiných jednotkách atp. Pro lepší názornost lze využít takzvané datové mapy, které pomáhají vyjádřit data ve společném formátu. Ve třetí části nastává popis dat, kde je nutné se seznámit s daty, se kterými se bude pracovat. Poslední část je zaměřena na zhodnocení kvality dat. V této části se řeší, jak vyplnit prázdná pole a jak naložit s extrémními hodnotami.
- V dalším kroku nastává příprava dat, což může být často velmi zdlouhavý, leč zásadní proces. Je zapotřebí zhodnotit a transformovat data tak, aby vytvořily soubor v požadovaném formátu. S přípravou dat může nastat řada problémů, jako například nekompatibilnost struktury dat s algoritmy či to, že data jsou uchována na účetní úrovni a jejich přeuspořádání je často velkou výzvou. Nejčastějšími operacemi při přípravě dat jsou čištění dat, transformace dat, vypořádání se s chybějícími hodnotami, filtrování dat, abstrahování dat, redukce dat a odvozování dat.

- Jakmile jsou data připravena, může nastat modelovací proces. Nejprve je nutné zvolit modelovací techniky, algoritmy a architekturu celého modelu a specifikovat modelovací předpoklady. Poté se vytvoří experimentální návrh, který musí splňovat přísné kritérium – výsledky za normálních podmínek se musejí rovnat výsledkům, kdy je model vystaven různým úpravám. Jakmile experimentální model splňuje všechny podmínky, je možné přistoupit k tvorbě samotného modelu. Zde se musí určit parametry modelu a sestavit několik typů modelů. Následně je model zhodnocen, a to tak, že se porovná k tomu, co by tvůrce modelu očekával že se stane, pokud by model nepoužil. Jinými slovy se zhodnotí, jestli má model nějakou přidanou hodnotu.
- Po sestavení modelu je nutné model řádně vyhodnotit. Tento krok je nutné uskutečnit ještě před samotným spuštěním modelu, aby se předešlo případným nedostatkům či chybám, které by bylo složité řešit v již běžícím modelu. Je třeba opět projít všechny kroky vedoucí k vytvoření modelu a také se ujistit, jestli model odpovídá obchodním cílům stanoveným v prvním kroku procesu. Důležitým úkolem v této části je určit to, zdali nebyl opomenut nějaký obchodní problém, který je zásadní ke správnosti funkce modelu. Na konci vyhodnocovací části je nutné učinit rozhodnutí o použitelnosti modelu.
- Posledním krokem data miningového procesu je tzv. deployment neboli uvedení modelu do praxe. V této fázi procesu je zapotřebí vytvořit plán na monitorování a údržbu modelu, vytvořit finální report pro zadavatele tak, aby mu rozuměl, a



Obrázek 6 Grafické znázornění CRISP-DM. Zdroj: Wikipedia.

v neposlední řadě se v této fázi hodnotí celý projekt. DM proces však touto fází nekončí, neboť je neustále nutné získávat a analyzovat zpětnou vazbu modelu a tvůrci modelu jej neustále upravují a dolaďují do té doby, dokud nejsou možná žádná další vylepšení v předpovídajících schopnostech modelu.

3.2.9 **Prediktivní modelování**

Cílem všech výše zmíněných aktivit je vytvořit model, který bude odpovídat stanoveným požadavkům. Podle Logana et al. (2016) existují dva druhy modelů, a to deskriptivní a prediktivní. Deskriptivní modely slouží k porozumění kauzálních vztahů mezi proměnnými, zatímco prediktivní modely se snaží nalézt empirické vztahy, které poskytují dobré odhady budoucího vývoje (Logan et al., 2016). Logan et al. (2016) ve svém článku píší o predikci velikosti jablek při sklizni a říkají, že deskriptivní modely se dají využít například k tomu, aby farmářům sdělily, jak vypěstovat velká jablka. Oproti tomu prediktivní modely jim jsou schopny s určitou přesností předpovědět velikost jablek při sklizni neboli předpovědět nové nebo budoucí pozorování. Podle autorů tohoto článku je nejdůležitější částí prediktivního procesu validace modelu. Tato validace má dva důležité prvky:

- Validaci napříč různými soubory dat, která se provádí na souborech dat, které nebyly použity k vytvoření modelu. Tímto souborem může být například jeden rok záznamů (pokud je k dispozici více let záznamů) a pak je tento soubor využit pro validaci modelu a ostatní roky pro tvorbu modelu. Tento proces je následně opakován tak, aby každý rok sloužil pro validaci modelu právě jednou a výsledná chyba modelu je vypočítána jako průměr chyb všech modelů.
- Druhým prvkem validace je použití mimo-souborových metrik, jako je například RMSE (root mean square error) nebo MAE (mean absolute error), což jsou ukazatele prediktivní síly modelu. Logan et al. je radí používat raději než koeficient determinace R^2, který měří, jak dobře regrese reprezentuje data.

$$RMSE = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2}{n} \qquad MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n}$$

Obrázek 7 Výpočet RMSE a MAE. Zdroj: Logan et al., 2016.

Celý proces prediktivního modelování pak lze rozdělit do několika kroků. Logan et al. (2016) ve svém článku uvádějí následující kroky:

- Úplně prvním krokem je určení cíle. Cíl prediktivního modelování určí nejen to, čeho se modelem má dosáhnout, ale i to, jakým způsobem toho bude dosahováno.
- V druhém kroku nastává shrnutí dat, která budou využita k vytvoření modelu. V této části tak nastává nejen popis dat, ale i jejich grafické znázornění pomocí grafů a také se zde odehrává nutná úprava dat. V případové studii zmíněné v tomto článku byly chybějící hodnoty jedné proměnné nahrazeny průměrem dostupných hodnot.
- V další části procesu jsou zváženy různé modely, které lze v daném případě využít. Při porovnávání modelů se podle Logana et al. hledí na dva aspekty, kterými jsou druh modelu a kombinace použitých proměnných. Preferován je takový model, který pro dosažení podobné přesnosti předpovědi využívá méně proměnných.
- Následuje validace za využití dat nevyužitých při sestavování modelu.
- Dalším krokem je výběr proměnných a určení jejich důležitosti a vlivu.
- V posledním kroku se již sestaví finální model, který bude využit pro předpovězení dané proměnné.

3.2.10 Technologie pro práci s Big Data

Stejně tak jako techniky, tak i technologie pro práci s velkými soubory se neustále vyvíjejí a vylepšují, aby byly schopny pracovat s neustále většími objemy dat, a aby usnadňovaly práci datovým specialistům. Je jich tolik, že je nelze vypsat všechny, a proto jich je popsáno jen pár. V této části je také důležité zmínit, s jakými druhy dat lze pracovat. Jsou v základě tři typy dat – strukturovaná, nestrukturovaná a semi-strukturovaná data. Při práci s prvním typem se data nacházejí v daných polích. Příkladem může být relační databáze či data v tabulkách. Nestrukturovaná data jsou opakem strukturovaných, tudíž se data nenachází v daných polích. Příkladem je volná forma textu (knihy, emaily), obrazová a video data. Semi-strukturovaná data jsou něčím mezi. Data nespadají do přesně daných polí, nicméně obsahují určité značky, které umožňují jednotlivé datové prvky oddělovat. Příkladem může být XML text (Manyika et al.., 2011). Informace o technologiích pocházejí od Manyiky et al. (2011).

- *Business intelligence* je druhem aplikačního softwaru, který je určen k hlášení, analýze a prezentaci dat. Tyto nástroje jsou používány k práci s daty, která byla již uložena v datovém skladu či trhu.
- *Cassandra* je open-sourcový systém na práci s daty vytvořen pro obrovská množství dat na distribuovaném systému. Tento systém byl původně vyvinut ve Facebooku.
- *Cloud computing* je v zásadě výpočetní systém, ve kterém jsou zdroje velkého výpočetního výkonu k dispozici jako služba skrze síť.
- *Datový sklad a trh* jsou specializované databáze optimalizované pro reporting a často jsou využívány pro uchovávání velkých objemů dat.
- *Hadoop* je dalším open-sourcovým rámcem, který je určen ke zpracovávání velkých souborů zaměřených na nějaký problém. Hadoop byl původně vyvinut společností Yahoo.
- *SQL*, zkratka pro strukturovaný dotazovací jazyk, je počítačovým jazykem pro práci s daty v relačních databázích, což jsou databáze, ve kterých jsou data uchovávána ve sloupcích a řádcích. SQL technologie umožňuje širokou škálu operací s daty, od jednoduchých jako je vložení, vymazání či dotazování dat, až po práci s databázovými strukturami.
- *Vizualizace* je používána k tvorbě obrázků, diagramů či animací k vyjádření nějakého sdělení, které je pak použito k syntéze výsledků big data analýzy. Vizualizace se stává čím dál tím více důležitějším oborem, neboť dobrá prezentace je nadmíru důležitá. Lidé mají problémy s porozuměním velkých datových či textových souborů, a přesně z tohoto důvodu je vizualizace důležitá – umožňuje lidem lepší a jednodušší pochopení. Na poli vizualizace se z těchto důvodů vědci snaží o co nejvíce vylepšení. Jednou z nových metod vizualizace je značkový mrak (tag cloud), což je v podstatě vážený vizuální list, ve kterém jsou nejčastěji

vyskytující se slova větší než ta, která se vyskytují méně často, což čtenáři umožňuje relativně rychle rozpoznat ty nejdůležitější koncepty.

Obrázek 8 Příklad vizualizace – Tag Cloud, Manyika et al., 2011.

3.3 Chování zákazníka

Chování zákazníka jsou podle Kardese et al. (2014) všechny aktivity spojené s nákupem, užíváním a likvidací výrobků nebo služeb, včetně zákazníkových emocionálních a duševních reakcí a změn v chování, které buď předcházejí nebo následují dané aktivitě. Naproti tomu Kotler (2012) ve své definici nezmiňuje reakce a změny v zákazníkově chování, ale zabývá se pouze tím, jak lidé vybírají, nakupují, užívají a zbavují se výrobků, služeb atd. Jinými slovy, studium zákazníkova chování se zabývá nákupními aktivitami, spotřebními aktivitami a likvidačními aktivitami. Nákupní aktivity zahrnují nejen samotný nákup, ale také všechny aktivity, které nákupu předcházejí, jako je například sběr informací, hodnocení alternativ, služeb a platebních možností. Spotřební aktivity zahrnují otázky kdy, kde, kdo a jak týkající se spotřeby. Likvidační aktivity zahrnují to, jakým způsobem se zákazník zbavuje produktu a jeho obalu a mohou také zahrnovat případný přeprodej použitých výrobků. Toto odvětví také věnuje velkou pozornost reakcím zákazníků. Ty mohou být tří druhů – emocionální, duševní a změny v chování. Emocionální reakce zákazníků zahrnují pocity a nálady, duševní zahrnují jejich myšlenkový pochod a změny v chování se zaměřují na pozorovatelné změny ve vztahu ke koupi daného výrobku či služby (Kotler, 2012).

V dnešní době je každý člověk spotřebitelem a zákazníkem od té doby, co se narodí, až do té doby, než zemře. Díky neustálému vystavení reklamním sdělením se lidé po celý život musí aktivně rozhodovat. Kotler (2012) říká, že spotřebitelské chování sleduje to, jak jednotlivci, skupiny a organizace vybírají, nakupují, používají a zbavují se zboží, služeb, nápadů a zkušeností k uspokojení svých potřeb a chtíčů. Oblast, kde se tato rozhodnutí odehrávají, se nazývá spotřební trh, který je tvořen lidmi, výrobky, službami a penězi (Kotler, 2012). Faktorů, které ovlivňují toto rozhodování, je hodně, počínaje duševními vlastnostmi přes zkušenosti jednotlivců až po působení druhých lidí. Tyto faktory lze rozdělit podle několika faktorů, kdy například Kotler (2012) faktory rozděluje do tří skupin – kulturní, sociální a osobní. Lze je však rozdělit do dvou základních, obecnějších kategorií, kterými jsou vnější stimuly a osobnost zákazníka. Osobnost neboli persona, je pojem pocházející z latiny, kde persona byla maska, kterou si při představení nasazovali herci (Vysekalová, 2011). V tomto smyslu je však osobnost brána jako označení pro jedinečnost každého člověka. Jde o jednotlivé rysy, vlastnosti, schopnosti, potřeby, zájmy, temperament a další charakteristiky, které jako celek určují to, čím daný člověk je. Chování zákazníka je také velmi usměrněno kulturním pozadím, kterému je důležité porozumět z pohledu marketérů. Dokonalé porozumění zákazníkova chování může být pro daný podnik obrovskou výhodou. V dnešní datové době se tyto poznatky dají využít i k modelování zákazníkova chování, z čehož plynou další užitečné pomůcky. Při úspěšném modelování zákazníkova chování lze určit, který typ zákazníků se vrací, který naopak po prvním užití již nechce mít s produktem nic společného apod. Toto všechno je možné pouze při dokonalém porozumění zákazníkova chování a dobrém využití dostupných dat.

3.3.1 Udržení zákazníka

Kvalitní analýza zákaznických dat může marketingovým specialistům pomoci i v další velmi důležité oblasti, kterou je udržování stávajících zákazníků. Udržování zákazníků se v posledních několika letech stalo hlavním faktorem prozkoumávaným CRM a obecně je v této oblasti pozornost zaměřena na tvorbu a kontrolu loajálních, profitujících a trvajících vztahů se zákazníky (Hassouna et al., 2015). Podle Kotlera (2012) je získání nového zákazníka až pětkrát dražší než udržení toho stávajícího. Z toho důvodu může společnosti poznání faktorů zapříčiňujících odchody stávajících zákazníků ušetřit spoustu peněz. Proto se současné firmy snaží nabízet svým zákazníkům co nejširší škály výrobků a

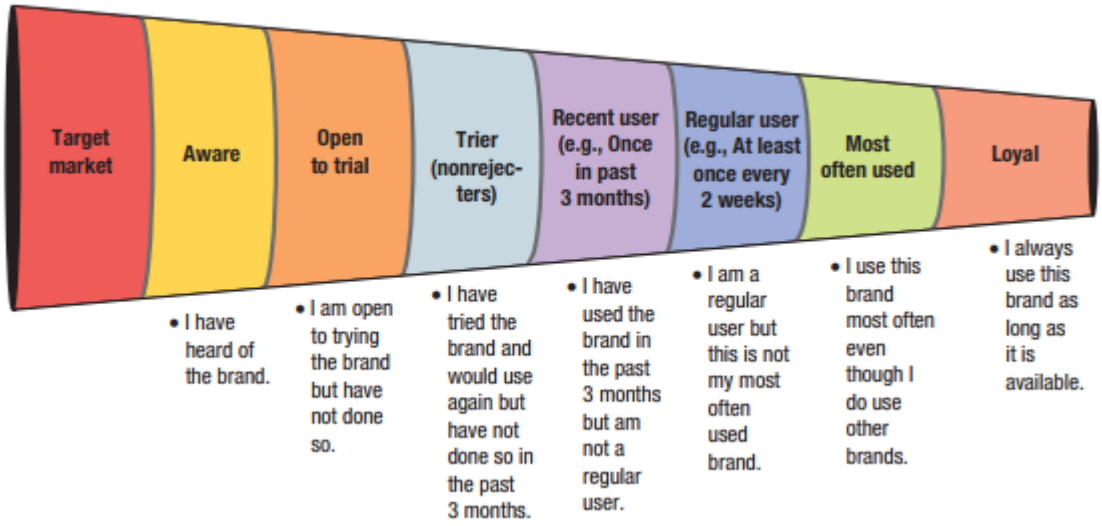
také trénují své zaměstnance v cross-sellingu a up-sellingu (Kotler, 2012). Získávání nových zákazníků je velmi důležité, nicméně udržení těch stávajících je minimálně na stejné úrovni, avšak pro některé firmy může být velmi těžké. Podle Kotlera (2012) tímto trpí převážně telefonní operátoři a poskytovatelé televizních služeb, kde někteří lidé mění poskytovatele několikrát do roka s cílem neustále využívat nejlepší nabídky na trhu. Mnoho z těchto poskytovatelů ročně ztratí až 25 % zákazníků, což vede ke ztrátám 2 až 4 miliard USD (Kotler, 2012). Hassouna et al. (2015) ve svém článku zmiňují dokonce až 40 % ztráty zákazníků u mobilních operátorů ročně, což je velmi vysoká hodnota. Nejčastějšími důvody pro změnu poskytovatele pak jsou nenaplněné potřeby a očekávání, nízká kvalita služby či zákaznického servisu, či chyby ve vyúčtování (Kotler, 2012). Aby společnost předešla vysokému počtu odcházejících zákazníků, podle Kotlera (2012) by si měla určit a spočítat míru udržení zákazníků (například počty prodloužených smluv), odhalit příčiny odchodů a pokusit se je řešit, a porovnat zákazníkovu hodnotu (customer lifetime value) s cenou za snížení míry odchodů.

Cílem společností by mělo být vybudovat širokou zákaznickou základnu, která je však věrná. Nejen že udržet zákazníky je levnější než získávání nových, ale věrný zákazník také mnohem více odolává tlakům odchodu ke konkurenci (Kotler, 2012). Průměrná firma podle Kotlera (2012) ztratí ročně 10 % zákazníků, ale kdyby se podařilo toto číslo snížit na polovinu, mohlo by to zvýšit zisky dané firmy o 25 až 80 % v závislosti na odvětví. Britský operátor Orange například během jednoho roku ušetřil 25 milionů liber díky snížení míry odchodů z 20 na 10 % (Hassouna et al., 2015). Loajální zákazníci také přinášejí více peněz, a to nejen z důvodu vyšších odběrů, ale i díky sníženým nákladům (Kotler, 2012).

Věrnost zákazníků je tedy velmi důležitým kritériem, které může radikálně změnit směr, kterým se podnik ubírá. Podle Kotlera (2012) je „budování silného a blízkého vztahu se zákazníky snem mnoha marketérů a často klíčem k dlouhodobému marketingovému úspěchu“. Budování věrnosti lze provádět mnoha cestami, jako například skrze interakci se zákazníky, a to nejen naslouchání, ale také bojováním za jejich práva, nebo skrze budování věrnostních programů. Kotler (2012) uvádí sedm způsobů, jak vytvořit silné pouto se zákazníky. Mezi tyto způsoby patří například vytvoření lepšího produktu, služeb a zkušeností pro cílový trh, brát ohled na názory a připomínky zákazníků, či například oceňování nejlepších pracovníků podniku, což zdánlivě nemá se zákazníky nic společného,

nicméně to vypovídá o charakteru společnosti a zlepšuje to její postavení z pohledu zákazníků.

Pro ohodnocení, v jaké fázi vztahu ke společnosti se zákazník nachází, se používá marketingový trychtýř, který je vyobrazen na následujícím obrázku.



Obrázek 9 Marketingový trychtýř. Zdroj: Kotler, 2012.

Marketingový trychtýř je nástroj, který zobrazuje hlavní kroky v získávání a udržování zákazníků (Kotler, 2012). Tento trychtýř pomáhá určovat, kolik procent zákazníků z potencionálního cílového trhu se nachází v jaké části rozhodovacího procesu (Kotler, 2012). Při práci s tímto nástrojem se využívá konverzních poměrů, které určují procento zákazníků jedné fáze, které se posune do následující fáze procesu (Kotler, 2012). Analýza těchto poměrů může marketérům pomoci identifikovat klíčové oblasti, kde například dochází k odlivu zákazníků nebo jich velké procento nepostupuje do další části (Kotler, 2012). Pokud je například velké procento lidí nepřecházejících z fáze „otevřen k vyzkoušení“ do fáze „vyzkoušel a nezavrhnul“, lze předpokládat, že je problém například s distribucí nebo s cenou. Z obrázku je patrné, že nejvíce zákazníků je právě v potenciálním cílovém trhu a čím hlouběji v procesu se zákazníci nacházejí, tím jich je méně. Nejmenší procento zákazníků se dostane až do poslední části, kde se ze zákazníků stali ti nejvěrnější. Věrní zákazníci jsou tací, kteří danou značku budou používat pokaždé, dokud je na trhu dostupná. Věrní zákazníci jsou pro společnosti klíčoví. Nejen že stálými odběry přinášejí kontinuální příliv peněz, ale často značce dělají reklamu, čímž lákají nové zákazníky (Kotler, 2012). Základna věrných zákazníků je také důležitá z pohledu hodnoty společnosti, neboť

kdyby měla být prodána, kupující by získával nejen samotnou společnost, ale právě i její zákazníky (Kotler, 2012).

Faktorů, které přímo ovlivňují to, zdali zákazník zůstane u dané společnosti či ne, je velké množství a nelze je přesně určit. Existuje několik důvodů, proč tomu tak je. Každé odvětví je jiné a zákazníci hledají jiné kvality, každý zákazník je jedinečný a nelze tedy přesně určit, že když je jedna věc důležitá pro zákazníka X, bude důležitá i pro zákazníka Y. To, jaká kombinace faktorů vede k odhodu zákazníka, také velmi záleží na tom, jaké data jsou využita. Například Hassouna et al. (2015) ve své studii využili kombinaci mnoha faktorů, jenže nikde není dáno, že tato kombinace je optimální či nejlepší. Faktory využity v jejich studii byly následující:

Category	Variable Name	Description
Demographics	Lifestage_Segment	Subscribers' age stage and gender
	Gender	Subscribers' gender
	Post_Code	Post code in which subscribers live
Cost	Package_Cost	Cost of the package of services chosen by subscribers
	Contract_Length	Number of months of the contract
	Tenure	Number of months with the present mobile operator
Features/Marketing	Tariff	The package of services chosen by subscribers
	Device_Desc	Handset model and manufacturer
	sales_channel	The first channel where the relationship with the customer was established
Usage Level	Q2_bytes	Data usages in the second quarter
	Q3_bytes	Data usages in the third quarter
	Q2_voice	Voice usages in the second quarter
	Q3_voice	Voice usages in the third quarter
Customer Services	No_of_Repairs	Number of times handset has been in for repair in a 12-month period
	Prob_Handset	Known issues with existing handset
	No_of_Complaints	Number of customer complaints regarding billing in a 12 month period

Obrázek 10 Přehled faktorů použitých ve studii. Zdroj: Hassouna et al., 2015.

Z tabulky je vidět, že do modelu vstupovalo velké množství různých faktorů, kde některé měly na výslednou proměnnou větší vliv a jiné vliv minimální. Určit vliv jednotlivých proměnných lze za pomocí jednoho z nástrojů softwaru Rapidminer, který zváží proměnné dle jejich informačního přínosu, a na tomto základě pak lze vyloučit proměnné, které jsou v modelu zbytečné.

3.3.2 **Udržování zákazníka a Data Mining**

Jak již bylo zmíněno dříve, udržování zákazníků je pro firmy zásadní. V posledních několika letech došlo k mnoha zlepšením na poli statistických analýz, které se začínají využívat i v této oblasti. Jak říká Kotler (2012), „marketéři musí znát své zákazníky“. A aby toto bylo možné, je třeba o nich sbírat data a uchovávat je v databázích (Kotler, 2012). Takovéto databázi se říká zákaznická, a obsahuje širokou škálu informací jak o stávajících zákaznících, tak i o potenciálních, a obsahuje aktuální a přístupná data pro tvorbu a údržbu vztahů s novými zákazníky či pro zlepšení prodejů zboží a služeb (Kotler, 2012). Podle Kotlera (2012) ideální zákaznická databáze obsahuje informace o zákazníkových minulých nákupech, demografické (věk, příjem atd.), psychografické (zájmy, názory), informace o preferovaných médiích a další užitečné informace. Data z této databáze pak mohou být využita za pomocí data miningu k odhalení užitečných informací o individuálních zákaznících, trendech i celých segmentech. Podle Hassouna et al. (2015) se donedávna tyto informace do databází získávaly pomocí dotazníků, které jsou však spojeny s vyššími náklady a omezeným přístupem k zákazníkům. Naproti tomu data mining poskytuje znalost celé zákaznické populace založené na analýze současných a minulých dat.

Jedny z nejčastěji používaných metod pro analýzu zákaznických dat jsou logistická regrese a rozhodovací stromy, které právě ve svém článku Hassouna et al. (2015) zkoumají. Obě tyto metody lze využít k předpovědi pravděpodobnosti zákazníkova opuštění dané společnosti. Další často využívanou metodou ke zjišťování tohoto kritéria jsou neuronové sítě. Výhodami a nevýhodami prvních dvou metod se práce zabývá již dříve, co je však stejné pro všechny metody je důležitost vyhodnocování jejich výkonnosti. Hassouna et al. (2015) uvádějí tři metody k vyhodnocení kritéria výkonnosti:

1) Přesnost klasifikace

Tento ukazatel pomáhá určit procentuální podíl pozorování, která byla klasifikována správně, ku celkovému počtu klasifikací. Ukazatel se spočítá následovně:

$$CA = \frac{TP+TN}{TP+FP+TN+FN}$$

Obrázek 11 Přesnost klasifikace. Zdroj: Hassouna et al., 2015.

kde

CA …. Classification accuracy (přesnost klasifikace)

TP …. True positive (správně určená pozitiva)

TN …. True negative (správně určená negativa)

FP …. False positive (nesprávně určená pozitiva)

FN …. False negative (nesprávně určená negativa)

V případě tohoto článku byla cílová proměnná binárního typu, která se snažila předpovědět odchod zákazníků. TP byli tedy zákazníci, kteří byli modelem určeni jako odcházející a ve skutečnosti opravdu odešli. FP byli zákazníci, kteří byli modelem označeni jako odcházející, nicméně ve skutečnosti u společnosti zůstali.

2) Citlivost (Sensitivity) a specifičnost (Speficity)

Citlivost ukazuje podíl skutečných pozitiv, která byla správně identifikována, a specifičnost ukazuje podíl skutečných negativ, která byla správně identifikována. Například telefonní operátoři preferují modely, které mají vysokou citlivost než ty, které mají vysokou specifičnost. Je to dáno tím, že cena spojená s nesprávným zařazením odcházejících zákazníků je vyšší než ta, která je spojena s nesprávným zařazením neodcházejících zákazníků (Hassouna et al., 2015). Ukazatele se vypočítají následovně:

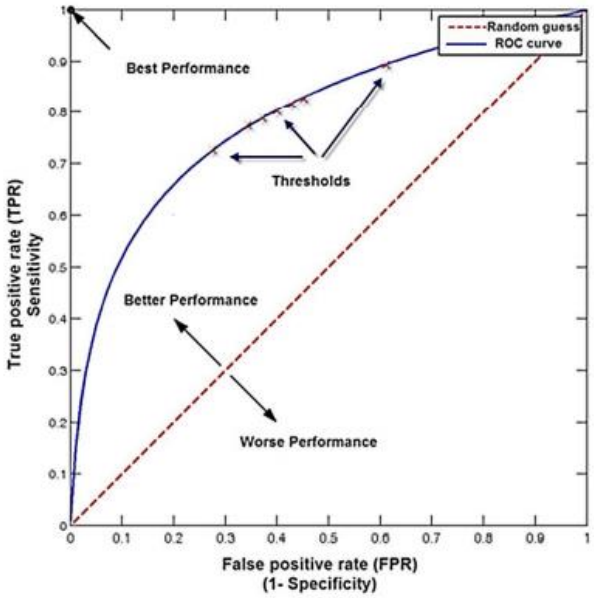
$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

Obrázek 12 Citlivost a specifičnost. Zdroj: Hassouna et al., 2015.

3) ROC křivka

Tato křivka zachycuje vztah mezi poměrem správných pozitiv a poměrem nesprávných pozitiv. V tomto případě křivka zachycuje vztah mezi poměrem odcházejících správně určených jako odcházející a neodcházejících nesprávně určených jako odcházející. ROC křivka může vypadat následovně a nejlepší výkonnost modelu je dosažena, když křivka prochází body (0; 1).



Obrázek 13 Příklad ROC křivky. Zdroj: Hassouna et al, 2015.

Za využití metod data miningu a prediktivního modelování tedy lze z dané databáze určit, které faktory mají největší vliv na odliv zákazníků a následně dle toho jednat. Marketingoví specialisté pak na základě výsledků analýz mohou nejen pracovat na odstraňování nedostatků, ale i vědět, kterým zákazníkům věnovat zvýšenou pozornost nebo naopak neplýtvat penězi za marketing na zákaznících, u kterých je dle výsledků pravděpodobné, že rychle odejdou. Odhalování klíčových faktorů z dostupných dat je ve středu pozornosti CRM a může vést k razantnímu zvýšení zisků. Praktická část této práce bude zaměřena na předpovídání odchodu zákazníků telefonické společnosti, které bude založeno na vzorku zákazníků.

4 Vlastní práce

Pro praktickou část této diplomové práce byla zvolena data, která se týkají vzorku zákazníků jedné nejmenované americké telekomunikační společnosti. Celkem soubor obsahuje záznamy o 7043 zákaznících, o kterých poskytuje jak kategorické, tak numerické informace. Soubor poskytuje pro každého zákazníka 20 proměnných, které jsou následující:

- Pohlaví (Muž/Žena),
- Je zákazník důchodce (Ano/Ne),
- Má zákazník partnera (Ano/Ne),
- Má zákazník na něm závislé osoby (Ano/Ne),
- Jak dlouho je zákazníkem (v měsících),
- Platí zákazník za telefonní služby (Ano/Ne),
- Má zákazník více telefonních linek (Ano/Ne),
- Platí zákazník za internetové služby, pokud ano, jaký typ (DSL/Optický kabel)
- Platí zákazník za online bezpečnost (Ano/Ne),
- Využívá zákazník online zálohování (Ano/Ne),
- Využívá zákazník ochrany elektronických zařízení (Ano/Ne),
- Využívá zákazník internet ke sledování televize (Ano/Ne),
- Využívá zákazník internet ke sledování filmů (Ano/Ne),
- Využívá zákazník technické podpory (Ano/Ne),
- Jaký druh smlouvy zákazník uzavřel (z měsíce na měsíc/roční/dvouletá)
- Využívá zákazník online vyúčtování (Ano/Ne),
- Způsob platby (poštovní šek, elektronický šek, platební příkaz, platební kartou),
- Měsíční platba (v USD),
- Celkové platby (v USD),
- Zákazník v posledním měsíci odešel (Ano/Ne).

Poslední zmíněná proměnná, tedy jestli zákazník změní poskytovatele, je v této analýze tou stěžejní proměnnou, která bude modelována. Cílem této práce je předpovědět, jestli zákazník zůstane u daného poskytovatele služeb, nebo přejde ke konkurenci. Před samotným modelováním je však třeba provést mnoho přípravných kroků a také datový soubor detailně prozkoumat. Pro práci s tímto souborem byl použit software Rapidminer,

což je platforma pro práci s daty, která nabízí mnoho možností, jak s daty pracovat. Rapidminer, což je i název firmy, která software vyvíjí, nabízí jak spousty naučných videí, tak i detailní nápovědu uvnitř programu.

Celý proces bude následovat kroky a postupy metody CRISP Data Mining, která byla zmíněna v předchozí části práce.

4.1 Porozumění byznysové stránce

Data využita pro tuto práci reprezentují vzorek zákazníků telekomunikační společnosti, která nabízí zákazníkům telefonické připojení s možností mít více linek a připojení k internetu. Společnost také nabízí velké množství doplňkových služeb, kterými jsou online bezpečnost, online záloha, ochrana zařízení, technická podpora a online streamování televize a filmů. Zákazníci si sami mohou vybrat kombinaci služeb, které chtějí využívat. Vzhledem k tomu, že se jedná o telekomunikační společnost, které jsou obecně náchylné k častým odchodům zákazníků, je znalost faktorů ovlivňujících toto chování pro danou společnost zásadní. Tyto faktory se budou lišit zákazník od zákazníka, nicméně prediktivní model může společnosti dopomoci s určitou pravděpodobností předpovídat, který typ zákazníků je náchylný k odchodu, a dle této informace se patřičně zachovat. Cílem tohoto data miningového procesu je sestavit tři modely schopné předpovědět pravděpodobnost odchodu zákazníka a vybrat ten model s nejvyšší přesností předpovědi. Sestavenými modely budou logistická regrese, rozhodovací strom a neuronová síť.

První krok této části DM procesu se zaměřuje na definici obchodního cíle modelu, případně modelů, určení toho, co vytvořilo potřebu pro vznik modelu a stanovení kritérií úspěchu. Obchodním cílem tohoto DM procesu tedy je z dostupných dat získat faktory, které největší měrou přispívají k odchodu zákazníků. Na základě výsledků analýz pak budou navrhnuta možná řešení, jak co nejvíce snížit počet odcházejících zákazníků, čímž by se společnosti mohlo podařit navýšit zisky.

Druhým krokem je posouzení okolí celého procesu, kam patří inventura zdrojů, identifikace risků a posouzení okolí, kde bude model působit. Zdroje pro tento DM proces jsou následující:

- Databáze zákazníků
- Rapidminer Studio 8.0 Educational Edition
- DELL PC s Intel Core i3-6100U CPU @ 2.30 GHz, 4 GB RAM

- Cloud computing skrze Rapidminer Studio (k dispozici 8 minut operací v cloudu)

S každým DM procesem souvisí několik rizik. Existuje riziko, že data nebudou dostatečně průkazná, budou obsahovat špatný vzorek zákazníků (například s příliš velkým počtem odcházejících zákazníků), což může vést ke zkresleným výsledkům, které nebude možné použít do budoucna na předpovědi pro nové zákazníky. Dalším rizikem je nedostatečná přesnost modelu. Modely musí mít určitou přesnost, aby je bylo možné použít. Například aby byla logistická regrese považována za dobrý model, její přesnost zařazování musí být alespoň o 25 % lepší než u náhodného modelu (náhodný model je model s přesností 50 %) (Hassouna et al., 2015). Pokud model nemá dostatečnou přesnost, je lepší a levnější použít náhodný výběr a kategorizaci namísto DM modelů.

Nejlepší model, který vyjde z tohoto DM procesu, bude použit v prostředí telekomunikačních společností a mobilních operátorů. Z toho plyne, že bude muset být lehce použitelný lidmi pracujícími v těchto odvětvích, nicméně u těchto lidí se počítá s jistou znalostí IT technologií, takže model nebude muset být úplně zjednodušen. Model bude možné použít nejen společností, pro kterou je primárně sestavován, ale i ostatními společnostmi v oboru. Nicméně podmínkou použití bude existence zákaznických dat, které z celého procesu vyjdou jako klíčová pro určování probability odchodu zákazníka, neboť bez těchto dat nebude model fungovat.

Posledním krokem první části DM procesu je stanovení cílů a úkolů. Primárním cílem je sestavit co možná nejpřesnější model, který splňuje obchodní cíl. Jinými slovy, cílem tohoto DM procesu je správné předpovídání toho, zdali zákazník od společnosti odejde, založené na analýze dostupných dat. Výsledný model by měl společnosti generovat dostatečnou přidanou hodnotu, aby se vůbec vyplatilo celý proces absolvovat. Do úkolů se řadí získání vhodných dat, sestavení modelů a jejich následná evaluace a vyhodnocení toho nejlepšího nebo spuštění modelu. Úkoly pro tento proces jsou následující:

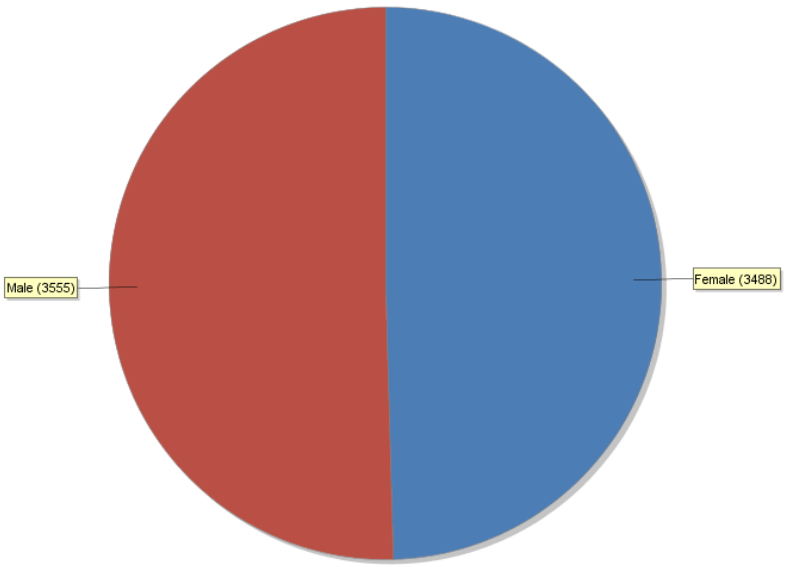
- Získání dat
- Příprava dat
- Sestavení logistické regrese
- Verifikace a vyhodnocení modelu
- Sestavení rozhodovacího stromu a náhodného lesa
- Verifikace a vyhodnocení modelu
- Sestavení neuronové sítě

- Verifikace a vyhodnocení modelu
- Výběr nejlepšího modelu a jeho následná aplikace
- Analýza výsledků a návrh opatření

4.2 Porozumění dat

Porozumění dat je druhou částí DM procesu, která je dále rozdělena na čtyři pod kroky. První částí je akvizice dat. Data pro tuto práci byla získána od společnosti IBM. Společnost IBM se zabývá big data analýzami již řádku let a vyvinula pro tyto účely počítač Watson, který odpovídá na položené otázky pomocí analýzy vložených dat. Soubor využit v této práci slouží jako ukázkový soubor pro zájemce o práci s Watsonem. Datový soubor obsahuje informace o vzorku 7043 zákazníků a 20 proměnných. Druhou částí porozumění dat je integrace dat. Zde je důležité prozkoumat formát a jednotky proměnných. U nominálních proměnných není nutné nic zkoumat, proměnné jako pohlaví či seniorita zákazníka jsou jasně vyjádřeny. Nicméně i zde může nastat rozdíl ve vyjádření. Například seniorita je vyjádřena pomocí 0 a 1, zatímco například to, zdali má zákazník partnera, je vyjádřeno pomocí Ano a Ne. Soubor dále obsahuje tři kvantitativní proměnné, a to počet měsíců strávených u stávajícího poskytovatele, což je vyjádřeno v měsících, a měsíční a celkové poplatky, které jsou vyjádřeny v amerických dolarech. Třetí částí porozumění dat je seznámení se s daty. To lze provést například pomocí základních statistických metod. Rapidminer nabízí velmi rychlý a přehledný náhled na základní informace o datovém souboru, kdy stačí pouze spojit vstupní datový soubor s uzlem výsledků a program nabídne základní statistiky pro všechny proměnné.

4.2.1 Základní informace o souboru



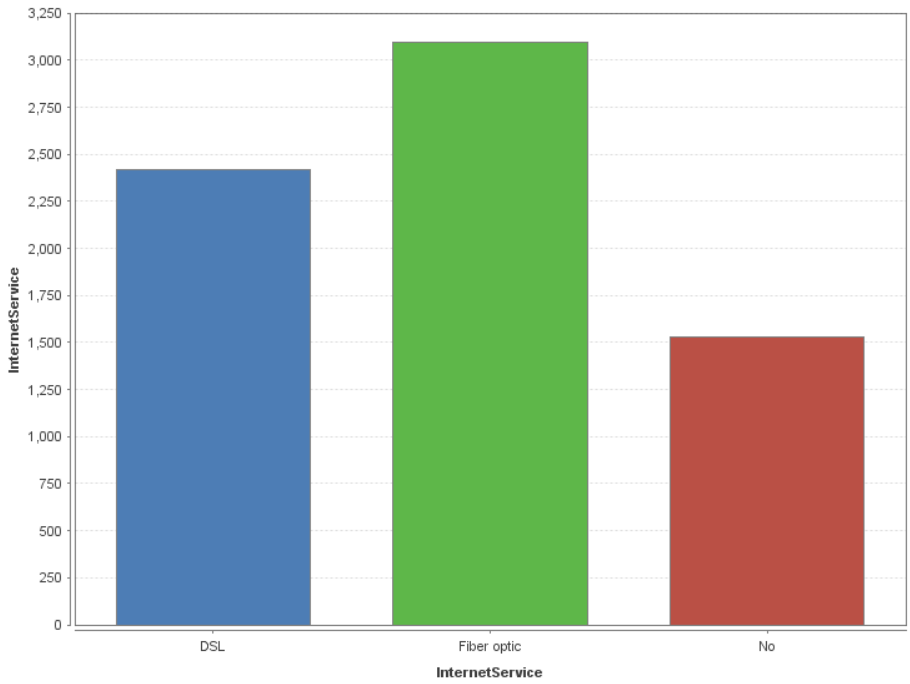
Obrázek 14 Rozdělení podle pohlaví. Zdroj: IBM, vlastní zpracování.

Z obrázku 14 je patrné, že rozdělení pohlaví je v daném souboru relativně stejné, muži (červeně) zaujímají 50,4 % a ženy (modře) 49,6 % z celkového počtu zákazníků. Rapidminer poskytuje základní informace o každé z proměnných. Informace o proměnných nabývajících hodnoty 0-1 (Ano – Ne) jsou obsaženy v následující tabulce.

PROMĚNNÁ	POČET ZÁKAZNÍKŮ ODPOVÍDAJÍCÍCH:		
	ANO	NE	NEMAJÍ INTERNET
DŮCHODCE	1134 (16,1 %)	5909 (83,9 %)	
PARTNER	3402 (48,3 %)	3641 (51,7 %)	
ZÁVISLÍ	2110 (29,95 %)	4933 (70,05 %)	
TELEFON	6361 (90,31 %)	682 (9,69 %)	
ONLINE BEZPEČNOST	2019 (28,67 %)	3498 (49,67 %)	1526 (21,66 %)
ONLINE ZÁLOHA	2429 (34,49 %)	3088 (43,85 %)	1526 (21,66 %)
OCHRANA ZAŘÍZENÍ	2422 (34,39 %)	3095 (43,95 %)	1526 (21,66 %)
TECHNICKÁ PODPORA	2044 (29,02 %)	3473 (49,32 %)	1526 (21,66 %)
ONLINE TV	2707 (38,44 %)	2810 (39,90 %)	1526 (21,66 %)
ONLINE FILMY	2732 (38,79 %)	2785 (39,55 %)	1526 (21,66 %)
ONLINE VYÚČTOVÁNÍ	4171 (59,22 %)	2872 (40,78 %)	

Tabulka 1 Základní informace o souboru. Zdroj: IBM, vlastní zpracování.

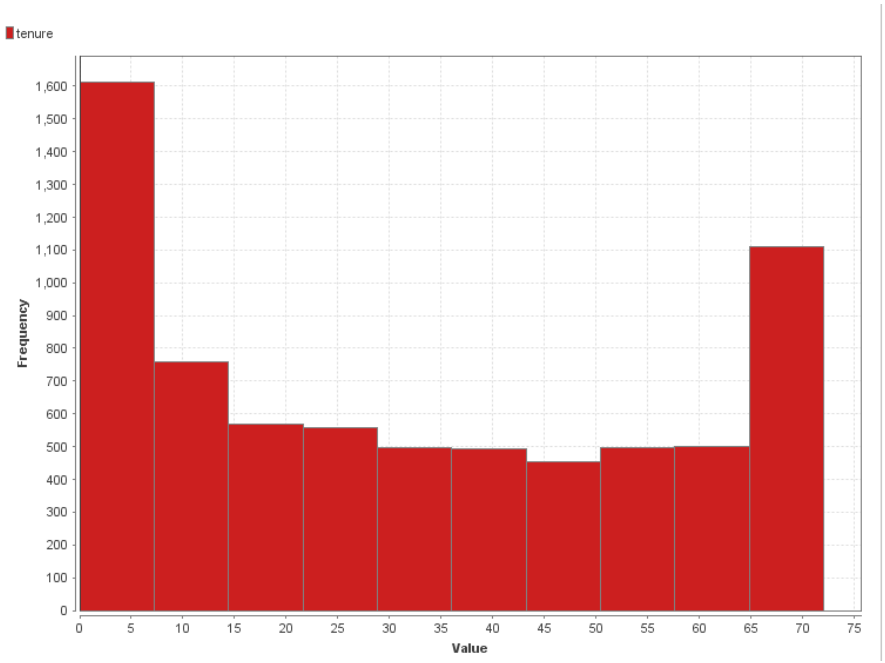
4.2.2 Způsob připojení k internetu



Obrázek 15 Histogram délky kontraktu. Zdroj: IBM, vlastní zpracování.

Na obrázku 15 je vidět, že zhruba 1500 lidí internetové služby od tohoto poskytovatele neodebírá vůbec, a z těch, kteří ano, tak většina má připojení přes optický kabel.

4.2.3 Délka kontraktu



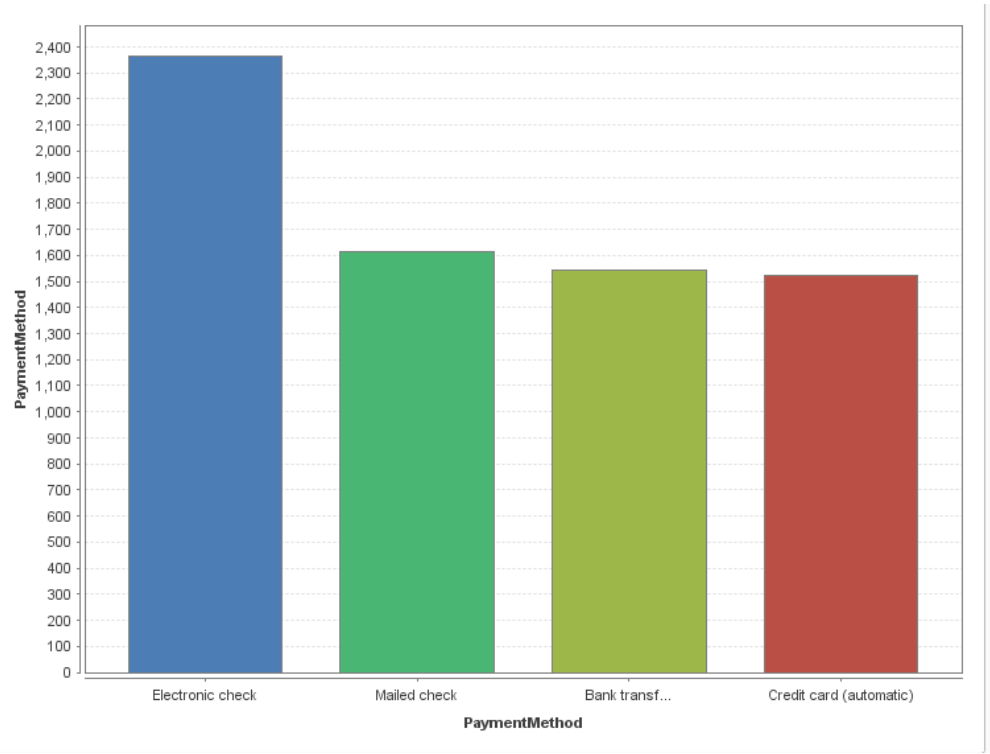
Obrázek 16 Histogram délky kontraktu. Zdroj: IBM, vlastní zpracování.

Obrázek číslo 16 zobrazuje histogram, na kterém je vidět, jak dlouhé smlouvy mají zákazníci této firmy, kdy na ose x jsou zobrazeny měsíce a na ose y počty zákazníků mající smlouvy vyjádřené v délce měsíců. Z grafu je patrné, že největší procento zákazníků je nových, tudíž mají kontrakty dlouhé maximálně do 7,5 měsíců. Druhou nejpočetnější skupinou jsou zákazníci s nejdelšími smlouvami, a to od 65 do 72,5 měsíců. To, jak dlouho je zákazník u jedné společnosti, je z pohledu teorie velmi důležitá věc, neboť na nové zákazníky se musí působit jinak než na ty, kteří už jsou u firmy alespoň rok. Délka kontraktu bude mít také důležitou roli při modelování toho, zdali zákazník zůstane, či odejde.

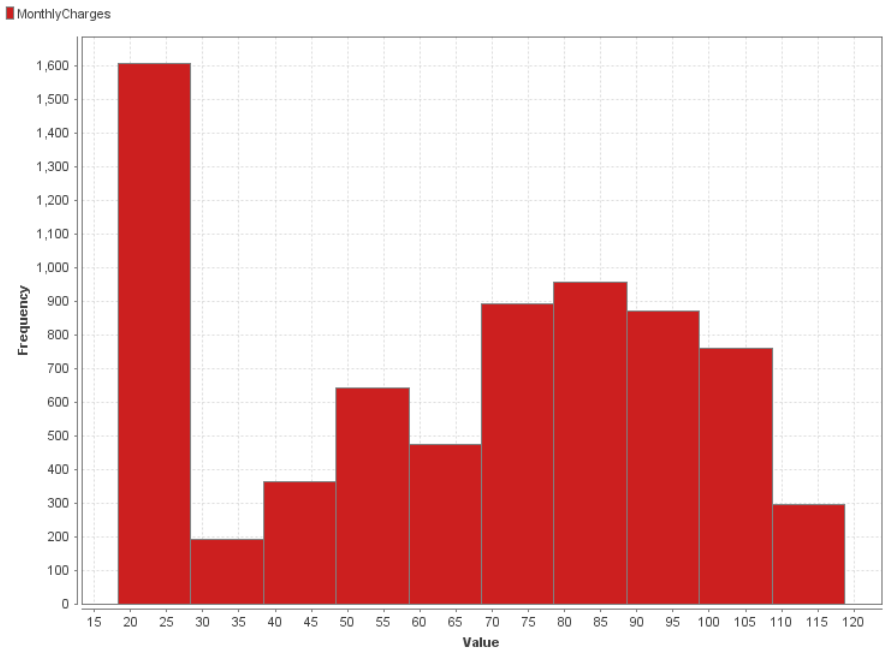
4.2.4 Způsob platby

Z obrázku 17 je patrné, že největší procento lidí upřednostňuje platbu pomocí elektronického šeku a ostatní metody mají zhruba stejnou oblíbenost. Z grafu lze také vyčíst, že více jak polovina zákazníků upřednostňuje platbu šekem, zatímco platba převodem či kreditní kartou není tak populární. Kreditní karty jsou v USA velmi populární, nicméně se

dá předpokládat, že v tomto případě je myšlena platební karta obecně, a ne pouze kreditní karta. Datový soubor se vztahuje na americké zákazníky a lze očekávat, že v Čechách by tento graf vypadal jinak, už jen kvůli nepopulárnosti šeků (elektronických i papírových).



Obrázek 17 Metoda placení. Zdroj: IBM, vlastní zpracování.

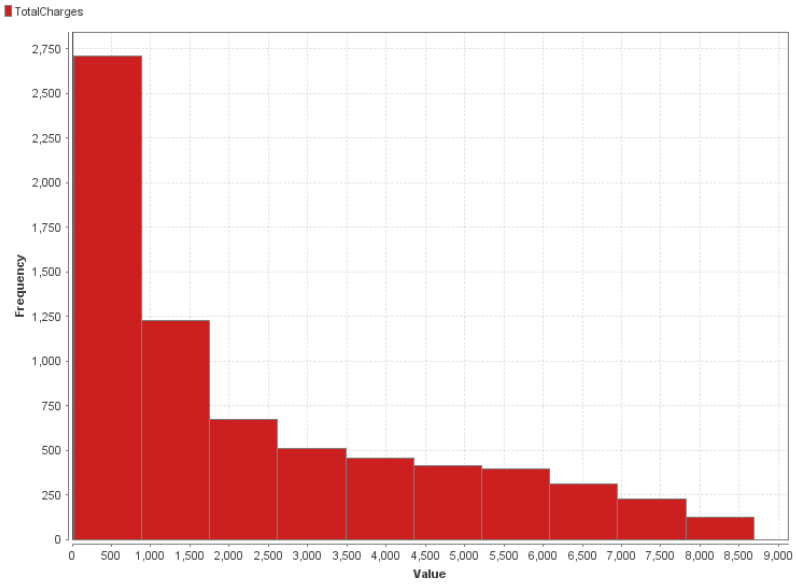


Obrázek 18 Histogram výše měsíčních poplatků. Zdroj: IBM, vlastní zpracování.

4.2.5 Měsíční a celkové platby

Obrázek 18 vystihuje frekvenci měsíčních poplatků za poskytované služby. Poplatky v rozmezí 17,5 až 27,5 USD jsou nejčastější, kdy přes 1600 zákazníků platí poplatky právě v tomto rozmezí. Právě kvůli této skupině není možné mluvit o normálním rozdělení. Průměrná výše měsíčního poplatku je 64,762 USD, nejmenší hodnotou je 18,25 USD a nejvyšším měsíčním poplatkem jej 118,75 USD. Směrodatná odchylka této proměnné je 30,09 USD. Všechny hodnoty byly spočítány softwarem Rapidminer.

Na obrázku 19 je pak histogram celkových plateb. Nejvíce lidí celkově poskytovateli služeb za celou dobu trvání kontraktu zaplatilo v rozmezí 0 až 850 USD. Nejmenší hodnotou u této proměnné je 18,80 USD, nejvyšší 8684,80 USD a průměrnou hodnotou 2283,30 USD. Směrodatná odchylka tohoto souboru dat je 2266,771 USD, což vypovídá o velkém rozdílu v hodnotách jednotlivých záznamů.

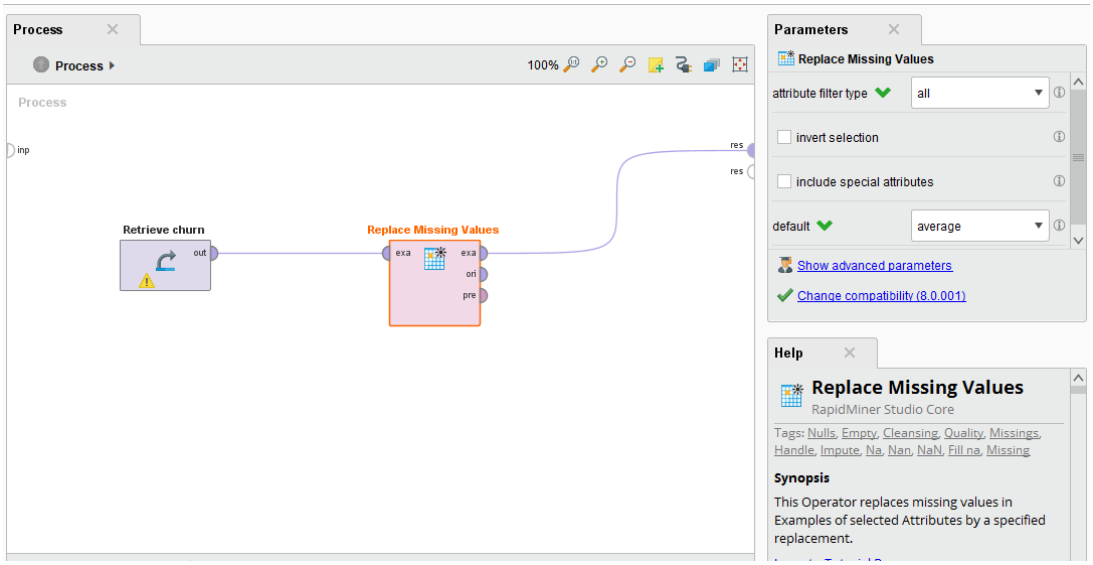


Obrázek 19 Histogram celkových plateb. Zdroj: IBM, vlastní zpracování.

Software Rapidminer také zaznamenal 11 chybějících hodnot, které bude nutné odstranit pro další analýzy a modelování, což vede k poslednímu kroku části o porozumění dat. V této části se řeší, jak naložit právě s chybějícími hodnotami a s extrémními hodnotami. Samotný Rapidminer nabízí několik možností, jak se s chybějícími hodnotami vypořádat. V první řadě je však nutné vytvořit proces na úpravu dat. Rapidminer nástroje na čištění a úpravu dat řadí do kategorie Cleansing a v pod kategorii Missing se nachází 6 nástrojů na práci s chybějícími daty. Těmito nástroji jsou:

- Nahrazení chybějících hodnot, kde si lze určit, čím budou chybějící data nahrazena. Přednastavené je nahrazení průměrem všech dostupných záznamů daného atributu, nicméně chybějící hodnoty lze nahradit i minimální či maximální hodnotou dostupných záznamů, nulou, nebo uživatelem stanovenou hodnotou.
- Dopočtení chybějících hodnot, kde program odhadne chybějící hodnoty daného atributu aplikováním modelu, který program sestaví z hodnot všech dalších proměnných kromě proměnné označené „label“ (cílová proměnná). Rapidminer se tedy porozuměním ostatních dat snaží doplnit chybějící data.
- Prohlášení chybějících hodnot je nástroj, který je schopen označit uživatelem vybrané hodnoty dané proměnné jako chybějící hodnoty.
- Nahrazení nekonečných hodnot je nástrojem, který nahradí nekonečné hodnoty proměnných jednou z nabízených možností: Pokud se jedná a kladnou hodnotu nekonečna, lze ji nahradit horní hranicí variačního rozpětí neboli největší hodnotou. Pokud se jedná o záporné nekonečno, lze jej nahradit spodní hranicí variačního rozpětí. Dále lze hodnoty nekonečna nahradit nulou, ničím, nebo funkcí „missing“, která nahradí nekonečno nečíslem.
- Nástroj odstranění nepoužívaných hodnot odstraní každou nominální hodnotu, která není spojena s žádným příkladem.
- Posledním nástrojem pro práci s chybějícími daty je vyplnění datových mezer. Tento nástroj doplní chybějící pole na základě atributu ID. Tato nová pole budou obsahovat nulové hodnoty.

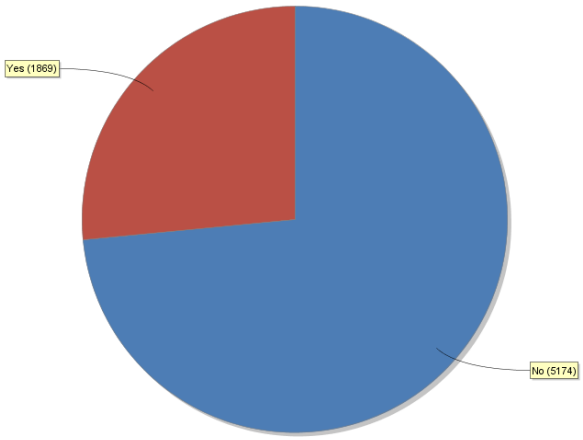
Jak bylo řečeno dříve, proměnná celkové platby obsahovala 11 chybějících hodnot. Pro další práci se souborem je tedy nutné se těchto chybějících hodnot zbavit. V tomto případě bylo využito prvního nástroje, což je nahrazení chybějících hodnot. Chybějící hodnoty v tomto souboru byly nahrazeny průměrem všech ostatních hodnot.



Obrázek 20 Nahrazení chybějících hodnot. Zdroj: vlastní zpracování.

4.2.6 Odchod zákazníků v posledním měsíci

Z celkového počtu 7043 zákazníků zahrnutých v tomto vzorku jich za minulý měsíc odešlo 1869, což představuje 26,54 % a zůstalo jich 5174, což je 73,46 %. Procento odchodů je vysoké a pohybuje se v rozmezí 20-40 % stanoveném Hassounou et al. (2015). Výsledný model by měl odhalit, na kterých oblastech je třeba zapracovat, aby se toto procento do budoucna snížilo.



Obrázek 21 Poměr odešlých (červeně) a zůstalých (modře) zákazníků za poslední měsíc. Zdroj: IBM, vlastní zpracování.

4.3 Příprava dat

Příprava dat je velmi důležitým krokem celého DM procesu. V tomto kroku je třeba dostat data do požadovaného formátu. Vzhledem k tomu, že budou vytvořeny celkem 3 modely a každý má trochu odlišné požadavky, bude transformace dat probíhat až před samotným modelováním u každého modelu zvlášť. Rapidminer při sestavování procesů nutných k vytvoření modelu sám říká, co je s daty špatně a jak tuto chybu odstranit. Většinou lze chybu odstranit jedním z mnoha nástrojů stvořených právě pro přípravu dat. V tomto kroku také nastává filtrování a redukce dat. Ne všechny proměnné mají na výslednou proměnnou vliv a jejich přítomnost by tak pouze model zatěžovala. Dalším důvodem pro odstranění některých dat je přítomnost multikolinearity. V takovém případě je možné odstranit jednu z proměnných způsobujících multikolinearitu. Jednotlivé kroky nutné pro přípravu dat pro každý model budou detailně probrány u příslušných modelů.

4.4 Modelování

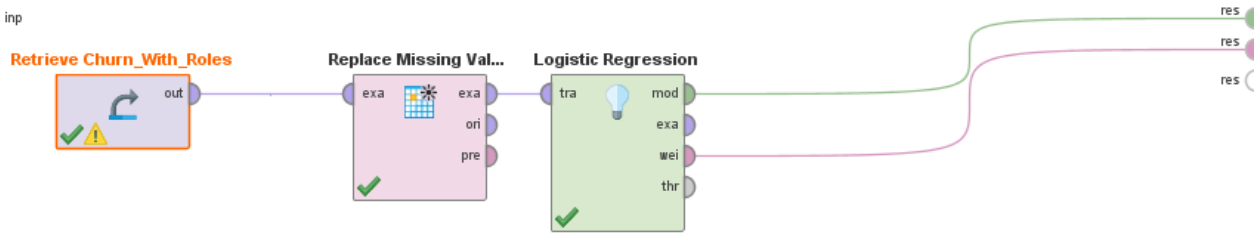
4.4.1 Logistická regrese

Prvním zvoleným modelem pro analýzu zákaznických dat je logistická regrese. Tento model je využívaný relativně často, avšak má své nedostatky. Mezi ně patří fakt, že logistická regrese může vést k zavádějícím výsledkům při vyhodnocování kauzality a váhy jednotlivých proměnných. Také sem patří multikolinearita, která může být výsledkem silných korelací mezi nezávislými proměnnými. Z těchto důvodů je nutné tuto regresi důkladně testovat, aby byly zajištěny co nejlepší výsledky. Cílem tohoto modelu je sestavit rovnici, která bude poskytovat co nejpřesnější pravděpodobnost zařazení zákazníka podle toho, jestli u společnosti zůstane, či odejde, a také určit faktory, které na toto chování mají největší vliv.

Prvním krokem v procesu sestavování logistické regrese musí být příprava dat. Podle obrázku číslo 20 byly chybějící hodnoty nahrazeny průměrem přítomných hodnot. Dále je třeba určit role jednotlivých proměnných a zajistit jejich správný formát. Tento krok lze provést již při importování datového souboru do modelovacího prostředí. Pro tento soubor byly provedeny následující úpravy:

- Proměnná Customer ID byla označena jako ID neboli identifikační proměnná, a to z důvodu, že je to proměnná sloužící pouze k rozlišování jednotlivých zákazníků a na jejich chování nemá žádný vliv.
- Proměnná Churn neboli odchod v posledním měsíci byla označena jako cílová proměnná, jelikož právě tato proměnná je cílem modelu a bez této znalosti by software nerozlišil, kterou proměnnou má modelovat.
- V poslední řadě byl změněn typ proměnné Důchodce, která nabývá hodnot 0 nebo 1, nicméně v původním souboru byl typ této proměnné určen jako celé číslo, což značně změnilo práci s touto proměnnou. Proměnná Důchodce byla nastavena jako polynominální.

Nyní jsou data připravena a je možné přistoupit k samotnému sestavení modelu, které je znázorněno na obrázku 22.



Obrázek 22 Schéma modelu logistické regrese. Zdroj: Vlastní zpracování.

Pro sestavení logistické regrese byl využit nástroj Logistic Regression, díky kterému lze určit koeficienty jednotlivých proměnných. Při interpretaci výsledků logistické regrese je cílem být na základě modelu schopen s určitou pravděpodobností predikovat zákazníkovo chování, a to že buď zůstane, nebo odejde. Výsledek logistické regrese je uveden v následující tabulce (proměnné jsou seřazeny podle p-hodnoty). V tabulce jsou uvedeny pouze statisticky významné parametry, tabulka obsahující veškeré proměnné je součástí přílohy (příloha č. 1).

Proměnná	Označení	Koeficient	Směrodatná chyba	p-hodnota
Délka kontraktu	X_1	-0,0585	0,0061	0,0000
Smlouva na 2 roky	X_2	-1,3962	0,1758	0,0000
Smlouva na 1 rok	X_3	-0,6685	0,1074	0,0000
Online vyúčtování-Ne	X_4	-0,3416	0,0745	0,0000
Celkové platby	X_5	0,0003	0,0001	0,0000
Platební metoda-karta (automaticky)	X_6	-0,3924	0,0973	0,0001
Platební metoda-poštovní šek	X_7	-0,3615	0,0963	0,0002
Platební metoda-trvalý příkaz	X_8	-0,3062	0,0945	0,0012
Důchodce-Ano	X_9	0,2147	0,0845	0,0111
Internet-optický kabel	X_{10}	1,7530	0,7976	0,0280

Tabulka 2 Výsledek logistické regrese. Ponechány pouze statisticky významné proměnné. Zdroj: IBM, vlastní zpracování.

Z logistické regrese byly z důvodu přítomnosti multikolinearity odstraněny následující proměnné:

- Online bezpečnost-Nemá internet,
- Online záloha-Nemá internet,
- Ochrana zařízení-Nemá internet,
- Technická podpora-Nemá internet,
- Online TV-Nemá internet,

- Online filmy-Nemá internet a
- Telefon-Ano.

V úplné tabulce výsledků, která je v příloze, mají tyto proměnné koeficient 0 a jejich směrodatné odchylky a p-hodnoty nejsou číselnými záznamy (pozn. NaN-not a number). V tabulce 2 jsou tak uvedeny pouze ty proměnné, jejichž p-hodnota je menší než hladina významnosti α=5 %, a jsou tedy statisticky významné. Ostatní proměnné byly na základě vyšších p-hodnot vyloučeny. Mezi vyloučené patří i konstanta, jejíž p-hodnota je 0,075. Na základě těchto výsledků by rovnice logistické regrese předpovídající pravděpodobnost odchodu zákazníka vypadala následovně:

$$\text{logit}(y) = -0{,}0585x_1 - 1{,}3962x_2 - 0{,}6685x_3 - 0{,}3416x_4 + 0{,}0003x_5 - 0{,}3924x_6 - 0{,}3615x_7 - 0{,}3062x_8 + 0{,}2147x_9 + 1{,}7530x_{10}$$

Rovnice 6 Logistická regrese. Zdroj: Vlastní zpracování.

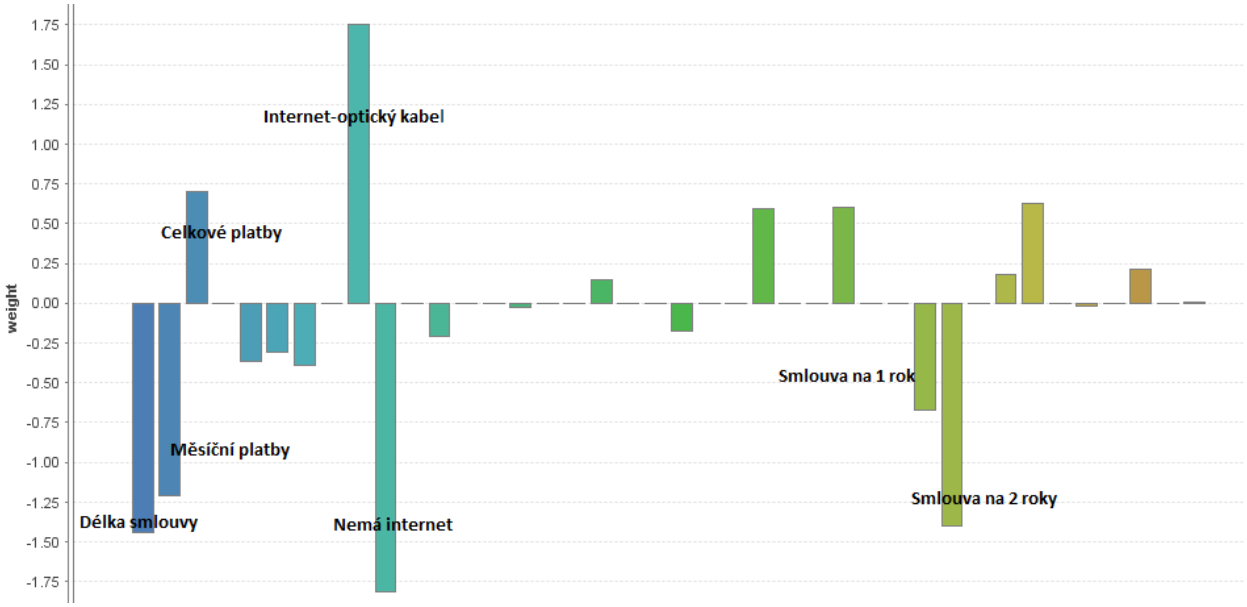
Na základě této rovnice lze nejen identifikovat faktory, které mají na výslednou proměnnou *y* velký vliv a které naopak vliv zanedbatelný, ale je možné i na základě informací získaných od nových zákazníků určit, jestli zákazník u společnosti zůstane, či odejde. V následující části je popsáno to, jak daná proměnná působí na cílovou proměnnou *y* neboli odchod zákazníka.

- x_1 – čím delší je délka kontraktu v měsících, tím se snižuje pravděpodobnost odchodu zákazníka. Logit se sníží o 0,0585 s každým přibývajícím měsícem. Násobek, o který se změní šance zákazníkova odchodu je exp (-0,0585) =0,9432. Toto číslo znamená, že se šance odchodu zmenší. Velikost i směr působení koeficientu této proměnné jsou v souladu s očekávaným chováním zákazníka, který čím déle je u jednoho poskytovatele služeb, tím je jeho odchod méně pravděpodobný.
- x_2 – pokud má zákazník smlouvu sjednanou na 2 roky, logit se snižuje, a to o 1.3962. Při přepočtu na šanci pomocí vztahu exp (-1,3962) je získáno číslo 0,2475, které značí násobek, o který se šance odchodu zákazníka zmenší. Opět je velikost i směr působení koeficientu proměnné v souladu s očekáváním i v souladu s předchozí proměnnou.
- x_3 – velmi podobné proměnné x_2, pouze koeficient je menší. Pokud má zákazník smlouvu na 1 rok, logit se snižuje o 0,6685 a šance se zmenšuje exp (-0,6685) = 0,5125krát.

- x_4 – pokud zákazník nevyužívá online vyúčtování, logit se snižuje o 0,03416. Šance odchodu se pak snižuje exp (-0,03416) = 0,9664krát. Toto může souviset s několika faktory, například s věkem, kdy starší zákazníci nevyužívají online vyúčtování tak často, jelikož nejsou dostatečně technologicky znalí, což může ve výsledku vést i k tomu, že je pro ně změna poskytovatele služeb náročnější, a tak preferují zůstat u toho stávajícího.
- x_5 – velikost koeficientu proměnné celkové platby působí na cílovou proměnnou v kladném směru, a to tak, že za každý dolar zaplacený zákazníkem se logit zvyšuje o 0,0003. Velikost násobku, o který se zvyšuje šance odchodu je rovna exp (0,0003) = 1,0003. Tento koeficient je v rozporu s očekáváním i s předchozími koeficienty, jelikož lze očekávat, že čím déle je zákazník u jednoho poskytovatele, tím více peněz mu zaplatí, a tím pádem by se měla pravděpodobnost jeho odhodu snižovat.
- x_6 – pokud zákazník platí za služby automaticky pomocí platební karty, logit se snižuje o 0,3924. Velikost násobku, o který se zvyšuje šance odchodu je rovna exp (-0,3924) = 0,6754.
- x_7 – pokud zákazník platí pomocí šeku zaslaného poštou, logit se snižuje o 0,3615. Velikost násobku, o který se zvyšuje šance odchodu je rovna exp (-0,3615) = 0,6966.
- x_8 – pokud zákazník platí pomocí trvalého příkazu, logit se snižuje o 0,3062. Velikost násobku, o který se zvyšuje šance odchodu, je rovna exp (-0,3062) = 0,7362.
- x_9 – pokud je zákazník důchodcem, logit se zvyšuje o 0,2147. Velikost násobku, o který se zvyšuje šance odchodu je rovna exp (0,2147) = 1,2395. Tento koeficient je tedy v rozporu s předchozí teorií i s očekávaným chováním.
- x_{10} – jestliže má zákazník internet vedený přes optický kabel, vysoce to ovlivňuje celý logit. Velikost koeficientu je 1,753, což z něj dělá největší koeficient. Velikost násobku, o který se zvyšuje šance odchodu je rovna exp (1,753) = 5,7719. Lze očekávat, že takovýto zákazník se snaží využívat ty nejmodernější technologie a ty nejlepší nabídky na trhu, a z toho důvodu často mění poskytovatele služeb.

Další funkcí, kterou nabízí software při modelování, je určení váhy jednotlivých proměnných. Na obrázku 22 je vidět, že z operátoru logistické regrese vedou dva vývody – jedním je výsledek logistické regrese a tím druhým je právě váha jednotlivých atributů. Na obrázku 23 je pak vidět, které proměnné mají na cílovou největší vliv, ať už negativní, či

pozitivní. Dle obrázku má na setrvání zákazníka také velký vliv proměnná Měsíční platby a Nemá internet, nicméně tyto proměnné byly regresí určeny jako statisticky nevýznamné.

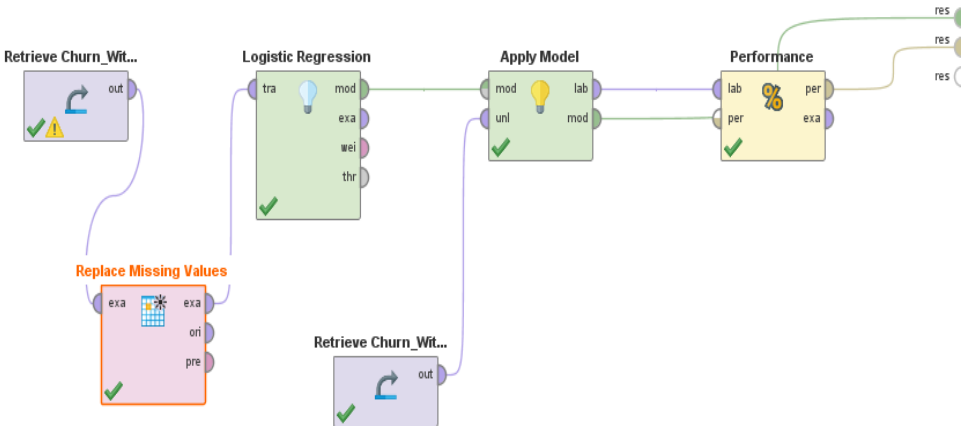


Obrázek 23 Váha jednotlivých proměnných na cílovou v logistické regresi. Zdroj: IBM, vlastní zpracování.

Z grafu plyne, že dle logistické regrese mají na odchod či setrvání zákazníka největší vliv tyto statisticky významné proměnné: Internet-optický kabel, Délka smlouvy, Smlouva na 2 roky, Celkové platby a Smlouva na 1 rok.

Dalším krokem v procesu modelování za využití logistické regrese je změření výkonnosti modelu. Výkonnost modelu lze změřit za využití operátoru Performance neboli Výkonost. Výkonnost modelu lze zjistit pouhou aplikací modelu. Aplikace modelu znamená, že je výsledek logistické regrese aplikován na data, a porovnává se, zdali je

předpověď modelu shodná se skutečností. Výsledkem je takzvaná confusion matice, která udává počet správných a špatných klasifikací.



Obrázek 24 Schéma procesu měření výkonnosti. Zdroj: Vlastní zpracování.

Na tomto schématu je vidět, jak celý proces probíhá. Nejdříve je nutné provést samotnou logistickou regresi a poté je nutné aplikovat model na původní data. Následně je změřena výkonnost modelu, jejímž výsledkem je následující tabulka:

accuracy: 80.70%

	true No	true Yes	class precision
pred. No	4650	835	84.78%
pred. Yes	524	1034	66.37%
class recall	89.87%	55.32%	

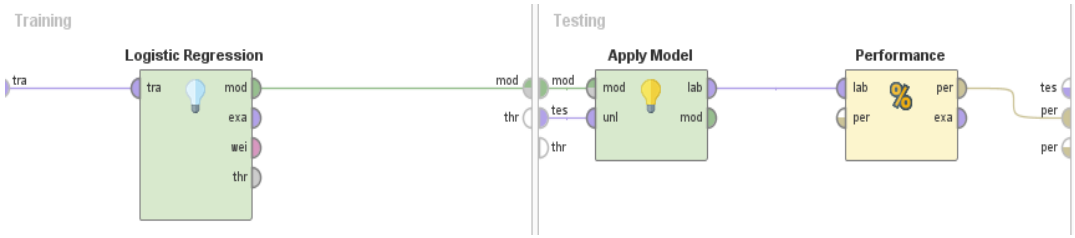
Tabulka 3 Výkonnost modelu. Zdroj: IBM, vlastní zpracování.

Z tabulky vyplývá následující:

- celková přesnost předpovědí je 80,70 %
- jako zůstávající zákazníci bylo správně označeno celkem 4650 záznamů a jako zůstávající zákazníci bylo chybně označeno 835 záznamů, což vede k celkové přesnosti předpovědi setrvání u společnosti 84,78 %
- jako odcházející zákazníci bylo správně označeno 1034 záznamů a jako odcházející zákazníci bylo chybně označeno 524 záznamů, což vede k celkové přesnosti předpovědi odchodu zákazníka pouze 66,37 %.
- Celkem se podařilo zachytit 89,87 % zákazníků, kteří ve skutečnosti u společnosti zůstali, ale pouze 55,32 % zákazníků, kteří ve skutečnosti odešli.

Existují však metody, jak přesnost modelu určit přesněji. V předchozím kroku byla výkonnost měřena na stejném souboru dat, na kterém byl model vytvořen, což nepředstavuje

ideální situaci. Jak ve svém článku zmiňují Logan et al. (2016), nejvhodnějším nástrojem na validaci modelu je validace napříč různými soubory dat (cross validation). Software Rapidminer na provedení tohoto kroku nabízí velmi vhodný nástroj nazvaný Cross-Validation. Tento nástroj rozdělí soubor dat na předem určený počet podsouborů a následně provádí logistickou regresi na n-1 souborech a podsoubor, který nebyl použit pro modelování, je následně použit pro validaci modelu. Celková přesnost modelu je pak určena jako průměr přesností jednotlivých pod-modelů. Standardní počet validací je 10, což bylo využito i u této validace. Proces validace napříč soubory obsahuje jeden sub-proces, ve kterém je nutné určit, jaký model bude validován, jaká data použít a jaké soubory budou sloužit k validaci. Tento sub-proces vypadá následovně:



Obrázek 25 Cross-validation. Zdroj: Vlastní zpracování.

Výsledkem této validace je také confusion matice, která je však lehce odlišná od té, která byla vytvořena za využití celého souboru dat.

accuracy: 80.39% +/- 1.12% (mikro: 80.39%)

	true No	true Yes	class precision
pred. No	4631	838	84.68%
pred. Yes	543	1031	65.50%
class recall	89.51%	55.16%	

Tabulka 4 Výsledek validace napříč 10 podsoubory dat. Zdroj: IBM, vlastní zpracování.

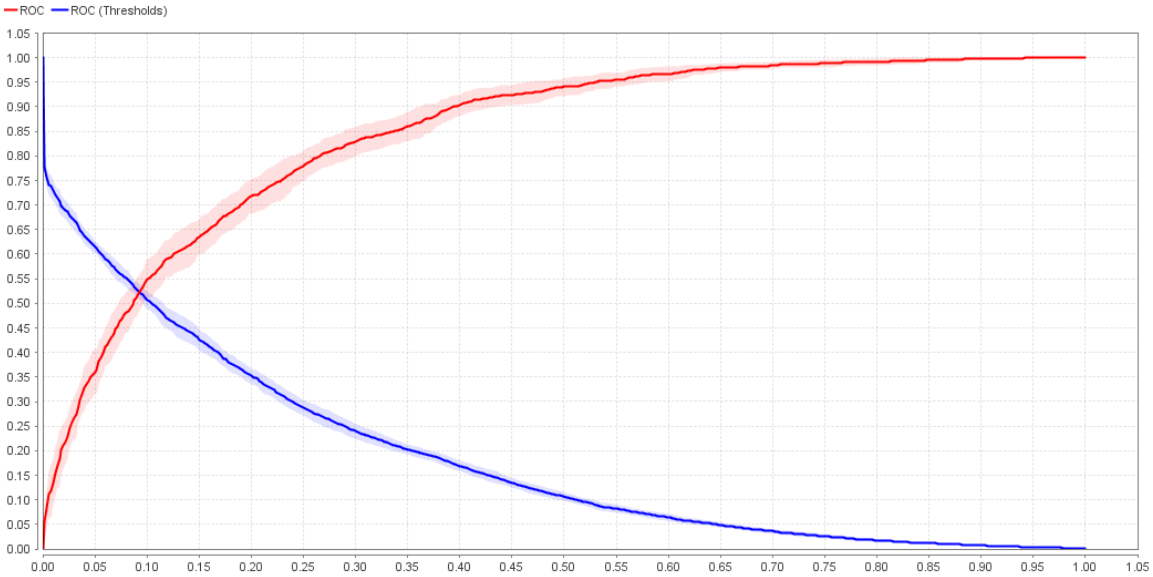
Celková přesnost modelu neboli přesnost klasifikace zmíněná v teoretické části práce, se nepatrně snížila, a to o 0,31 % na 80,39 % a směrodatná odchylka měření je pouze 1,12 %, což značí vysokou stabilitu modelu. Pokud by tato odchylka byla vysoká, jednotlivá měření by se velmi lišila a model by nebyl tak průkazný. Dále je z tabulky patrná citlivost a specifičnost.

- Citlivost = $\frac{1031}{1031+838} = 0{,}5516$

- Specifičnost = $\frac{4631}{4631+543} = 0{,}8951$

Z citlivosti vyplývá, že celkem 55,16 % skutečných pozitiv bylo správně zařazeno a ze specifičnosti vyplývá, že celkem 89,51 % skutečných negativ bylo správně zařazeno. Vyšší citlivost je v modelu žádanější než vyšší specifičnost, což je popsáno Hassounou et al. (2015) v jejich článku, kde zmiňují, že náklady asociované se špatnou klasifikací odcházejícího zákazníka jsou vyšší než ty, které jsou asociované s nesprávným zařazením zůstávajícího zákazníka. V tomto modelu je však specifičnost vyšší, a to o 34,35 %.

Dalším nástrojem k vyobrazení výkonnosti modelu a přesnosti jeho predikcí je křivka ROC. Tato křivka popisuje vztah mezi mírou pravých pozitiv na ose y a mírou nepravých pozitiv na ose x. Pro logistickou regresi vypadá následovně:



Obrázek 26 ROC křivka logistické regrese. Zdroj: IBM, vlastní zpracování.

Pro ROC křivky obecně platí, že čím blíže bodu (0,1) křivka probíhá, tím je model lepší. Křivka tohoto modelu se od tohoto bodu vzdaluje relativně rychle, což naznačuje ne úplně vysokou předpovídací schopnost. Tento fakt je v souladu s tabulkou číslo 4, ve které je zmíněna celková přesnost předpovědí modelu 80,39 %.

4.4.1.1 Shrnutí logistické regrese

Pro zajištění správného chodu modelu bylo třeba nejprve upravit data, a to za pomocí doplnění chybějících hodnot a určení cílové proměnné. Následně byla provedena samotná logistická regrese, která vygenerovala koeficienty jednotlivých proměnných a také jejich p-

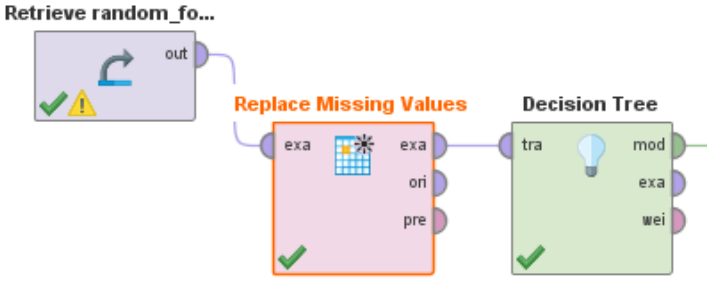
hodnoty, které pomohly k určení významnosti v modelu. Jako významné se ukázaly tyto proměnné: délka kontraktu, smlouva na 2 roky, smlouva na 1 rok, online vyúčtování-ano, celkové platby, platební metoda-kartou (automaticky), platební metoda-poštovní šek, platební metoda-trvalý příkaz, důchodce-ano a internet-přes optický kabel. Jako proměnné nejvíce ovlivňující cílovou proměnnou se ukázaly internet-přes optický kabel, délka smlouvy a smlouva na 2 roky. Dále byly z modelu odstraněny všechny kolineární proměnné, kterými byly všechny (až na jednu) proměnné s možností odpovědi „nemá internet".

Poté byla změřena výkonnost modelu, která vyšla 80,71 %, a následně byla provedena i jeho validace, kdy byl původní soubor dat rozdělen na 10 podsouborů a následně model sestaven na 9 a na posledním podsouboru byl aplikován. Touto metodou byla zjištěna výkonnost modelu 80,39 % neboli z 10 000 zákazníků jich bude správně zařazeno do příslušné kategorie 8 039 a 1 961 jich bude zařazeno chybně. Tato přesnost byla stvrzena i křivkou ROC, která se relativně brzy odchyluje od osy y.

4.4.2 **Rozhodovací strom a náhodný les**

4.4.2.1 Rozhodovací strom

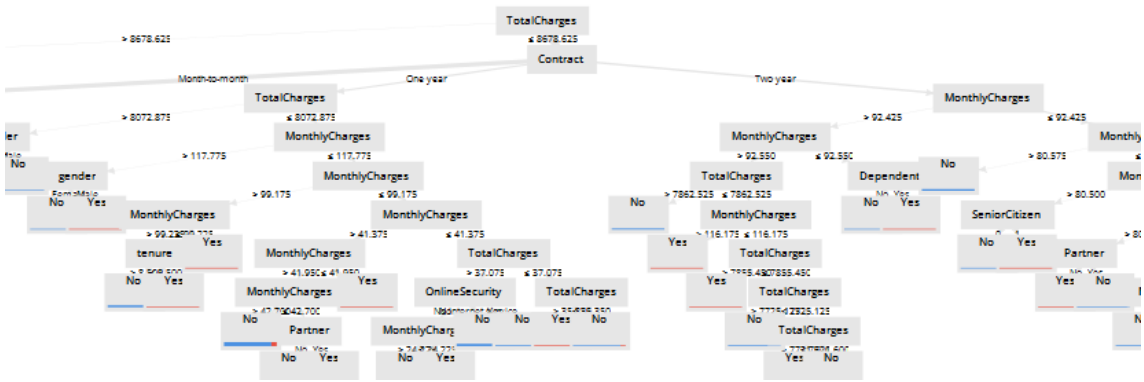
Druhým typem modelů aplikovaných na daný datový soubor jsou rozhodovací stromy. Nejprve bude aplikován základní rozhodovací strom a poté náhodný les a bude porovnána přesnost jejich předpovědí, a to nejen mezi sebou, ale i v porovnání s logistickou regresí. Opět bude k práci využit soubor, ve kterém už jsou role proměnných určeny. Co je naopak nutností je doplnění chybějících hodnot u proměnné celkové platby, čehož je opět docíleno pomocí nástroje Replace Missing Values, kde hodnoty byly nahrazeny průměrem dostupných záznamů. Proces pro tvorbu rozhodovacího stromu v Rapidmineru vypadá následovně:



Obrázek 27 Schéma procesu rozhodovacího stromu. Zdroj: Vlastní zpracování.

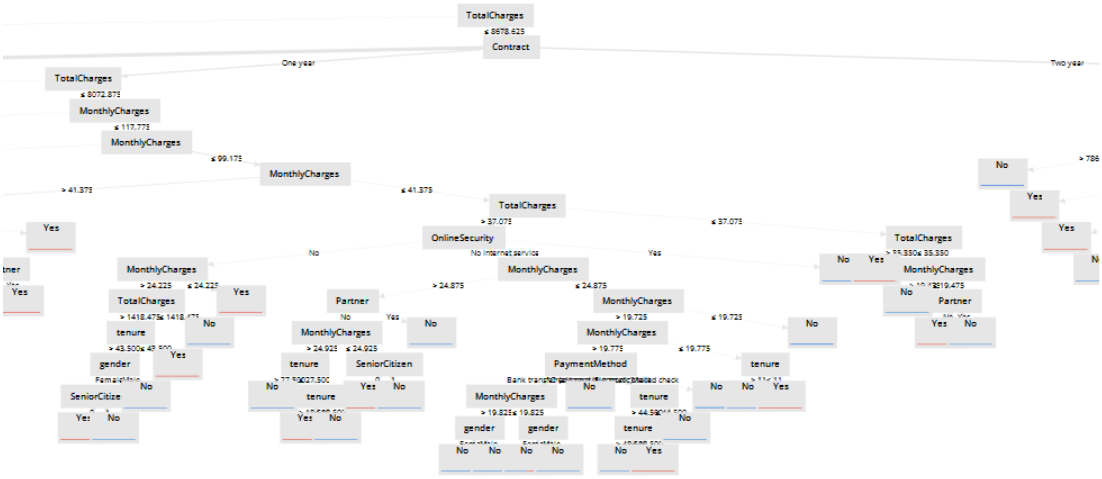
To, jak bude výsledný strom vypadat, lze ovlivnit několika parametry v nastavení rozhodovacího stromu. Lze zvolit maximální hloubka stromu, což je jinými slovy maximální počet úrovní uzlů ve stromu. Čím je hloubka větší, tím se stává strom méně přehledným, poněvadž obsahuje velké množství uzlů, ale celková předpovídací schopnost modelu se se zvyšováním maximální hloubky modelu zvyšuje. Je to logické, protože se zvyšujícím se počtem zařazovacích kritérií se zvyšuje počet možností pro správné zařazení. Lze například nastavit i neomezený počet úrovní rozhodovacího stromu. V takovém případě je sice přesnost modelu 100 %, nicméně strom je těžce použitelný, a to kvůli nepřehlednosti. Toto nastává hlavně v případě numerických proměnných, jako je v tomto případě výše plateb či délka kontraktu. Každá z těchto proměnných má velké množství možných hodnot a vzhledem k neomezenému počtu úrovní stromu je každá hodnota vedena jako jeden list stromu. Rozhodovací strom byl sestaven dvakrát, a to jednou s 10 a jednou s 15 podúrovněmi k názorné ukázce, jak tento faktor ovlivňuje celý model. Další možností úpravy modelu je omezení modelu (pruning). Tento parametr do stromu nezařadí nadbytečné parametry. To, jestli je daný parametr nadbytečný, je určeno hladinou významnosti. Přednastavená je hladina 0,25, nicméně při ponechání této hladiny obsahuje strom pouze jeden uzel. Při snížení hladiny na 0,05 je výstupem strom, který obsahuje pouze relevantní faktory. Schéma rozhodovacího stromu je k nalezení v příloze č. 2.

Kořenovým uzlem tohoto rozhodovacího stromu je proměnná Celkové platby. Pokud je hodnota této proměnné vyšší než 8678,25 USD, nastane odchod zákazníka. Vzhledem k tomu, že takový zákazník se v celém souboru nacházel pouze jeden, je to velmi zavádějící. Pokud je hodnota proměnné Celkové platby menší než tato hodnota, následuje posun do dalšího uzlu. Dalším uzlem, kterému lze přiřadit větší váhu, obsahuje proměnnou Smlouva.



Obrázek 28 Výřez rozhodovacího stromu s 10 úrovněmi. Zdroj: IBM, vlastní zpracování.

Tento uzel se pak větví do tří dalších uzlů, a to podle typu kontraktu – smlouva z měsíce na měsíc, smlouva na 1 rok a smlouva na 2 roky.



Obrázek 29 Výřez rozhodovacího stromu s 15 úrovněmi. Zdroj: IBM, vlastní zpracování.

Při znalosti informací o zákazníkovi je možné pomocí tohoto stromu dojít s danou pravděpodobností k závěru, jestli daný zákazník u společnosti zůstane, nebo ji naopak opustí. Software Rapidminer poskytuje funkci zhodnocení váhy jednotlivých proměnných v rozhodovacím stromu. Právě pomocí určení hladiny významnosti při jeho konfiguraci se z modelu dají vyloučit některé proměnné. Pro strom, kde je 10 podúrovní, jsou proměnné seřazeny v tabulce 5, pro strom s 15 podúrovněmi jsou proměnné v tabulce 6.

Proměnná	**Váha proměnné**
Pohlaví	0,2480
Celkové platby	0,1995
Měsíční platby	0,1946
Partner	0,1653
Důchodce	0,0827
Závislí	0,0826
Délka smlouvy	0,0178
Smlouva	0,0082
Online bezpečnost	0,0013

Tabulka 5 Váha proměnných rozhodovacího stromu (10 podúrovní). Zdroj. IBM, vlastní zpracování.

Proměnná	Váha proměnné
Celkové platby	0,2237
Měsíční platby	0,1763
Pohlaví	0,1409
Délka smlouvy	0,1401
Partner	0,0997
Důchodce	0,0969
Online bezpečnost	0,0809
Závislí	0,0323
Technická podpora	0,0035
Smlouva	0,0032
Filmy online	0,0020
Platební metoda	0,0004

Tabulka 6 Váha proměnných rozhodovacího stromu (15 podúrovní). Zdroj: IBM, vlastní zpracování.

Tabulky jsou velmi podobné, avšak jednotlivé váhy stejných proměnných jsou odlišné a při použití 15 podúrovní je proměnnou s nejvyšší váhou Celkové platby, zatímco v modelu s 10 podúrovněmi to je Pohlaví.

Dále je zapotřebí změřit výkonnost modelu neboli přesnost předpovědí. To je opět provedeno pomocí operátoru Performance a následně je provedena validace napříč 10 podsoubory. Výkonnosti modelu velmi úzce souvisí s počtem umožněných podúrovní, a tak jsou rozdíly mezi modelem s 10 a modelem s 15 podúrovněmi lehce viditelné. Výkonnost modelu s 10 podúrovněmi je v tabulce 7, výkonnost modelu s 15 podúrovněmi pak v tabulce 8.

accuracy: 73.79%

	true No	true Yes	class precision
pred. No	5174	1846	73.70%
pred. Yes	0	23	100.00%
class recall	100.00%	1.23%	

Tabulka 7 Výkonnost rozhodovacího stromu (10 podúrovní). Zdroj: IBM, vlastní zpracování.

accuracy: 74.16%

	true No	true Yes	class precision
pred. No	5174	1820	73.98%
pred. Yes	0	49	100.00%
class recall	100.00%	2.62%	

Tabulka 8 Výkonnost rozhodovacího stromu (15 podúrovní). Zdroj: IBM, vlastní zpracování.

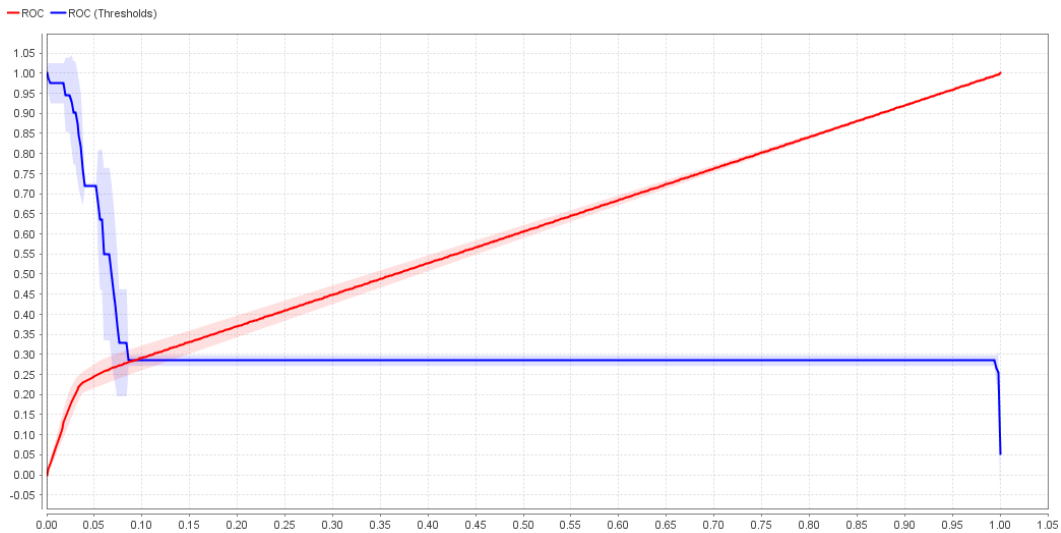
Při použití 10 podúrovní je přesnost zařazení 73,79 %, při navýšení počtu podúrovní se také navýší jeho přesnost. Z tohoto důvodu bude dále počítáno již jen s modelem s 15 podúrovněmi. Celková výkonnost modelu s 15 podúrovněmi testovaná na celém souboru dat je 74,16 %, přičemž model odhalil sice jen velmi malé procento odcházejících zákazníků (2,62 %), nicméně ani jeden zákazník nebyl modelem špatně označen jako odcházející. Při aplikaci validace napříč 10 soubory vypadala výkonnost modelu (15 podúrovní) následovně:

accuracy: 77.01% +/- 0.81% (mikro: 77.01%)

	true No	true Yes	class precision
pred. No	5008	1453	77.51%
pred. Yes	166	416	71.48%
class recall	96.79%	22.26%	

Tabulka 9 Výkonnost stromu při využití cross-validation. Zdroj: IBM, vlastní zpracování.

Výkonnost modelu narostla na 77,01 % a směrodatná odchylka měření je 0,81 %, což opět značí velmi stabilní model. Citlivost modelu je 22,26 %, což značí proporci správně zařazených pravých pozitiv a specifičnost 96,79 %, která značí proporci správně zařazených pravých negativ. Z těchto tabulek je vidět, že na daném souboru dat má větší přesnost



Obrázek 30 ROC křivka rozhodovacího stromu. Zdroj: IBM, vlastní zpracování.

zařazení logistická regrese, což je také potvrzeno ROC křivkou, která je od bodu (0,1) mnohem dál než ta u logistické regrese.

Celková přesnost předpovědí rozhodovacího stromu s 15 podúrovněmi je tedy 77,01 %. Při znalosti zákaznických dat lze tedy postupovat po uzlech rozhodovacího stromu a s více než 77 % přesností bude zákazník správně zařazen buď jako zůstávající či jako odcházející. Rozhodovací strom správně označil pouze 22 % odcházejících zákazníků, nicméně jako zůstávající správně označil necelých 97 % zákazníků. Hlavním uzlem je v případě tohoto stromu Výše celkových plateb následovaný uzlem Smlouva, který strom rozvětví na tři hlavní větve podle druhu smlouvy. Největší váhu má proměnná Celkové platby a druhou největší váhu Měsíční platby.

4.4.2.2 Náhodný les

Druhým typem modelu sestaveným v této kategorii bude náhodný les. Nástroj Náhodný les nabízí mnoho možností specifikace daného modelu. Tím základním nastavením je počet vygenerovaných stromů, kde přednastavený počet je 10. Dále lze nastavit, podle kterého kritéria budou proměnné vybírány na větvení ve stromech. Zde je možnost několika variant. Tou první je information gain neboli informační zisk. Při zvolení této varianty software spočítá entropie jednotlivých proměnných (průměr informací, které vygenerují) a proměnná s nejmenší hodnotou entropie je určena k rozvětvení. Další možností je poměr zisku, což je obdoba předešlé možnosti s tím rozdílem, že informační zisk je přizpůsoben každé proměnné, tak aby byla umožněna šíře a uniformita hodnot proměnných. Dále lze proměnné větvit například podle toho, které nejvíce přispívají k přesnosti celého modelu. Následujícím nastavením modelu je, stejně jako u rozhodovacího stromu, počet podúrovní a hladina významnosti.

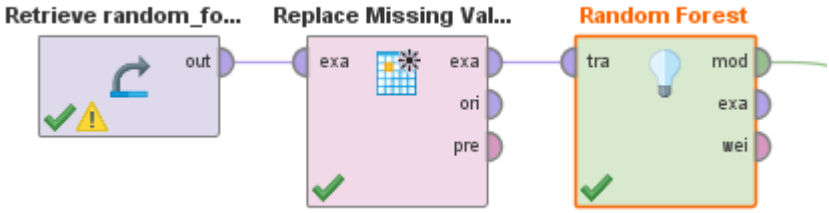
Je patrné, že možností nastavení náhodného lesa je mnoho a určení toho správného má velký vliv na to, jak bude výsledný model vypadat i na to, jaká bude jeho přesnost. Software Rapidminer nabízí nástroj i pro tento problém a jmenuje se Optimize Parameters. Tento nástroj doporučí optimální nastavení modelu na základě dostupných dat. Je však nutné specifikovat, pro jaký model má operátor parametry specifikovat a také jaké parametry mají být specifikovány. V případě tohoto modelu byly optimalizovány parametry pro počet stromů, hloubku modelu a hladinu významnosti.

Výsledky této optimalizace jsou následovné:

Random Forest.number_of_trees	25
Random Forest.maximal_depth	40
Random Forest.confidence	0,0943845023899301

Tabulka 10 Výsledek optimalizace parametrů náhodného lesa. Zdroj: IBM, vlastní zpracování.

Z tabulky číslo 10 vyplývá, že optimální počet stromů v náhodném lese je 25, přičemž každý z těchto stromů má maximální počet podúrovní 40. Jako optimální hladina významnosti byla určena hladina 0,0944. Pro vytvoření modelu náhodného lesa byly použity právě tyto parametry. Schéma procesu je na obrázku číslo 31.



Obrázek 31 Schéma procesu náhodného lesa. Zdroj: Vlastní zpracování.

Vzhledem k vysokému počtu jednotlivých stromů (25) v náhodném lese nelze určit jeden hlavní kořenový uzel, neboť každý rozhodovací strom je jiný. Náhodné lesy jsou z tohoto důvodu mnohem složitější k porozumění a náročnější na použití člověkem, a proto se model aplikuje převážně elektronicky na data obsahující informace o zákaznících nepoužitých pro tvorbu modelu. Z modelu lze určit váhu jednotlivých proměnných, které jsou uvedeny v tabulce 11.

Proměnná	Váha proměnné
Měsíční platby	0,2194
Délka kontraktu	0,1837
Celkové platby	0,1757
Pohlaví	0,0592
Partner	0,0451
Důchodce	0,0353
Více linek	0,0349
Závislí	0,0321
Ochrana zařízení	0,0318

Online záloha	0,0290
Online vyúčtování	0,0288
Způsob platby	0,0281
Sledování TV	0,0233
Sledování filmů	0,0221
Technická podpora	0,0162
Online bezpečnost	0,0141
Internet	0,0083
Telefon	0,0072
Smlouva	0,0056

Tabulka 11 Váhy jednotlivých proměnných v náhodném lese. Zdroj: IBM, vlastní zpracování.

Z tabulky vyplývá, že největší váhu má proměnná Měsíční platby, druhou v pořadí je Délka smlouvy a třetí je proměnná Celkové platby. Lze tedy říci, že v modelu náhodného lesa mají tyto tři proměnné největší vliv na to, zdali zákazník u dané společnosti zůstane nebo odejde.

Výkonnost modelu byla opět změřena dvakrát, a to jak na celém souboru dat využitých k sestavení modelu, tak i pomocí validace napříč podsoubory, kde bylo využito rozdělení hlavního soboru na 10 částí.

accuracy: 98.20%

	true No	true Yes	class precision
pred. No	5157	110	97.91%
pred. Yes	17	1759	99.04%
class recall	99.67%	94.11%	

Tabulka 12 Výkonnost náhodného lesa. Zdroj: IBM, vlastní zpracování.

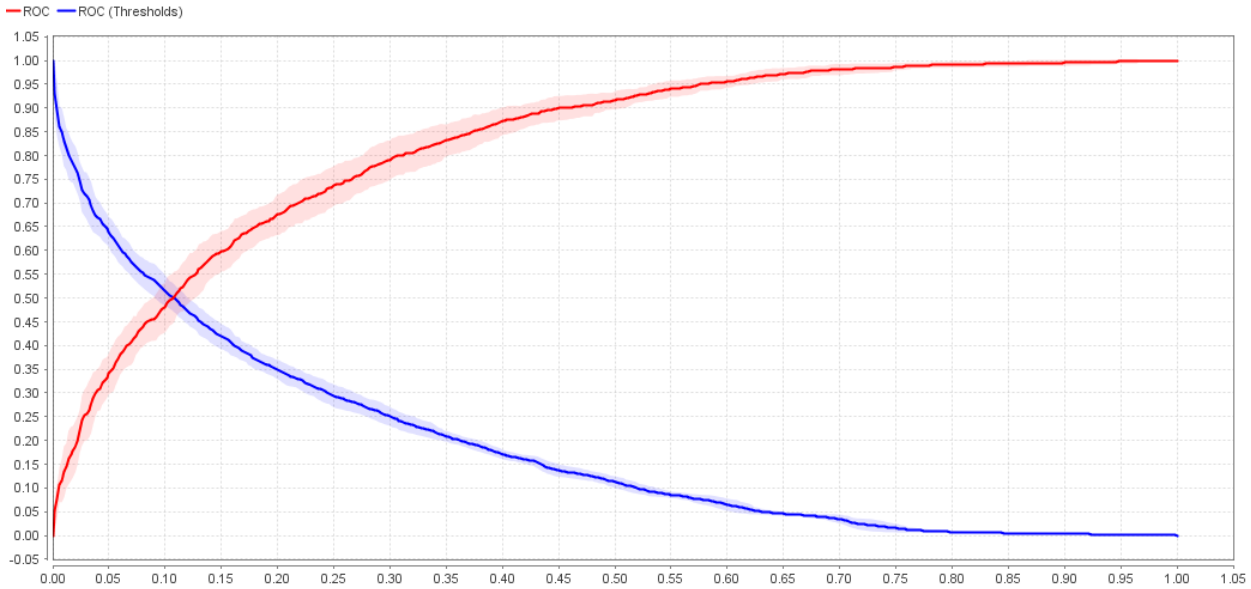
Při měření výkonnosti na celém souboru dat vyšla velmi vysoká výkonnost modelu, a to 98,20 %. Z těchto výsledků je patrné, že nastavení modelu bylo velmi dobré, neboť přesnost predikcí je skoro 100 %. Nicméně tato výkonnost je zavádějící z důvodu nepoužití neznámých dat. Pro odhalení skutečné výkonnosti je zapotřebí provést validaci napříč podsoubory.

accuracy: 79.11% +/- 1.33% (mikro: 79.11%)

	true No	true Yes	class precision
pred. No	4631	928	83.31%
pred. Yes	543	941	63.41%
class recall	89.51%	50.35%	

Tabulka 13 Výkonnost modelu při využití cross-validation. Zdroj: IBM, vlastní zpracování.

Je patrné, že výkonnost razantně poklesla, a to o více jak 19 %. Výkonnost 79,11 % je však průkaznější a lze ji považovat za skutečnou výkonnost modelu, neboť byl model aplikován na předem nepoznaná data, tudíž lze předpokládat, že s touto přesností by zařazoval zákaznické záznamy předem neobsažené v databázi. Směrodatná odchylka měření je 1,33 % značící stabilitu modelu. Model správně zařadil 50,35 % pravých pozitiv a 89,51 % pravých negativ. ROC křivka na obrázku 32 opět ukazuje vztah mezi mírou pravých pozitiv na ose y a mírou nepravých pozitiv na ose x. Křivka stále není ideální, neboť je daleko od bodu (0,1).



Obrázek 32 ROC křivka náhodného lesa. Zdroj: IBM, vlastní zpracování.

4.4.2.3 Shrnutí rozhodovacího stromu a náhodného lesa

Nastavení obou modelů má velký dopad na jejich výkonnost. U rozhodovacího stromu byla při 15 podúrovních docílena výkonnost 75,28 %. Nebyla zde provedena optimalizace parametrů jako tomu bylo u náhodného lesa, a to z toho důvodu, že výsledkem optimalizace bylo využití 56 podúrovní, což by sice zaručilo vyšší předpovídací přesnost

modelu, nicméně rozhodovací stromy se používají nejen stroji, ale také lidmi, pro něž by použití 56 podúrovní udělalo model příliš složitý a nepoužitelný. Oproti tomu náhodný les, kde bylo sestaveno 25 rozhodovacích stromů s maximální hloubkou 40 a hladinou významnosti 0,0944, měl přesnost předpovědí 79,11 %. Tento model lze použít převážně za pomoci výpočetní techniky, kdy model s 79,11 %-ní přesností zařadí zákazníka do příslušné kategorie.

Nejvýznamnějšími faktory, které v náhodném lese ovlivňují příslušnost zákazníka, jsou výše měsíčních plateb, délka kontraktu v měsících a celkové platby. Naproti tomu v jednoduchém rozhodovacím stromu těmito faktory jsou pohlaví, celkové a měsíční platby. Na základě těchto výsledků mohou společnosti podstoupit řádná opatření, aby co nejvíce snížily počty odcházejících zákazníků. Výsledky budou hlouběji prozkoumávány v další části práce.

4.4.3 Neuronová síť

Posledním modelem aplikovaným na daný datový soubor je neuronová síť. Jak je již zmíněno dříve, neuronová síť je systém, který je inspirován biologickými neuronovými sítěmi, které jsou přítomné ve zvířecích mozcích. Umělá neuronová síť za pomocí algoritmů nachází vztahy mezi vstupními proměnnými a cílovou. V tomto případě je cílovou proměnnou odchod zákazníka, což je binomická proměnná. Z toho plyne, že neuronová síť bude zakončena dvěma neurony.

Před sestavením samotného modelu je opět nutné provést přípravu dat tak, aby model mohl být spuštěn. Doplnění chybějících hodnot u proměnné Celkové náklady je samozřejmostí, v případě neuronové sítě však přibyla jedna nutná úprava, a to převedení nominálních hodnot na hodnoty numerické. Neuronová síť totiž neumí pracovat s polynominálními hodnotami a je tak třeba využít nástroje Nominal to Numerical, který tento krok provede.

Obrázek 33 Schéma přípravy dat pro neuronovou síť. Zdroj: Vlastní zpracování.

Jakmile jsou data pro neuronovou síť připravena, lze přejít k nastavení parametrů neuronové sítě. Rapidminer nabízí celou řadu možností, jak model nastavit. V první řadě lze nastavit počet skrytých vrstev, kde uživatel určí jméno skryté vrstvy a také její velikost, což je počet neuronů obsažených v této vrstvě. Dále je možné nastavit počet tréninkových cyklů, který určí počet cyklů použitých k vytvoření modelu. Počet cyklů určuje to, kolikrát se bude opakovat proces, ve kterém jsou výsledné hodnoty porovnány s opravdovými hodnotami, z čehož je následně vypočítána chybovost a ta je následně poslána zpět skrz neuronovou síť a na základě této informace se pak model snaží pozměnit váhy jednotlivých spojů tak, aby chyba byla ve výsledku co nejmenší. Pro tento model nebyla použita žádná skrytá vrstva a 500 cyklů. Výsledná neuronová síť je v příloze číslo 3.

Výkonnost modelu byla opět změřena dvakrát, a to jak na celém souboru dat využitých k sestavení modelu, tak i pomocí validace napříč podsoubory, kde bylo využito rozdělení hlavního soboru na 10 částí.

accuracy: 86.10%

	true No	true Yes	class precision
pred. No	4688	493	90.48%
pred. Yes	486	1376	73.90%
class recall	90.61%	73.62%	

Tabulka 14 Výkonnost neuronové sítě. Zdroj: IBM, vlastní zpracování.

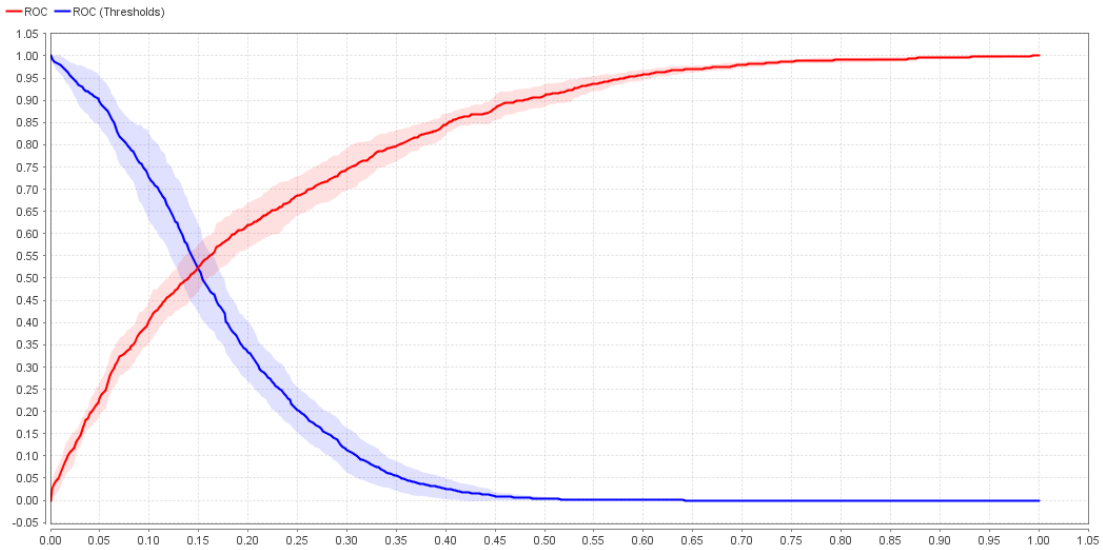
Výkonnost neuronové sítě není zdaleka tak vysoká, jako tomu bylo u náhodného lesa, nicméně byla lepší než u všech ostatních modelů. Při měření výkonnosti sítě na celém souboru zkušebních dat byla přesnost klasifikace 86,20 %.

accuracy: 76.40% +/- 1.67% (mikro: 76.40%)

	true No	true Yes	class precision
pred. No	4382	870	83.43%
pred. Yes	792	999	55.78%
class recall	84.69%	53.45%	

Tabulka 15 Výkonnost sítě při využití cross validation. Zdroj: IBM, vlastní zpracování.

Při zkoumání výkonnosti na předem neviděných datech nastal pokles o necelých 10 % na výslednou přesnost klasifikace 76,40 %. Směrodatná odchylka měření byla 1,67 %, jednalo se opět o stabilní model. Neuronová síť správně zařadila 54,45 % pravých pozitiv a 84,69 % pravých negativ. ROC křivka na obrázku 34 znázorňuje vztah mezi mírou pravých pozitiv na ose y a mírou nepravých pozitiv na ose x a je z ní patrné, že se opět nejedná o optimální model, neboť plocha nad křivkou je relativně velká.



Obrázek 34 ROC křivka neuronové sítě. Zdroj: IBM, vlastní zpracování.

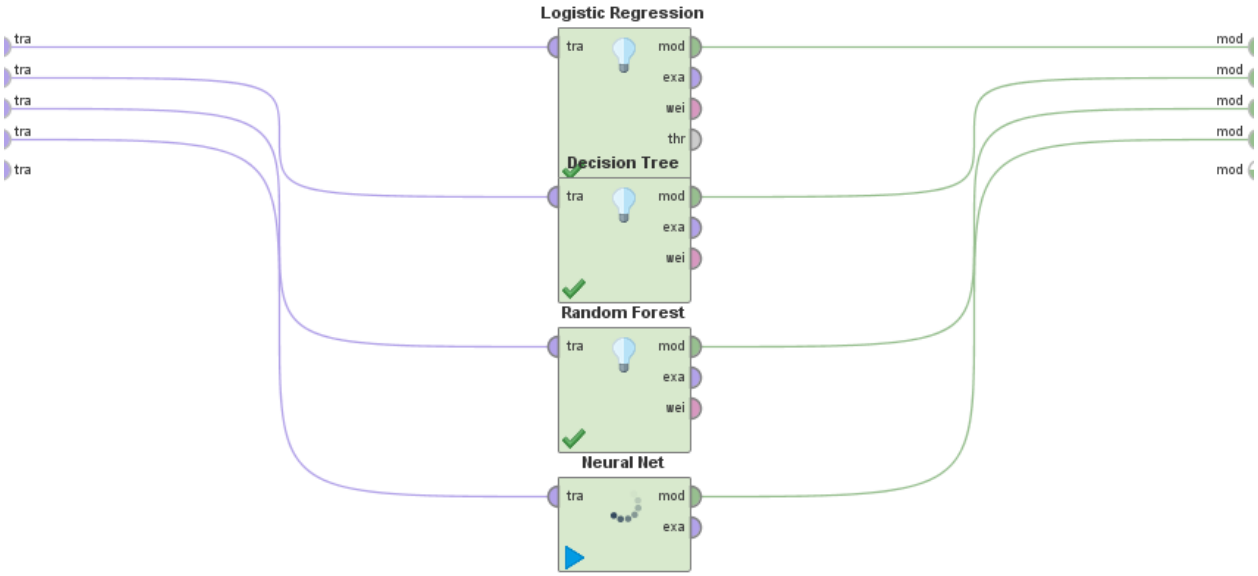
4.4.3.1 Shrnutí neuronové sítě

Neuronová síť poskytla přesnost klasifikace 76,40 %, což je méně než logistická regrese či náhodný les. Opět platí pravidlo, že nastavení modelu má zásadní vliv na to, jak dobře bude model fungovat. Při nastavení 500 cyklů a žádné skryté vrstvy vyšla výkonnost 76,40 %, při nastavení vyššího počtu cyklů nebo skrytých vrstev však již nedostačoval výpočetní výkon a model tak nemohl být spuštěn a změřena jeho výkonnost (proces cross-validation s pouze 500 cykly trval zhruba půl hodiny).

4.5 Srovnání a vyhodnocení modelů

Vzhledem k použití ROC křivek ke zjištění přesnosti zařazení u jednotlivých modelů byla tato metoda použita i pro jejich celkové srovnání. Software Rapidminer přímo k tomu nabízí nástroj, který se nazývá Compare ROCs neboli srovnání ROC křivek. Stejně tak jako ostatní operátory, i tento má mnoho možností nastavení, které významně ovlivňují, jak

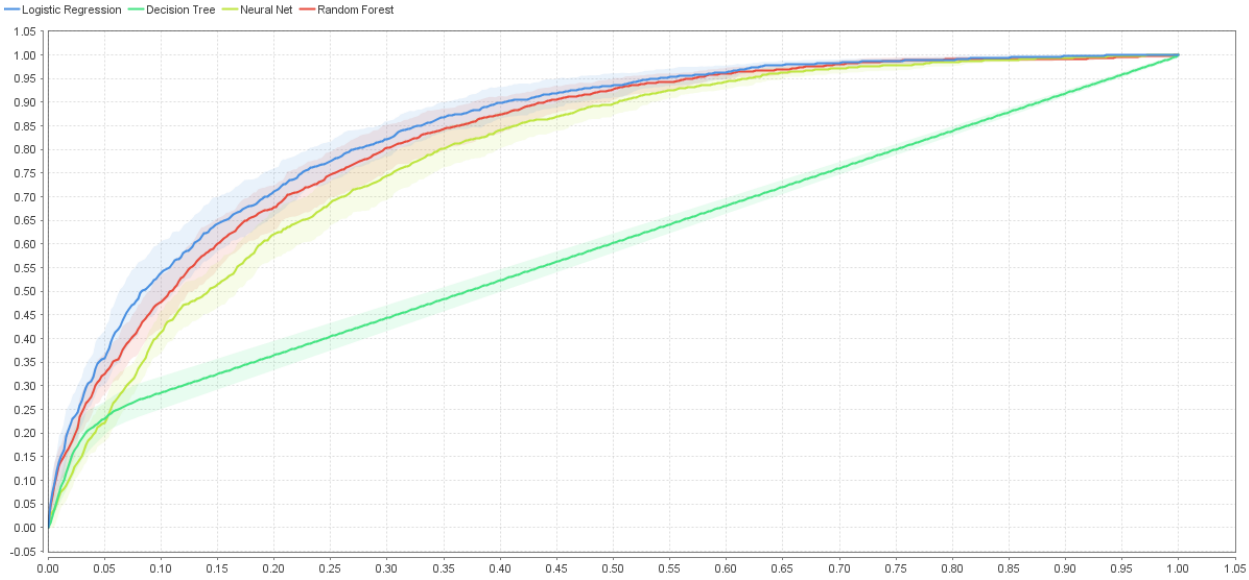
budou dané křivky vypadat. Lze tak nastavit, jestli budou křivky nějakým způsobem zaujaty (bias), a to buď optimisticky, pesimisticky, či neutrálně. Vzhledem k použití neutrálního zaujetí u všech předchozích křivek, bylo toto nastavení ponecháno i zde. Do tohoto operátoru jsou zahrnuty všechny modely, a to logistická regrese, rozhodovací strom s hloubkou 15 a hladinou významnosti 0,05, náhodný les s 25 stromy s hloubkou 40 a hladinou významnosti 0,0944 a neuronová síť s 500 cykly. Schéma procesu je na obrázku 35. Tento nástroj provede validaci napříč deseti podsoubory pro každý model a výsledky zobrazí v jednom grafu. Je to výpočetně náročný proces, neboť validace probíhá desetkrát pro každý ze čtyř modelů. Vzhledem k potřebě numerických dat u neuronové sítě byly hodnoty celého souboru pomocí operátoru Nominal to Numerical převedeny na numerické hodnoty.



Obrázek 35 Schéma podprocesu srovnání ROC křivek. Zdroj: Vlastní zpracování.

Výsledkem je graf, na kterém jsou vidět ROC křivky jednotlivých modelů, které jsou barevně odlišeny. Logistická regrese je modře, rozhodovací strom zeleně, náhodný les červeně a neuronová síť žlutě. Čím lepší je předpovídací schopnost modelu, tím víc se jeho křivka přibližuje bodu (0,1). Výsledek je na obrázku 36. Z grafu, stejně tak jako z dosavadního průběhu modelování, je patrné, že na daném souboru dat má největší přesnost zařazení logistická regrese. I když Hassouna et al. (2015) ve svém článku zmiňují, že rozhodovací strom je v případě jejich výzkumu tím lepším modelem, v tomto případě logistická regrese svou přesností zařazení zákazníků předčila všechny ostatní modely. Logistická regrese je ze všech čtyř modelů nejlepší i co se týče citlivosti, což je pro takovouto

společnost důležitým faktorem. Citlivost logistické regrese byla v tomto případě 55,16 %. Druhým nejlepším modelem byl podle ROC křivky náhodný les a podle citlivosti neuronová síť.



Obrázek 36 Porovnání ROC křivek jednotlivých modelů. Zdroj: IBM, vlastní zpracování.

V následující tabulce jsou shrnuty přesnosti, citlivosti a specifičnosti všech zkoumaných modelů.

Model/Kritérium	Přesnost	Citlivost	Specifičnost
Logistická regrese	80,39 %	55,16 %	89,51 %
Rozhodovací strom	77,01 %	22,26 %	96,79 %
Náhodný les	79,11 %	50,35 %	89,51 %
Neuronová síť	76,40 %	53,45 %	84,69 %

Tabulka 16 Srovnání modelů. Zdroj: IBM, vlastní zpracování.

Z tabulky vyplývá, že logistická regrese je opravdu tím nejlepším modelem, který byl na daný soubor použit. Dle srovnání ROC křivek je druhým nejlepším modelem náhodný les, následovaný rozhodovacím stromem a nejhorším modelem pro tento soubor dat byla vyhodnocena neuronová síť. Kdyby se však modely srovnávaly podle citlivosti, neuronová síť by zaujímala druhé místo za logistickou regresí. Pro další část práce tedy bude využita logistická regrese, jakožto nejlepší model na daný soubor.

4.6 Aplikace logistické regrese

Statisticky významné faktory, které mají na chování zákazníka, v tomto případě jeho odchod či setrvání, největší vliv, jsou podle výsledků logistické regrese tyto (seřazeny jsou sestupně dle velikosti vlivu):

- Internet-optický kabel (x_{10}) – koeficient této proměnné je 1,753, což značí zvyšující se šanci zákazníkova odchodu, pokud má internet veden právě přes optický kabel. Šance se zvyšuje 5,7719krát.
- Délka smlouvy (x_1) – koeficient této proměnné je -0,059 neboli s každým dalším měsícem, který zákazník stráví pod smlouvou u daného poskytovatele, se šance jeho odchodu snižuje, a to 0,9432krát.
- Smlouva na 2 roky (x_2) – koeficient této proměnné je -1,396. Pokud je zákazník vázán dvouletou smlouvou, šance jeho odchodu se snižuje 0,2475krát.
- Celkové platby (x_5) – koeficient této proměnné je 0,0003, tudíž s každým dolarem, který zákazník zaplatí navíc, se šance jeho odchodu zvyšuje, a to 1,0003krát.
- Smlouva na 1 rok (x_3) – koeficient této proměnné je -0,699. Pokud je tedy zákazník vázán smlouvou na 1 rok, šance jeho odchodu se snižuje 0,5125krát.

Další statisticky významné proměnné v logistické regresi jsou tyto:

- Online vyúčtování-Ne (x_4) – velikost koeficientu této proměnné je -0,342. Pokud zákazník nevyužívá online vyúčtování, šance jeho odchodu se snižuje 0,9664krát.
- Platební metoda-karta (automaticky) (x_6) – koeficient této proměnné je -0,392, takže pokud zákazník platí automaticky kartou, šance jeho odchodu se snižuje 0,6754krát.
- Platební metoda-poštovní šek (x_7) – pokud zákazník platí poštovním šekem, šance jeho odchodu se snižuje 0,6966krát.
- Platební metoda-trvalý příkaz (x_8) – pokud zákazník využívá trvalého příkazu k platbě, šance jeho odchodu se snižuje 0,7362krát.
- Důchodce-Ano (x_9) – pokud je zákazník důchodcem, šance jeho odchodu se zvyšuje 1,2395krát.

Celá rovnice logistické regrese, která zahrnuje pouze statisticky významné proměnné, pak vypadá následovně:

$$\text{logit}(y) = -0{,}0585x_1 - 1{,}3962x_2 - 0{,}6685x_3 - 0{,}3416x_4 + 0{,}0003x_5 - 0{,}3924x_6 - 0{,}3615x_7 - 0{,}3062x_8 + 0{,}2147x_9 + 1{,}7530x_{10}$$

Rovnice 7 Logistická regrese. Zdroj: Vlastní zpracování.

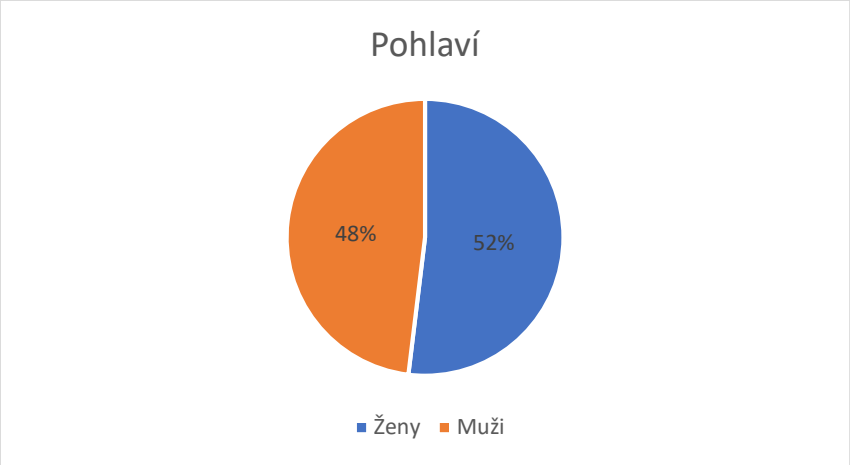
Dalším krokem je aplikace modelu. Ta proběhne aplikováním modelu na původní soubor dat, kdy se pomocí logistické rovnice bude u jednotlivých zákazníků určovat to, jestli odejdou či ne. Opět je využito jednoho z nástrojů softwaru Rapidminer, a to Apply model. Výsledkem tohoto procesu je tabulka, která je velmi podobná původnímu souboru dat, jen s tím rozdílem, že navíc obsahuje tři sloupce. Tím prvním je předpověď toho, jestli zákazník odejde či ne. Tato předpověď je vlastně cílem celého modelování a určí se tak, že se jednotlivé informace o zákazníkovi dosadí do logistické regrese vytvořené modelováním, a model s určitou pravděpodobností zákazníka zařadí do dané kategorie. Další dva nové sloupce v souboru jsou jistoty zařazení, jeden sloupec tedy ukazuje, s jakou jistotou byl daný zákazník označen jako zůstávající, a druhý sloupec ukazuje, s jakou jistotou byl zákazník označen jako odcházející. Nejdůležitější z nových sloupců je sloupec předpovědi, který ukazuje, kam byl zákazník na základě jeho informací zařazen. Na základě tohoto zařazení byly určeny obecné charakteristiky zákazníků jak těch, kteří zůstávají, tak těch, kteří odcházejí. Díky těmto obecným charakteristikám bude lehčí identifikovat zákazníky, u kterých je vysoká pravděpodobnost odchodu.

	ANO	**NE**
Odchod-skutečnost	1869	5174
Odchod-předpověď	1558	5485

Tabulka 17 Porovnání skutečnosti s předpovědí. Zdroj: IBM, vlastní zpracování

Z tabulky 17 je vidět, kolik celkem bylo modelem určeno odcházejících a zůstávajících zákazníků. Informace o 1558 zákaznících, kteří byli modelem označeni jako odcházející, budou využity k vytvoření obecných charakteristik odcházejícího zákazníka u

všech proměnných. Nejdříve bude pozornost věnována binominálním proměnným, pak polynominálním a poté proměnným nabývajících číselných hodnot.



Obrázek 37 Graf pohlaví odcházejících zákazníků. Zdroj: IBM, vlastní zpracování.

Z grafu v obrázku 37 je vidět, že pohlaví zákazníků, kteří odcházejí, je relativně rovnoměrné, kdy z 1558 zákazníků, kteří podle modelu odešli, bylo 809 (51,93 %) žen a 749 (48,07 %) mužů.

Jak ukázala důležitost jednotlivých proměnných, způsob připojení k internetu je faktorem, který má na zákazníka velký vliv. Obrázek 38 se tedy zaměřuje právě na tuto proměnnou ve vztahu k odcházejícím zákazníkům. Z celkového počtu 1558 odcházejících zákazníků jich 1371 mělo internet zřízený přes optický kabel a pouze 187 přes DSL. Je tedy patrné, že připojení k internetu přes optický kabel je faktorem, který má velký vliv na odchod zákazníků.

Obrázek 38 Graf rozdělující odcházející zákazníky dle způsobu připojení k internetu. Zdroj: IBM, vlastní zpracování.

Další binominální proměnné jsou zobrazeny v tabulce 18.

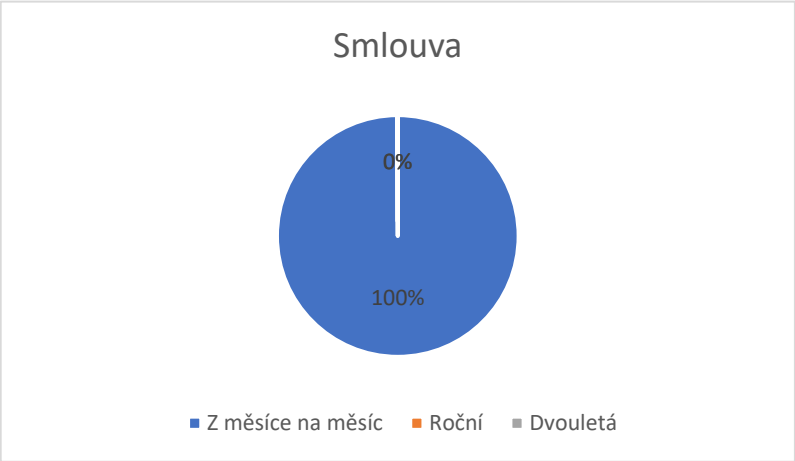
Proměnná/Hodnota	ANO	NE
Důchodce	489 (31,39 %)	1069 (68,61 %)
Partner	461 (29,59 %)	1097 (70,41 %)
Závislí	171 (10,98 %)	1387 (89,02 %)
Telefonní služba	1426 (91,53 %)	132 (8,47 %)
Online bezpečnost	93 (5,97 %)	1465 (94,02 %)
Online záloha	337 (21,63 %)	1221 (78,37 %)
Ochrana zařízení	382 (24,52 %)	1176 (75,48 %)
Technická podpora	134 (8,60 %)	1424 (91,40 %)
Sledování TV online	728 (46,73 %)	830 (53,27 %)
Sledování filmů online	744 (47,75 %)	814 (52,25 %)
Online vyúčtování	1313 (84,27 %)	245 (15,73 %)

Tabulka 18 Počty odcházejících zákazníků dle daných proměnných. Zdroj: IBM, vlastní zpracování.

Z tabulky vyplývají následující informace:

- Většina ze zákazníků, kteří od daného poskytovatele služeb v posledním měsíci odešli, využívala online vyúčtování (přes 84 % ze všech odchozích zákazníků) a dále jich také většina využívala telefonních služeb (přes 91 %). Nicméně je nutné zmínit, že celkem 6361 zákazníků z celkového počtu 7043 využívá telefonních služeb.
- Přes 89 % odchozích zákazníků nemá na sobě závislé členy rodiny.
- Většina odchozích zákazníků nevyužívá online služby. Přes 94 % jich nevyužívá online bezpečnosti, více jak 78 % nevyužívá online zálohy, nad tři čtvrtiny odchozích zákazníků nevyužívá ochrany zařízení a více jak 91 % z nich nevyužívá technické podpory.
- Více jak 68 % odchozích nejsou zákazníci důchodového věku a přes 70 % jich nemá partnera.

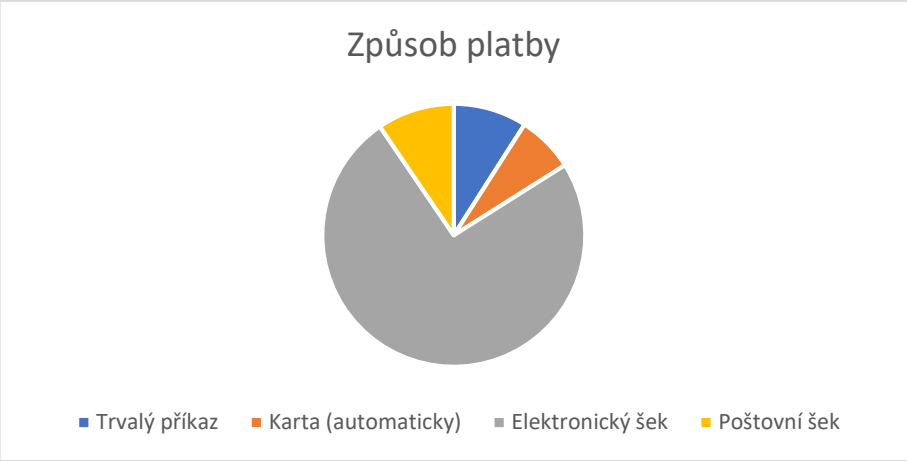
V další části jsou analyzovány polynominální proměnné ve vztahu k odchozím zákazníkům.



Obrázek 39 Graf rozdělující odchozí zákazníky podle typu smlouvy. Zdroj: IBM, vlastní zpracování.

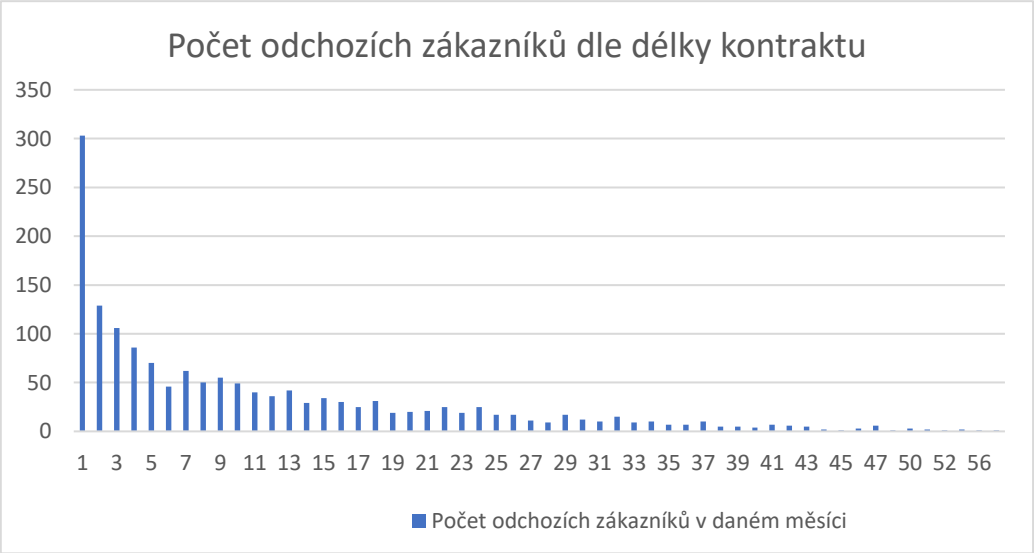
Z celkového počtu 1558 odchozích zákazníků měl pouze jeden roční smlouvu, ani jeden neměl dvouletou smlouvu a 1557 jich mělo smlouvu z měsíce na měsíc. Je tedy patrné, že nejvíce odchozích zákazníků je se smlouvou z měsíce na měsíc neboli nemají žádný závazek. V původním souboru bylo celkem 3875 zákazníků právě s tímto typem smlouvy, 1473 jich mělo smlouvu na jeden rok a 1695 smlouvu na dva roky. Zákazníků s nejkratším typem smlouvu je tedy více než polovina, nicméně 99,94 % ze všech odchozích byli právě tito zákazníci.

Na grafu v obrázku 40 jsou odcházející zákazníci rozděleni podle toho, jakým způsobem platí za služby. Naprostá většina těch, co odešli, využívala elektronický šek. Celkem to bylo 1160 ze všech zákazníků, co odešli. 141 z odchozích využívalo trvalého příkazu, 109 platilo kartou a 148 poštovním šekem.



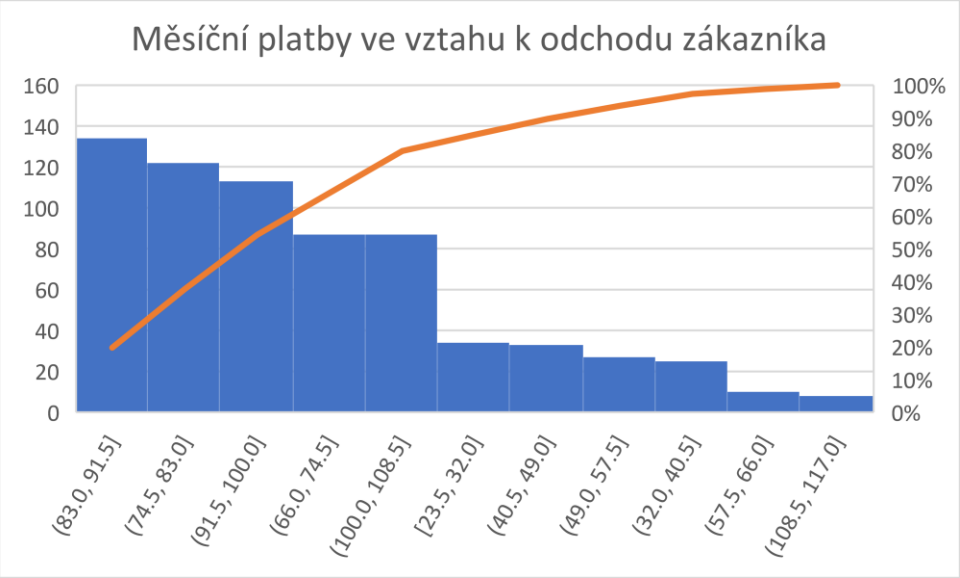
Obrázek 40 Graf rozdělující odcházející zákazníky dle způsobu platby. Zdroj: IBM, vlastní zpracování.

Nyní bude pozornost věnována proměnným Délka kontraktu, Měsíční platby a Celkové platby. Jsou to proměnné, které jejichž hodnoty jsou číselné.



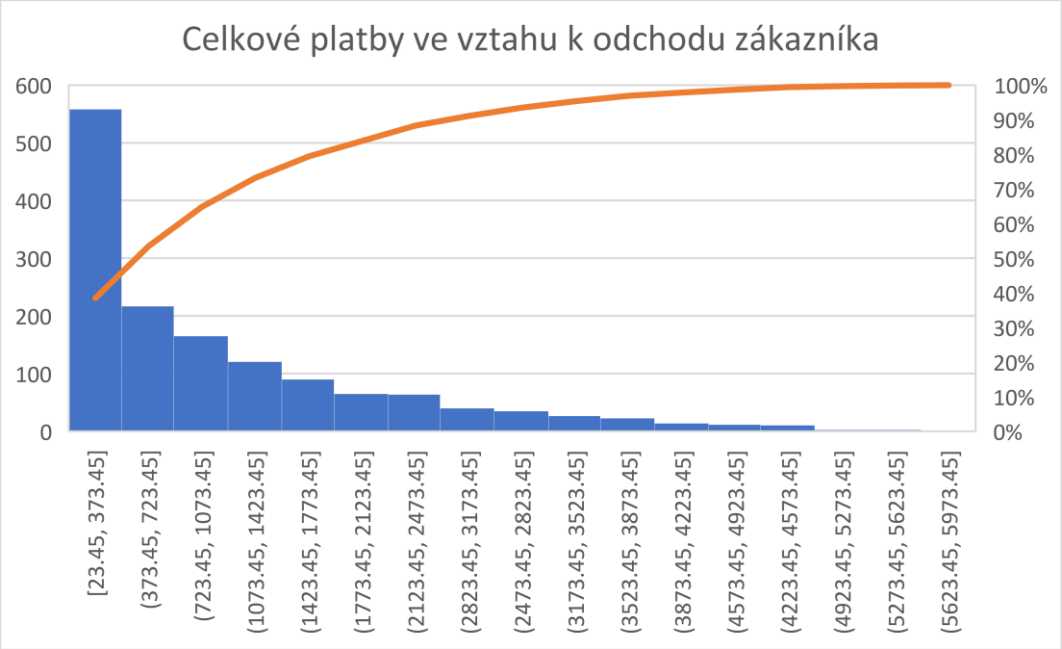
Obrázek 41 Graf zobrazující, kolik zákazníků odešlo v daném měsíci smlouvy. Zdroj: IBM, vlastní zpracování.

Z obrázku 41, kde na ose x jsou měsíce strávené u společnosti a na ose y počty odchozích zákazníků, vyplývá, že nejvíce zákazníků odchází v prvních několika měsících smlouvy. Nejvíce, a to 303 (skoro 20 %), zákazníků, kteří odešli, bylo u poskytovatele jen jeden měsíc. 129 zákazníků bylo u poskytovatele jen dva měsíce a 106 tři měsíce. Z celkového počtu 1558 odcházejících zákazníků jich více než 34,5 % odešlo během prvních tří měsíců strávených u daného poskytovatele. V průměru strávili u tohoto poskytovatele odcházející zákazníci 11,1 měsíce. Naproti tomu ti zákazníci, kteří u poskytovatele zůstali, u něj v průměru jsou 38,4 měsíce čili více než tři roky.



Obrázek 42 Graf měsíčních plateb ve vztahu k odchodu zákazníka. Zdroj: IBM, vlastní zpracování.

V Paretově grafu na obrázku 42 je znázorněn vztah mezi výší měsíčních plateb a odchodem zákazníků. Na ose x jsou intervaly rozdělující zákazníky podle výše jejich plateb a na ose y jsou pak počty zákazníků, kteří odešli. Nejvíce zákazníků, kteří odešli, platí v rozmezí od 83 do 91,50 USD. Vysoké počty odchozích zákazníků se dále nacházely v rozmezí plateb 74,50 až 83 USD a také v rozmezí 91,50 až 100 USD. Lze tedy konstatovat, že největší procento odcházejících zákazníků platilo měsíčně v rozmezí 74,50 až 100 USD. Průměrně pak odcházející zákazník měsíčně platí 80,20 amerických dolarů. Naopak ti, kteří u společnosti zůstali, průměrně zaplatili 60,40 USD. V Paretově grafu na obrázku 43 je pak vidět vztah mezi tím, kolik zákazník za dobu trvání kontraktu již společnosti zaplatil, a mezi jeho odchodem od společnosti. Na ose x jsou opět intervaly celkových plateb a na ose y počty zákazníků, kteří odešli a náleží do daného intervalu. Nejvíce zákazníků zaplatilo společnosti mezi 23,45 a 373,45 USD. Čím se zvyšuje suma zaplacených peněz, tím se snižuje počet odcházejících zákazníků. Více jak polovina všech, co odešli, zaplatila v rozmezí 24,45 a 723,45 USD. Průměrně pak odcházející zákazník zaplatil 981,10 dolarů, zatímco ti, co zůstali, společnosti za celou dobu využívání služeb zaplatili v průměru 2653,90 dolarů.



Obrázek 43 Graf celkových plateb ve vztahu k odchodu zákazníka. Zdroj: IBM, vlastní zpracování.

4.7 Charakteristika odcházejícího zákazníka

Z výše získaných informací lze sestrojit průměrného zákazníka, u kterého je vysoká pravděpodobnost odchodu. Samozřejmě že nejlepší metodou stále zůstává využití modelu logistické regrese, nicméně je vhodné mít i modelového odcházejícího zákazníka, aby společnost byla schopná rychle zákazníka kategorizovat. Do charakteristiky odcházejícího zákazníka budou zařazeny pouze ty proměnné, které na ni mají průkazný vliv. Průměrný zákazník, který od společnosti odejde, vypadá tedy takto:

- má připojení k internetu zařízeno přes optický kabel,
- není důchodce,
- nemá partnera,
- nemá na sobě závislé členy rodiny,
- nevyužívá online bezpečnosti,
- nevyužívá online zálohy,
- nevyužívá ochrany zařízení,
- nevyužívá technické podpory,
- využívá online vyúčtování,

- má smlouvu z měsíce na měsíc,
- platí pomocí elektronického šeku,
- u společnosti je do tří měsíců,
- měsíčně platí v rozmezí 74,50 až 100 amerických dolarů a
- celkově zaplatil do 373,45 USD.

5 Výsledky a diskuse

Po sestavení jednotlivých modelů a vyhodnocení přesnosti jejich predikce pomocí změření výkonnosti, provedení cross-validation (viz tabulka č. 16) a sestavení ROC křivek (viz obrázek č. 36) vyšla pro daný soubor dat jako nejlepší model logistická regrese. Přesnost predikce logistické regrese je 80,39 %, tudíž tento model by z 10 000 zákazníků správně zařadil 8 039 z nich. Přesnost ostatních modelů byla nižší, nejvíce se logistické regresi, co se týče přesnosti predikce, přibližoval náhodný les s přesností 79,11 %. V logistické regresi mělo na chování zákazníka největší vliv to, pokud měl internet veden přes optický kabel, smlouvu na 2 roky, smlouvu na 1 rok, a také výše celkových plateb a délka smlouvy v měsících. Po aplikování výsledné logistické regrese na původní soubor dat byli zákazníci modelem rozděleni na odcházející a zůstávající a z těchto informací byly následně vyvozeny závěry týkající se odcházejících zákazníků a také charakteristika odcházejícího zákazníka. Ten není důchodcem, nemá partnera, nemá na sobě závislé členy rodiny, má smlouvu z měsíce na měsíc, je u společnosti do tří měsíců, nevyužívá žádných doplňkových služeb, má internet přes optický kabel, platí pomocí elektronického šeku, měsíčně platí od 74,50 do 100 USD a celkem společnosti zaplatil do 373,45 USD.

Na základě výsledků modelování a následné aplikace nejlepšího modelu může společnost poskytující telekomunikační služby přijmout jistá opatření, která jim ve výsledku mohou pomoci k udržení zákazníků, což je v tomto oboru velmi důležité. Opatření může společnost přijmout jak na základě samotných výsledků, tak i na základě případných dalších analýz plynoucích z těchto výsledků. Velmi vhodným nástrojem na prohloubení porozumění zákazníkova chování je dotazníkové šetření, které společnost musí provádět na odchozích zákaznících a pokusit se tak hlouběji porozumět důvodům jejich odchodu. Co se týče demografických charakteristik odcházejících zákazníků, s nimi společnost nic nenadělá a nezmění je; mohou tedy sloužit pouze jako takové ukazatele a varovné signály, které mohou

upozornit na vyšší pravděpodobnost zákazníkova odchodu. Díky těmto znalostem pak mohou lidé z telekomunikační společnosti těmto zákazníkům věnovat větší pozornost a nabídnout jim například zvýhodněné služby.

Z výsledků vyplynulo, že valná většina těch, co od společnosti odcházejí, využívají připojení k internetu přes optický kabel. Na základě této informace pak společnost musí dále zkoumat, proč tomu tak je. Je nutné zjistit hlavní příčinu odchodu těchto zákazníků, protože momentálně je optický kabel tím nejlepším způsobem připojení k internetu. Z toho vyplývá, že za lepší technologií zákazníci neodcházejí. Odchod těchto zákazníků může být dán lepšími cenovými nabídkami konkurence či poskytováním lepších doplňkových služeb. Pokud je tím hlavním důvodem cena, musí společnost změnit svoji cenovou politiku a zlevnit, případně nabízet balíček služeb, který překoná nabídku konkurence.

Dále z výsledků plyne, že většina odcházejících zákazníků nevyužívá žádné doplňkové služby, jako je online záloha, online bezpečnost, ochrana zařízení či technická podpora. Společnost tedy musí zvážit, čím je toto způsobeno. Důvodů může být hned několik, jako například nedostatečná kvalita těchto služeb, jejich vysoká cena, nebo nedostatečná informovanost zákazníků, kteří tak vyhledávají tyto služby a přecházejí ke konkurenci. Zaměření se na tyto služby a důvod, proč je odcházející zákazníci nevyužívají, by mohlo být pro společnost velmi přínosné. Nejen že by zákazníci využívající tyto služby mohli zvýšit příjmy, ale také by je to mohlo se společností více svázat, lépe naplňovat jejich předpoklady a celkově zvýšit jejich spokojenost. To by ve výsledku mohlo vést k jejich setrvání u společnosti.

Vzhledem k tomu, že většina odcházejících zákazníků využívá online vyúčtování, společnost musí věnovat pozornost i této oblasti. Opět zde může hrát roli nedostatečná kvalita této služby. Zákazníci by chtěli tuto službu využívat, nicméně její kvalita nedosahuje požadované úrovně, a klasické vyúčtování zasílané poštou je pro tyto zákazníky nepřijatelné. Proto zákazník odejde ke konkurenci, která nabízí tuto službu v lepší kvalitě. Je tedy nutné zajistit dostatečnou přístupnost a použitelnost této služby, zajistit plynulý chod online vyúčtování ve všech typech prohlížečů a také mít kvalitní mobilní aplikaci, která by měla být v dnešní době samozřejmostí. S online vyúčtováním také úzce souvisí metoda platby. Z analýzy vyplynulo, že velké procento odcházejících zákazníků platí pomocí elektronického šeku. Tato informace může posloužit specialistům zabývajícím se touto problematikou nejen k identifikaci zákazníků s vysokou pravděpodobností odchodu, ale také

k hlubšímu prozkoumání tohoto faktu. Jednou z možností řešení je lidem, kteří využívají online vyúčtování, ale platí elektronickým šekem, nabídnout možnost platit trvalým příkazem nebo kartou právě za využití online služeb umožňujících práci s účtem daného zákazníka.

Dále bylo zjištěno, že největší procento odcházejících zákazníků je u společnosti relativně krátkou dobu a má smlouvu z měsíce na měsíc. Jednou z možností, která je ale relativně drastická, je úplně zrušit možnost měsíčních kontraktů a nabízet místo nich alespoň půlroční smlouvy. Během této doby je vysoká pravděpodobnost, že by si zákazník na společnost zvykl, oblíbil si ji a zůstal tedy déle, což je i podloženo výsledky analýzy. Další možností je ponechat možnost měsíčního kontraktu, nicméně jej velmi znevýhodnit oproti ročnímu či dvouletému. Na základě těchto pobídek by se pak dalo očekávat, že by více lidí volilo delší kontrakty a u společnosti zůstávali déle. Jinou možností by také bylo nabízet zákazníkům s měsíčními smlouvami výhodné podmínky při prodloužení smlouvy a donutit tak zákazníky k delšímu setrvání.

Co se týče výše měsíčních a celkových plateb, ty také mohou sloužit pouze jako ukazatele k označování zákazníků s vyšší pravděpodobností odchodu, neboť se odvíjejí od mnoha dalších faktorů, jako je délka smlouvy či využívané služby.

6 Závěr

Cílem práce bylo na základě prediktivního modelování určit faktory, které mají největší vliv na chování zákazníků a následně navrhnout opatření. Jako faktory, které měly největší vliv na zákazníkovo chování, byly zjištěny tyto:

- internet přes optický kabel,
- smlouva na 2 roky,
- smlouva na 1 rok,
- výše celkových plateb a
- délka smlouvy v měsících.

Těchto pět faktorů nejvíce ovlivňovalo to, jestli zákazník u společnosti setrval, či odešel. Následnou aplikací modelu bylo zjištěno, že zákazníci využívající internet přes optický kabel více inklinovali k odchodu, zatímco zákazníci se smlouvou uzavřenou na jeden nebo dva roky zůstávali ke společnosti loajální. Dále bylo zjištěno, že nejvíce

odcházejících zákazníků zaplatilo do 373,45 USD a byli u společnosti relativně krátce, a to do tří měsíců. Opatření, která firma musí přijmout, je několik. Společnost určitě musí hlouběji prozkoumat, proč zákazníci odcházejí, a to například provedením dotazníkového šetření mezi odcházejícími zákazníky. Dále společnost musí zanalyzovat svoje nabízené služby ve vztahu ke konkurenci. Je nutné, aby společnost zjistila, proč ti, co od ní odcházejí, nevyužívají žádné doplňkové služby. Jedním z možných důvodů může být nedostatečná kvalita či příliš vysoká cena. Všechny tyto skutečnosti musí společnost zanalyzovat, aby její snahy mohly vést ke snížení počtu odcházejících zákazníků.

Dílčím cílem práce bylo sestavení čtyř prediktivních modelů a následný výběr toho nejlepšího. Postupně tak byly sestaveny logistická regrese, rozhodovací strom, náhodný les a neuronová síť. U každého modelu byla změřena přesnost klasifikace, a to pomocí cross-validation a následným sestrojením ROC křivek. ROC křivky všech modelů pak byly vloženy do jednoho grafu pro lepší názornost a vybrán nejlepší model, kterým se s přesností klasifikace 80,39 % stala logistická regrese. Tento model pak byl využit na další práci, kdy byl model aplikován na původní soubor a na základě této aplikace určeny faktory popsány v předchozím odstavci. Z výsledků vyhodnocování modelů také vyplynulo, že není možné určit univerzálně nejpřesnější model z těchto čtyř testovaných, a to z důvodu různorodosti dat používaných v jednotlivých výzkumech.

Posledním cílem práce bylo zhotovení charakteristiky odcházejícího zákazníka. Této charakteristiky bylo dosaženo aplikováním modelu na data a následnou analýzou. Ze souboru byli vybráni ti zákazníci, které model určil jako odcházející, a poté byly určeny průměrné charakteristiky těchto zákazníků. Průměrný odcházející zákazník tedy není důchodcem, nemá partnera, nemá na sobě závislé členy rodiny, má smlouvu z měsíce na měsíc, je u společnosti do tří měsíců, nevyužívá žádných doplňkových služeb, má internet přes optický kabel, platí pomocí elektronického šeku, měsíčně platí od 74,50 do 100 USD a celkem společnosti zaplatil do 373,45 USD.

Výsledkem práce je logistická regrese, která jakožto nejpřesnější model na daný soubor dat poskytla charakteristiku odcházejícího zákazníka a také faktory, které jeho chování nejvíce ovlivňují. Z těchto výsledků dále vyplynula doporučení, která společnost musí zvážit, pokud chce snížit počty odcházejících zákazníků. Model logistické regrese pak je společnosti schopen u nově příchozích zákazníků označovat ty, kterým, aby u společnosti zůstali, je nutno věnovat větší pozornost. Využití tohoto modelu má potenciál udržet

zákazníky déle, což je u telekomunikačních společností v současné době velkým problémem. Budoucí výzkum by se měl zabývat prohlubováním znalostí o chování zákazníka; v tomto případě hlavně na hlubší porozumění toho, proč většina zákazníků, kteří odcházejí, má internet veden přes optický kabel, a proč jich většina odchází v prvních měsících strávených u společnosti. Nemusí se jednat o výzkum pouze v akademické sféře, ale pro danou společnost by bylo nejlepší se pokusit objasnit chování jejích zákazníků interně. Tento krok by vedl k navýšení konkurenční výhody oproti dalším hráčům v odvětví. Na základě výsledků těchto analýz bude společnost opravdu schopná reagovat a přizpůsobit se potřebám a chtíčům svých zákazníků. Pro budoucí výzkum by také bylo vhodné zvýšit počet zákazníků, jejichž data jsou použita pro modelování, a to především k zajištění přesnějšího popisu skutečnosti.

7 Seznam použitých zdrojů

Big Data on AWS, *Amazon Web Services* [online]. [cit. 2017-09-05]. Dostupné z: https://aws.amazon.com/big-data/

DE MAURO,Andrea, Marco GRECO, Michele GRIMALDI. *A formal definition of Big Data based on its essential features*, 2016. Library Review [online]. 65(3), 122-135 [cit. 2017-11-03]. DOI: 10.1108/LR-06-2015-0061. ISSN 0024-2535. Dostupné z: http://www.emeraldinsight.com/doi/10.1108/LR-06-2015-0061

DIXON, Wilfrid J.,Frank J. MASSEY a JR. Introduction to statistical analysis. 2nd edition. New York (N. Y.): McGraw-Hill, 1957. ISBN 9780070170704.

EXECUTIVE OFFICE OF THE PRESIDENT, 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights* [online]. 2016 [cit. 2017-10-15]. Dostupné z: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

FREUND Rudolph, William J WILSON a Ping SA. *Regression Analysis*. 2nd ed. Burlington: Elsevier, 2006. ISBN 9780080522975.

GARTNER, 2011. *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*. Gartner.com [online]. Stamford, USA [cit. 2017-08-26]. Dostupné z: https://www.gartner.com/newsroom/id/1731916

GERASIMOU, Vicke, 2016. Big Data and the 3Vs: What is the fourth 'V' and what are the implications for not embracing it?. *Think Big Analytics* [online]. 29.3.2016 [cit. 2017-10-19]. Dostupné z: https://www.thinkbiganalytics.com/2016/03/29/big-data-3vs-fourth-v-implications-not-embracing/

GORDON, J.Edited by BCS a The Chartered Institute for IT. *Big Data Opportunities and challenges*. Swindon: BCS Learning & Development Limited, 2014. ISBN 9781780172620.

HASSOUNA, Mohammed, Ali TARHINI, Tariq ELYAS a Mohammad Saeed ABOU TRAB, 2015. Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research* [online]. **8**(6), - [cit. 2018-01-22]. DOI: 10.5539/ibr.v8n6p224. ISSN 1913-9012. Dostupné z: http://www.ccsenet.org/journal/index.php/ibr/article/view/47593

HENDL, Jan. *Přehled statistických metod: analýza a metaanalýza dat*. Páté, rozšířené vydání. Praha: Portál, 2015. ISBN 9788026209812.

CHEN, Hsinchun, Roger H. L. CHIANG a Veda C. STOREY, 2012. BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. *MIS Quarterly* [online]. **36**(4), 1165-1188 [cit. 2017-10-29]. Dostupné z: https://ai.arizona.edu/sites/ai/files/MIS611D/chen-bi-december-2012.pdf

KARDES Frank R., Thomas W. CLINE a Maria L. CRONLEY. *Consumer behavior*. 2nd edition. 2014. ISBN 9781133587675.

KERLNER, Dan, 2014. Empirical Rule. *Wikimedia Commons* [online]. [cit. 2017-10-10]. Dostupné z: https://commons.wikimedia.org/w/index.php?curid=36506025

KOTLER, Philip a KEVIN LANE KELLER. *Marketing management*. 14th [ed.]. Upper Saddle River, N.J: Prentice Hall, 2012. ISBN 9780132102926.

LEHMANN, E.L. a Joseph P. ROMANO. *Testing statistical hypotheses*. 3rd ed. New York: Springer, 2005. ISBN 9780387276052.

LOGAN, T.M., S. MCLEOD a S. GUIKEMA, 2016. Predictive models in horticulture: A case study with Royal Gala apples. *Scientia Horticulturae* [online]. (209), 201-213 [cit. 2018-01-20]. Dostupné z: https://www.sciencedirect.com/science/article/pii/S030442381630317X

MANYIKA, James, Michael CHUI, Brad BROWN, Jacques BUGHIN, Richard DOBBS, Charles ROXBURGH a Angela HUNG BYERS, 2011. *Big data: The next frontier for innovation, competition, and productivity* [online]. McKinsey Global Institute [cit. 2018-03-04]. Dostupné z: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx

MARR, Bernard, 2015. *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance* [online]. 1. John Wiley [cit. 2017-10-09]. ISBN 978-1-118-96578-8.

MARR, Bernard, 2017. 3 Massive Big Data Problems Everyone Should Know About. *Forbes* [online]. [cit. 2017-8-12]. Dostupné z: https://www.forbes.com/sites/bernardmarr/2017/06/15/3-massive-big-data-problems-everyone-should-know-about/#4eede8cc6186.

MCAFEE, Andrew a Eric BRYNJOLFSSON, 2012. *Big Data: The Management Revolution*. Hardvard Business Review [online]. 2012(Říjen) [cit. 2017-09-18]. DOI: Harvard Business Review. Dostupné z: https://hbr.org/2012/10/big-data-the-management-revolution

MONTGOMERY, Douglas C. *Introduction to linear regression analysis*. 5. ed. Oxford: Wiley-Blackwell, 2011. ISBN 9780470542811.

NIELSEN, Michael, 2017. Using neural nets to recognize handwritten digits. *Neural Networks and Deep Learning* [online]. [cit. 2017-12-27]. Dostupné z: http://neuralnetworksanddeeplearning.com/chap1.html

NISBET, Robert. *Handbook of statistical analysis and data mining applications*. Amsterdam: Elsevier, c2009. ISBN 978-0-12-374765-5.

PENN STATE. The 7 Step Process of Statistical Hypothesis Testing, *Penn State University* [online]. Penn State University [cit. 2017-11-11]. Dostupné z: https://onlinecourses.science.psu.edu/stat502/node/139

ŘEHÁKOVÁ, Blanka, 2000. Nebojte se logistické regrese. *Sociologický časopis* [online]. **36**(4), 475-492 [cit. 2018-02-28]. Dostupné z: http://sreview.soc.cas.cz/cs/issue/64-sociologicky-casopis-4-2000/1149

SAS. Decision Trees for Business Intelligence and Data Mining, 2008. In: *SAS Support* [online]. SAS [cit. 2017-12-11]. Dostupné z: https://support.sas.com/publishing/pubcat/chaps/57587.pdf

SAS. What is BIG DATA?, *SAS* [online]. [cit. 2017-12-07]. Dostupné z: https://www.sas.com/en_th/insights/big-data/what-is-big-data/.

SIMON, Phil. *Too big to ignore: the business case for big data*. Hoboken, New Jersey: John Wiley & Sons, 2013. ISBN 9781118641866.

STROME, Trevor L. *Healthcare Analytics for Quality and Performance Improvement.* Hoboken: Wiley, 2013. ISBN 9781118760178.

The Four V's of Big Data, 2014. IBM Big Data Hub [online]. [cit. 2018-03-03]. Dostupné z: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

VYSEKALOVÁ, Jitka. *Chování zákazníka: jak odkrýt tajemství "černé skříňky"*. Praha: Grada, 2011. Expert (Grada). ISBN 978-80-247-3528-3.

WIKIPEDIA. Cross-industry standard process for data mining, *Wikipedia* [online]. [cit. 2018-02-04]. Dostupné z: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

ZAIONTZ, Charles, 2013. Null and Alternative Hypothesis. *Real Statistics* [online]. [cit. 2017-12-04]. Dostupné z: http://www.real-statistics.com/hypothesis-testing/null-hypothesis/

Zdroj dat:

Using Customer Behavior Data to Improve Customer Retention, 2015. *IBM.com* [online]. [cit. 2018-03-13]. Dostupné z: https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/

8 Přílohy

Příloha č. 1 – výsledek logistické regrese.

Proměnná	Koeficient	Směrodatná odchylka	p-hodnota
Délka kontraktu	-0.0585	0.0061	0.0000
Smlouva na 2 roky	-1.3962	0.1758	0.0000
Smlouva na 1 rok	-0.6685	0.1074	0.0000
Online vyúčtování-Ne	-0.3416	0.0745	0.0000
Celkové platby	0.0003	0.0001	0.0000
Platební metoda-karta (automaticky)	-0.3924	0.0973	0.0001
Platební metoda-poštovní šek	-0.3615	0.0963	0.0002
Platební metoda-bankovní převod (automaticky)	-0.3062	0.0945	0.0012
SeniorCitizen.1	0.2147	0.0845	0.0111
Internet-optický kabel	1.7530	0.7976	0.0280
Internet-ne	-1.8135	0.9636	0.0598
Online filmy-Ano	0.6047	0.3264	0.0639
Online TV-Ano	0.5912	0.3262	0.0699
Konstanta	1.8057	-1.0159	0.0755
Závislí-Ano	-0.1553	0.0897	0.0833
Měsíční platby	-0.0402	0.0317	0.2054
Online bezpečnost-Ano	-0.2054	0.1786	0.2501
Technická podpora-Ano	-0.1784	0.1804	0.3229
Ochrana zařízení-Ano	0.1480	0.1763	0.4013

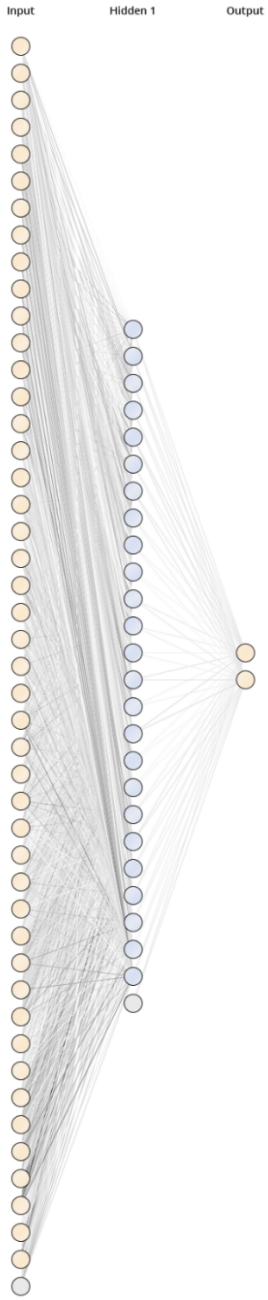
Více linek-Ano	0.6277	0.8051	0.4356
Pohlaví-muž	-0.0219	0.0648	0.7356
Více linek-Ne	0.1810	0.6479	0.7800
Online záloha-Ne	-0.0261	0.1752	0.8815
Partner-Ne	0.0032	0.0778	0.9667
Online bezpečnost-nemá internet	0	NaN	NaN
Online záloha-nemá internet	0	NaN	NaN
Ochrana zařízení-nemá internet	0	NaN	NaN
Technická podpora-nemá internet	0	NaN	NaN
Online TV-nemá internet	0	NaN	NaN
Online filmy-nemá internet	0	NaN	NaN
Telefon-Ano	0	NaN	NaN

Tabulka 19 Výsledek logistické regrese. Zdroj: Vlastní zpracování.

Příloha č. 2 – rozhodovací strom

Obrázek 44 Rozhodovací strom. Zdroj: Vlastní zpracování.

Příloha č. 3 – Neuronová síť



Obrázek 45 Neuronová síť. Zdroj: Vlastní zpracování.