Anotace

Diplomová práce s názvem Data mining a jeho využití v ekonomické praxi popisuje možnosti užití data miningových nástrojů primárně v odvětví financí a ekonomie. Úvod práce je zaměřen na pojmy související s disciplínou data miningu. Zde je definován pojem data minimg, jeho techniky a historie, pojem business intelligence a Big Data jakožto obory, které s data miningem úzce souvisí. Následující část práce se soustředí na data miningový model zaměřený na konkrétní finanční instituci. Závěr práce je věnován poznatkům zjištěným tvorbou modelu a jejich následným vyhodnocením.

Klíčová slova

Data mining, business intelligence, finance, modelování, Big Data, ekonomické pojetí, IBM SPSS Modeler, metodologie, efektivita práce, řešení problémů.

Annotation

A thesis entitled Data mining and its use in economic practice describes the possibilities of using data mining instruments primarily in the finance and economics sectors. The introduction of the thesis is focused on concepts related to the discipline of data mining. Here is defined the term minimg data, its techniques and history, the concept of business intelligence and Big Data as fields that are closely related to data mining. The following part of the work focuses on a data mining model focused on a particular financial institution. The conclusion of the work is devoted to the findings of model formation and their subsequent evaluation.

Key words

Data mining, business intelligence, finance, modelling, Big Data, economic concept, IBM SPSS Modeler, methodology, work efficiency, problem solving.

Obsah

Seznam obrázků

Seznam tabulek

Seznam grafů

Seznam zkratek

BI	Business Intelligence
ČAP	Česká asociace pojišťoven
DM	Data mining
DMA	Data Mart, datové tržiště
DWH	Data Warehouse, datový sklad
EAI	Enterprise Application Integration, integrační nástroje
ETL	Extraction Transformation Loading, transformační nástroje
IBM	International Business Machines Corporation
ICT	Informační a komunikační technologie
ML	Machine learning
OLAP	On-Line Analytical Processing
UI	Umělá inteligence

Úvod

Diplomová práce na téma Data mining a jeho využití v ekonomické praxi si klade za cíl vytvoření data miningového klasifikačního modelu pro určování odchodů klientů z banky XY.

Před samotným sestavováním modelu je však nutné pochopit teoretické souvislosti z odvětví data miningu. Data mining, neboli *„dolování z dat"* je oborem pracujícím s velkoobjemovými daty, proto je na počátku diplomové práce definován pojem Big data. Kapitola zaměřující se na Big data se zabývá jejich zpracováním, vývojem a příklady využití v praxi.

Dalším pojmem pro lepší zařazení data miningu je Business Inteligence, jakožto souhrn přístupů pro práci s daty, mezi které patří i data mining. Business Inteligence obsahuje několik vrstev, přičemž každá vrstva obsahuje jiné nástroje pro práci s daty. Mezi vrstvami se nachází vrstva pro nahrávání a transformaci dat. Data se nachází na různých úložištích a v různých formátech, které je potřeba sloučit a v rámci vrstvy pro ukládání dat uložit do určitého typu datového úložiště, například datového skladu či datového jezera. Uložení dat na jednom místě umožní firmě s daty pracovat efektivněji než v případě rozložení informací mezi více systémů.

Pokud jsou data na jednom místě, je jednodušší jejich analýza (vrstva analýzy dat). V této části jsou popisovány analytické činnosti prováděné v této vrstvě. Patří sem například reporting, či právě data mining. Po analýze dat je možné se přesunout k prezentační vrstvě, ve které jsou data vizualizována tak, aby mohla být poutavější formou předána například vedení společnosti. Prezentační vrstva může obsahovat aplikace a další systémy, které umožní propojení koncových uživatelů s výstupy z ostatních vrstev BI. Poslední vrstvou, která je v práci zmiňována je vrstva know-how.

Poté je již práce zaměřena čistě na téma data miningu, jakožto součásti Business intelligence. V rámci tématu data mining je definován samotný pojem, představena metodologie CRISP-DM a techniky data miningu. Dále jsou v této podkapitole stručně popsány podkategorie data miningu, kterými je text mining, web mining a zpracování obrazu.

Teoretické vymezení data miningu je v kapitole dvě doplněno o příklady využití ve vybraných odvětvích ekonomie a financí. Konkrétně jsou zde uvedeny příklady z odvětví marketingu, bankovnictví a pojišťovnictví.

Druhá část diplomové práce je zaměřena na samotný popis vytvořeného data miningového modelu zaměřeného na odchody klientů z banky XY. Při jeho sestavování je postupováno podle metodologie CRISP-DM. Posledním bodem práce je vyhodnocení výsledků vygenerovaných sestaveným modelem.

1 Vymezení základních pojmů z oblasti

1.1 Big Data

Již ze samotného pojmu vyplívá, že se jedná o soubor mnoha komplexních údajů, které přichází ve velkém množství. Konkrétně se jedná o velké datové sklady přicházející velkou rychlostí primárně z nových zdrojů, proto je nedokáže zpracovat běžný software pro zpracování dat. (Hendl 2021)

Big Data jsou definována třemi R (v angličtině třemi V):

- **Rozsah** (Volume).
  Rozsah dat se měří v rámci terabytů, petabytů, exabytů atd. Zpracovává se velké množství převážně nestrukturalizovaných dat. Data mohou obsahovat neznámé hodnoty získané například z mobilní aplikace, kliknutím na webovou stránku, z aktivity na sociálních sítích atd. (Anon. [b.r.])

  V dnešní době jsme informacemi doslova přehlceni, velké množství generovaných dat tlačí na vývoj výkonnějších software, zároveň se s objemem zvyšují náklady na datová úložiště. Objem dat roste také v závislosti na množství připojených zařízení k internetu, rozsah je tedy hlavním aspektem Big dat. Pro zajímavost lze uvést historický podklad, kdy bylo od vzniku civilizace do roku 2003 vytvořeno 5 exabytů informací, v současné době je toto množství vytvořeno za dva dny (Hendl 2021). Přičemž jeden exabyte je 1 000 petabytů, což je 10^{18} megabytů. (Anon. [b.r.])

- **Rychlost** (Velocity).
  Při zvyšujícím se objemu dat je potřeba data rychle přeměňovat. Data se rychle ukládají a zpracovávají. (Anon. [b.r.])
- **Různorodost** (Variety).
  Objevuje se mnoho formátů dat. Záleží na to, z jakého zdroje se data načítají. Může jít o text, webovou stránku, obrázek či například zvuk. Analyzují se strukturalizovaná i nestrukturalizovaná data. (Anon. [b.r.])

Během vývoje přístupu k datům se v posledních letech vytvořily další dvě definice dat: **hodnota** a **věrohodnost**. Nemůžeme říci, že by všechny datové sady měly stejnou hodnotu. Některá data jsou cennější než jiná. Interní data společností zobrazující transakce, platební morálku či jiné informace z profilu klienta nám mohou připadat cennější než například záznamy nálad uživatelů Facebooku. Cennost informací je občas těžké posoudit, nicméně je to

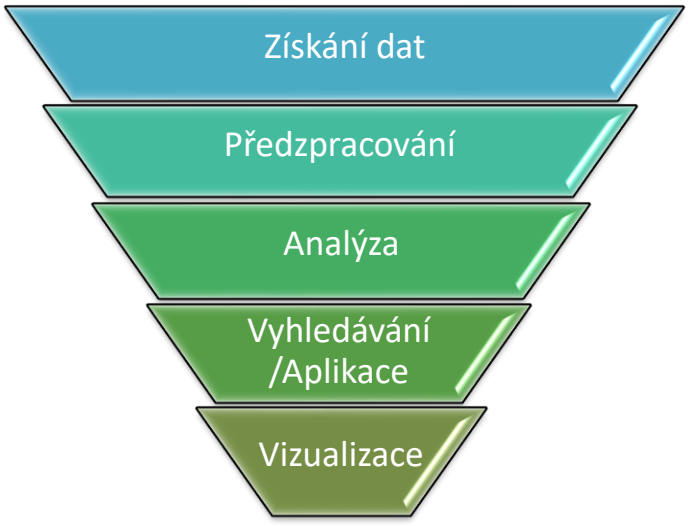
jeden z parametrů, který můžeme u jednotlivých sad určit. Věrohodnost dat není méně důležitým parametrem. Pokud využijeme nevěrohodná data, může se stát, že námi interpretované výsledky budou milné. Například u při zakládání profilů uživatelů či vyplňování dotazníků nemusíme mít jistotu pravdivých odpovědí, které mohou zkreslit výsledek, v případě, že jich je mnoho, či nemáme dostatek vstupních dat. (Anon. [b.r.])

1.1.1 Zpracování Big dat

Zpracování Big dat probíhá v několika krocích:

- Získání dat.
- Předzpracování.
- Analýza.
- Vyhledávání/Aplikace.
- Vizualizace.

V prvním kroku se data získávají, případně hledají, poté se předzpracují a následně na to analyzují. Po analýze dat přichází jejich aplikace a následná vizualizace viz obrázek číslo 1. V rámci předzpracování dat dochází k odstranění duplikátů a čištění nulových hodnot. Co se týče oblasti analýzy, uživatelé z různých odvětví využívají různé druhy analýzy. Zmíněné analytické techniky zahrnují například statistické modely, strojové učení, data mining. Vizualizace je výsledkem předchozích kroků, je to cesta, jak sdělení předat uživateli. Uživatelem je často oddělení firmy, vedoucí pracovník, či vedoucí útvar. (Hendl 2021)



Obrázek č. 1: *Proces zpracování Big dat*

Zdroj: Vlastní zpracování.

1.1.2 Big Data v praxi

Big Data je možné využít v široké škále odvětví, například v marketingu při analyzování reakcí uživatelů například na sociálních sítích, internetových vyhledávačích a dalších. Dále je možné využít velké množství dat při řešení obchodních problémů, které by bez takto rozsáhlých informací nebylo možné vyřešit. V otázce dat totiž platí, čím větší objem dat v určité kvalitě máme, tím přesnější analýzy a predikce lze tvořit. (Hendl 2021)

Dalším příkladem využití Big dat je například vývoj produktů na základě zákaznické poptávky. Společnost Netflix a Procter & Gamble vytvářejí pro nové produkty klasifikační modely založené na hodnocení klíčových vlastností produktů s přihlédnutím k jejich úspěšností na trhu. Po vyhodnocení se poté zaměřují na produkty, které mají vlastnosti vyžadované trhem. Na tomto základě poté mohou obsah personalizovat pro jednotlivé uživatele. (Anon. [b.r.])

Dobrým příkladem personalizace obsahu je společnost Amazon, ve které vznikl první matematický model doporučující uživatelům obsah, o který budou mít pravděpodobně zájem. Amazon byl na svém počátku e-shop nabízející knihy online. Jádro tohoto e-shopu tvořilo několik desítek knižních kritiků a editorů, kteří tvořili obsah webu Amazonu. Editoři měli na starosti titulky a vzhled e-shopu a knižní kritici doporučovali a komentovali obsah inzerovaných knih, což zákazníkům pomáhalo při rozhodování o koupi. Mayer-Schönberger a Cukier (2014, s. 59) říkají, že editoři a kritici *„odpovídali za to, čemu se říkalo „styl Amazonu" a co společnost považovala za jedno ze svých klíčových aktivit a svou konkurenční výhodu. Článek v tehdejším vydání Wall Street Journal je označil za nejvlivnější literární kritiky v celé zemi, protože se na základě jejich článků prodávalo tolik knih."*

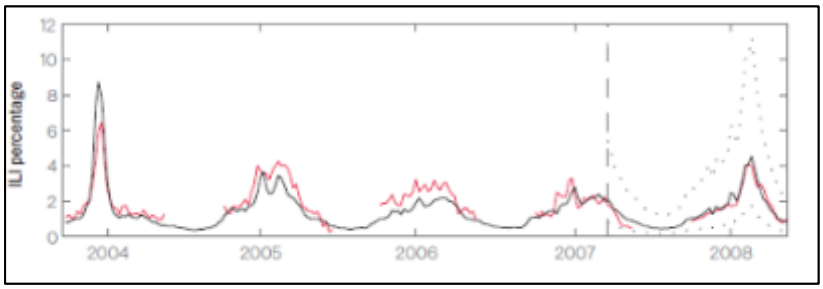
Poté se však společnost začala zajímat o možnosti automatického doporučování obsahu zákazníkům. Amazon po celou dobu své existence zaznamenával data o pohybu zákazníků na e-shopu. Zaznamenával jak prodeje, tak také vložení knih do košíku, obsah košíku, kliknutí a prohlížení knihy (zda vedlo, či nevedlo ke koupi), délku prohlížení dané knihy atd. První verze modelu nebyla příliš zdařilá, neboť docházelo k doporučování velmi podobného obsahu, byly zde zvoleny příliš malé odchylky. Zákazník si nechtěl koupit knihu o Kanadě, když vyhledával knihu o Novém Zélandu. Tento model porovnával zákazníky a jejich košíky a na základě toho doporučoval knihy. Nakonec však Amazon našel řešení, které nesrovnávalo zákazníky s potenciálními zákazníky, ale pouze hledalo souvislosti mezi jednotlivými knihami. (Mayer-Schönberger a Cukier 2014)

Tuto metodu bylo možné využít také na jiné produkty, takže ve chvíli, kdy Amazon začal nabízet další kategorie produktů, mohl model nasadit také pro zbylý sortiment. Tuto metodu poté převzala například společnost Netflix pro personalizaci filmů pro diváky.

Co se oblasti Big dat týče, je důležité zmínit společnost Google, která jako světový vyhledávač disponuje obrovským množstvím zaznamenaných dat. Google má v oblasti tvorby predikčních modelů i dalších produktů Big dat ohromnou sílu, kterou využil například roku 2009, kdy se v USA rozšířil virus kombinující prvky ptačí a prasečí chřipky. Tuto událost využila společnost Google ke shromáždění dat z vyhledávačů a následné predikci onemocněních chřipkového typu. (Hendl 2021)

Google nejprve pouze mapoval vyhledávané termíny vždy v lokalitách, ve kterých se sezónní nemoc objevovala. Po dostatečném množství analyzovaných dat našla společnost několik slovních spojení typu *„Co využít na kašel?“*, která vyhledávali nakažení lidé. Na základě těchto termínů, zjištěných z obrovského množství dat, nasadila společnost matematický model, který dokázal určit kde se nyní nemoc nachází. Díky implementovanému modelu mohla společnost Google informovat populaci o tom kde se nákaza aktuálně nachází, a to dříve než agentura spadající pod Ministerstvo zdravotnictví a sociální péče, které pro získání dat potřebuje fyzickou přítomnost každého jedince, kvůli následnému otestování. Tento model společnost využila i při výskytech jiného onemocnění chřipkového typu. (Hendl 2021)

Výsledný produkt Googlu je možné vyhledat pod názvem Google Flu Trends. Na grafu číslo 1 je znázorněno, jak odhady Google Flu Trends téměř kopírují výsledky zaznamenávané agenturou spadající pod Ministerstvo zdravotnictví a sociální péče, Centers for Disease Control and Prevention. Na ose Y jsou uvedena procenta, na ose X sledované roky. Černě je znázorněna křivka reálně zjištěných případů od agentury a červeně je křivka odhadu společnosti Google, korelace mezi nimi je 0,96. Ve druhé části grafu je tečkovaná křivka vystihující predikci společnosti Google. (Hendl 2021)



Graf č. 1: Data Google Flu Trends a Centers for Disease Control and Prevention
Zdroj: HENDL, J, [b.r.]. BIG DATA VE ZDRAVOTNICTVÍ – PERSPEKTIVY PROBLEMATIKY. 13.

Google Flu Trends určitě nebyl jediný produkt z dílny Google, při kterém byla využita velkoobjemová data. Okolo roku 2000, kdy Microsoft zdokonaloval kontrolu gramatiky programu Word, rozhodla se společnost Google posunout laťku ještě o kousek výš, a to vytvořením automatického překladače. Pokusy o tvorbu automatického překladače proběhly

již dříve, a to během studené války, kdy počítač sopečnosti IBM přeložil, první věty z ruštiny do angličtiny. Počítač přeložil, podle tiskové zprávy IBM, šedesát vět úspěšně. Po následném vývoji a testování však bylo zjištěno, že program neumí dosazovat vícevýznamová slova do správného kontextu věty. Společnost IBM se pokusila namísto přesné definice gramatických pravidel zapojit spíše výběr nejvhodnějšího slova na základě vypočítané pravděpodobnosti. Toto řešení sklidilo úspěch. Další inovace však úspěšnost překladače posouvaly stále pomaleji, proto se společnost rozhodla projekt ukončit. (Mayer-Schönberger a Cukier 2014)

Roku 2006 se o vývoj překladače pokusila společnost Google, která využila své přednosti- velkého množství dat získávaných z vyhledávačů. Společnost tedy využila velký, neuspořádaný datový soubor, a tím byl celý světový internet. Společnost tvořila překladač tím způsobem, že hledala podobná spojení slov v oficiálních dokumentech, firemních zprávách, na firemních webech či ve skenech knih. Tímto způsobem společnost vytvořila funkční překladač, který je již schopný překládat také na základě hlasového vstupu. Překladač stále není bezchybný, nicméně je úspěšnější než překladač společnosti IBM. Za úspěchem společnosti Google stojí zhruba deset tisíckrát větší datová sada, než kterou měla k dispozici společnost IBM, a to přes to, že společnost IBM měla data lepší kvality. (Mayer-Schönberger a Cukier 2014)

Další možností ochrany, ve které mohou data pomoci je prevence proti podvodům či porušování předpisů, například v rámci praní špinavých peněz. Pomocí Big dat je možné identifikovat vzorce chování, které se liší od normálu, díky tomu je možné rychle reagovat a předejít škodám. Často zmiňované využití je též ve zdravotnictví, kde mohou shromážděná data o pacientech doporučit nejvhodnější léčbu pro další pacienty. (Anon. [b.r.])

1.1.3 Big Data a jejich vývoj

První záznamy o zpracování dat pochází podle (Mayer-Schönberger a Cukier 2014) přibližně z roku 5 000 před naším letopočtem. Tato data byla zaznamenávána na hliněných destičkách sumerských obchodníků, kteří si takto zaznamenávali údaje o zboží. Počátky ve zpracování Big dat byly však doménou většího celku a to státu. Stát či vládnoucí vrstva se již před naším letopočtem snažili o to, nějakým způsobem vyčíslit velikost obyvatelstva a jejich majetku. Jako první využití Big dat je tedy možné uvést sčítání lidu, ke kterému docházelo ve velkých městech Egypta, Anglie, Číny či například Itálie.

V té době bylo sčítání lidu velmi nákladné. Každým sčítáním byli pověřeni sčítací komisaři, kteří obcházeli domácnosti a fyzicky sčítali osoby i majetek. Tato metoda byla velmi zdlouhavá a pouze přibližná. Například v Římě trvalo sčítání lidu dlouhých 8 let a s rychle rostoucím počtem obyvatel byla data zastaralá ještě před tím, než bylo sčítání zcela dokončeno. Tato problematika dala impuls pro vznik oboru statistiky. (Mayer-Schönberger a Cukier 2014)

Jednalo se o okruh statistiky, který se zabýval sledováním údajů o obyvatelstvu v delších časových úsecích. Tento okruh se tvořil v 17. století v Anglii a jeho hlavními představiteli byli britský obchodník John Graunt a William Petty. (Kříž et al. [b.r.])

Počet zpracovávaných dat rostl s rostoucím obyvatelstvem a jejich majetky. Najednou byly nástroje na zpracování dat doslova přehlceny. Impulzem na to byl vznik tabelačních strojů na děrné štítky. Děrné štítky byly vyplněny respondenty a poté vkládány do strojů. Vynalezení stroje znamenalo počátek strojového zpracování dat a zároveň položilo základy společnosti se současným názvem IBM. (Anon. [b.r.])

Tato metoda zkrátila čas sčítání lidu z osmi let na jeden rok, byla však tak nákladná, že ji nebylo efektivní využívat častěji než jednou za deset let. Proto se pozornost obrátila znovu ke statistice a jejímu výběrovému souboru, přičemž statistice zjistili, že pokud nejsou lidé vybráni náhodně, dosahují výsledky vyšší přesnosti. Zjistili, že nehledě na velikost populace, náhodně vybraný vzorek poměrně spolehlivě reprezentuje celek. Zjištění, že pomocí náhodného výběrového souboru je možné zjišťovat lepší výsledky, než při zohlednění celku byl klíčový. Tato metoda byla finančně nenáročná, byla přesnější a nezahlcovala systémy tak velkým množstvím dat, proto bylo možné provádět průzkumy několikrát do roka. Princip náhodného vzorku se zavedl také v dalších odvětví, ulehčil například kontrolu kvality a zavádění změn v podnicích, dal vzniknout předvolebním průzkumům či průzkumy spokojenosti zákazníků. (Mayer- Schönberger a Cukier 2014)

U vzniku náhodného výběru však příběh nekončí. Náhodný výběr byl pouze přechodným řešením problému, který nastal se zpracováním veškerých dat, celku. Náhodný vzorek má své využití v praxi, není však jednoduché ho stanovit a pokud dojde při výběru k omylu, promítne se tato skutečnost jako zkreslení výsledných dat. K chybě může dojít opomenutím určitého faktoru. Například při internetových průzkumech hrozí, že nebude oslovena starší generace lidí, kteří se na internetu příliš nepohybují. (Mayer-Schönberger a Cukier 2014)

Během sčítání lidu v 19. století lidstvo nebylo po technické stránce připraveno na zpracování velkoobjemových dat, proto si pomohlo úspornějším a v té době nejlepším řešením. Přelom pro zpracování Big dat přišel ve 20. století s internetem, který umožnil sdílet a zaznamenávat informace vyšší rychlostí. Následovalo spuštění prohlížeče společnosti Google, která stála za dalšími inovacemi, též z oblasti Machine learning a objemnými analýzami dat. Dále roku 1998 vývoj databázového systému NoSQL. 21. století přineslo například roku 2005 platformu Apache Hadoop, na které je možno ukládat strukturovaná i nestrukturovaná data, proto ji lze využít například u ML či pokročilých analýz. Roku 2006 došlo k dalšímu posunu v uchovávání dat, a to vznikem prvního cloudového úložiště od Amazonu. (Anon. [b.r.])

1.2 Business Intelligence

Business Inteligence (BI) je souhrn přístupů a aplikaci pro práci s daty. Využívá se při řízení a analýze dat ve společnosti. S využitím BI nástrojů získává společnost možnost efektivní práce s daty, která má dopad na správnost strategických rozhodnutí a na udržení kroku s konkurencí. Business Intelligence obsahuje několik vrstev, které jsou blíže popsány v kapitolách 1.2.1, 1.2.2, 1.2.3, 1.2.4 a 1.2.5. (Hendl 2021)

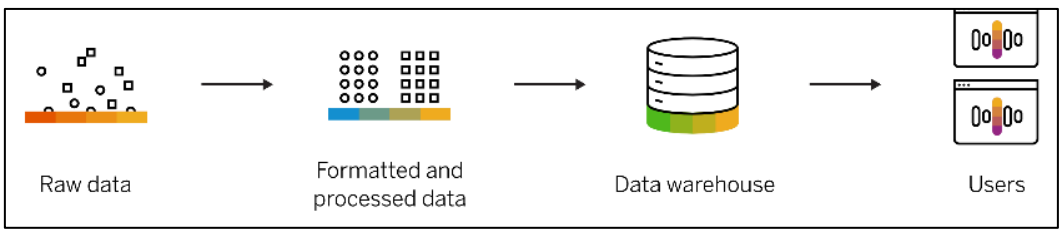
Jinými slovy se Business intelligence soustředí na shromažďování dat prostřednictvím databází, jejich následnou analýzou a vizualizací pro firemní potřeby a strategická rozhodnutí. BI tedy transformuje a zpracovává strukturovaná i nestrukturovaná data, tak aby byla pro management či jiné útvary společnosti čitelnější a pochopitelnější, k tomu využívá vizualizaci. (Anon. [b.r.])

Student Harvardu McCosh tvrdí, že pojem Business Intelligence zazněl poprvé z úst Scotta Mortona roku 1964 při obhajobě jeho disertační práce. Za propagací tohoto termínu však stojí Howard Dresner datový analytik z firmy Gatner Group, který termín začal propagovat v období okolo roku 1989. Howard Dresner vnímá BI jako *„soubor konceptů a metod ke zlepšení obchodního rozhodování pomocí podpůrných systémů.“* (Anon. [b.r.])

Koncem 20. století začal vznikat nový technologický trend datových skladů (Data Warehouse) a datových tržišť (Data Mart) sloužících k ukládání velkoobjemových dat. Za vývojem a propagací těchto systémů stojí Ralph Kimball a Bill Inmon. Vznik datových skladů a tržišť vytvořil dobré podmínky pro vznik dalších oborů, jako dolování z dat (Data mining) či strojového učení. (Anon. [b.r.])

1.2.1 Vrstva pro nahrávání a transformaci dat

Vrstva nahrávání a transformace dat je první vrstvou Business Intelligence. V této vrstvě je potřeba surová data nejprve získat z různých zdrojů, tedy extrahovat a poté ukládat například v datových skladech. (Novotný et al. 2005) Tento proces je znázorněn na obrázku číslo dva. Na počátku jsou surová data, která je potřeba upravit do určité formy, aby mohla být uložena do datového skladu, poté jsou dále používána.



Obrázek č. 2: *Transformace surových dat*

Zdroj: Anon., [b.r.]. Co je to datový sklad? | Definice, komponenty, architektura | SAP Insights. SAP [online] [vid. 2022c-12-23]. Dostupné z: https://www.sap.com/cz/insights/what-is-a-data-warehouse.html

V rámci této vrstvy se využívají dva systémy pro zpracování dat:

- **ETL systémy**. Jedná se o zkratky slov extrakce, transformace, load (přenos). V rámci extrakce jsou získávána data z různých zdrojů, poté probíhá transformace, která obsahuje činnosti typu čištění dat, agregace, odstraňování duplicit, filtrování atd. Finální fází je přenos dat do datového skladu. ETL je systém, který získává data ze systémů sloužících pro provoz společnosti. Data tedy sbírá například ze systému evidujícího zásoby na skladě, z účetních, ekonomických či výrobních systémů. (Schiller [b.r.])

  Každý z těchto systémů je jiný, něčím charakteristický, proto také data jsou často v různorodých formátech. Touto fází prochází společnost, pokud chce začít využívat BI nástrojů, je tedy potřeba vytvořit datový sklad jako jednotné místo pro všechny datové informace. Tato fáze sběru a transformace dat je velmi náročná a nákladná, podle webu o informačních technologiích (Schiller [b.r.]) může tato fáze stát společnost až 70 % z celkových nákladů na budování systému. Budování tohoto systému není radno podcenit. Pokud se po nasazení do provozu objeví nedbalostní chyby, zpomalí to běžné fungování společnosti a pracovníků obsluhujících datový sklad či data marty. (Schiller [b.r.])
- **EAI systémy**. V překladu Enterprise Application Integration, neboli integrace podnikových aplikací, jedná se tedy o integrační nástroje. EAI systémy tedy propojují a **přenášejí data** z různých zdrojů do datových úložišť, a to **v reálném čase**. Tímto způsobem společnost dojde ke zjednodušení a automatizaci podnikových procesů. Příkladem propojení může být integrace aplikace pro řízení dopravy, řízení vztahů se zákazníky (CRM) s business intelligence. Díky přenosu v reálném čase daly EAI systémy impuls pro vznik takzvaných Real-Time Data Warehouse (Anon. [b.r.])

1.2.2 Vrstva pro ukládání dat

Tato vrstva zajišťuje ukládání dat a jejich shromažďování a správu na jednotných místech kterými jsou:

- **Datové sklady (Data Warehouse).** Datový sklad, označující se často zkratkou DWH či DW je analytickou databází často relačního typu. V DWH dochází ke kumulaci historických dat vytvořených uvnitř podniku, která mohou mít měřítko v petabajtech. Z DWH jsou poté generovány sestavy (relace), které jsou následně využívány dalšími nástroji Business Intelligence. Dříve bylo nasazení datového skladu pro společnost

výhodou, nyní je to pro větší a často i střední podniky nezbytností pro udržení si dlouhodobé pozice na trhu. Datový sklad slouží ke spojování informací z vícero zdrojů. Každý systém či aplikace ve společnosti shromažďuje určité typy informací, například o klientech, o zaměstnancích, účetní záznamy, záznamy z controllingu a jiné. Všechny tyto informace je možné uchovávat pomocí datových skladů na jednom místě a dále s nimi pracovat. (Kroenke a Auer 2015)

Podle jednoho ze zakladatelů datových skladů Billa Inmona viz publikace (Novotný et al. 2005) je datový sklad *„integrovaný, subjektivně orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu.“*
Integrovaný ve smyslu ukládání dat z vícero zdrojů napříč celou společností. **Subjektivně orientovaný**, protože se soustředí na položky v datech, ne na dělení dle jednotlivých aplikací, ze kterých data přichází. Každá konkrétní položka, například informace o zákazníkovi by měla být v databázi datového skladu uložena pouze jednou, oproti produkčnímu systému, kde jsou data rozptýlena podle aplikací.
Stálým skladem se rozumí nemožnost ručního přepisování a vkládání informací do skladu. Informace jsou do skladu načítána z interních či externích zdrojů společnosti. Datový sklad musí být také **časově rozlišený**, neboť je potřeba v rámci databází pracovat s čísli za určitá období a analyzovat historické a současné hodnoty. Je tedy potřeba aby data obsahovala informaci o dimenzi času.

Alternativou k datovým skladům může být datové jezero (Data Lake). Datové jezero však může fungovat také v kooperaci s datovým skladem, a to jako zdroj dat pro datový sklad. Datové jezero se od DWH liší tím, že je v něm ukládáno velké množství dat v nestrukturovaného typu, tedy různých formátů. Data není potřeba pro uložení v datovém jezeře transformovat, ani definovat jejich strukturu. Formát se u datových jezer mění až takzvaně „při čtení“, což znamená, že až určitá aplikace bude načítat data z jezera, bude na něj až v této chvíli aplikovat vlastní vztahy. (Anon. [b.r.])

Tabulka 1: Datový sklad vs datové jezero

	Datové jezero	Datový sklad
Typ	Strukturované, částečně strukturované, nestrukturované	Strukturované
	Relační, nerelační	Relační
Schéma	Schéma při čtení	Schéma při zápisu
Formát	Nezpracováno, nefiltrováno	Zpracované, prověřené
Zdroje	Velké objemy dat, IoT, sociální média, streamování dat	Aplikace, obchodní data, transakční data, dávkové generování sestav
Škálovatelnost	Snadné škálování při nízkých nákladech	Obtížné a nákladné škálování
Uživatelé	Datoví vědci, datoví inženýři	Specialisté na datové sklady, obchodní analytici
Případy použití	Strojové učení, prediktivní analýza, analýza v reálném čase	Základní generování sestav, BI

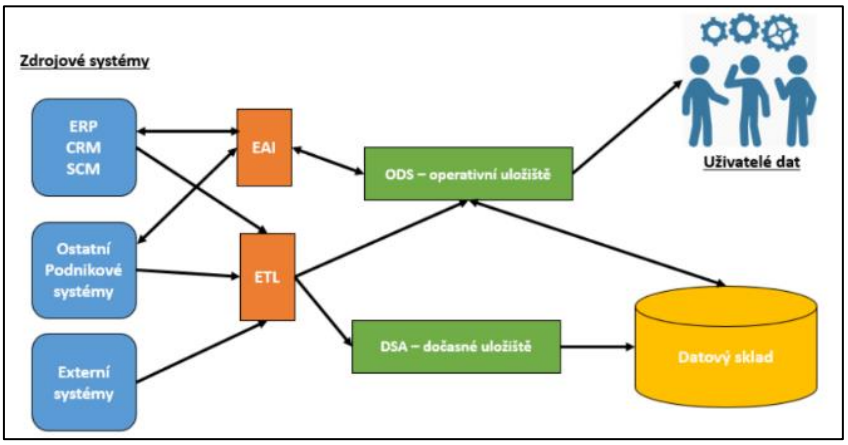
Zdroj: Anon., [b.r.]. Co je datové jezero? | Microsoft Azure [online] [vid. 2022c-12-26]. Dostupné z: https://azure.microsoft.com/cs-cz/resources/cloud-computing-dictionary/what-is-a-data-lake/

- **Datová tržiště (Data Marts)** mohou být závislé na datového skladu, nebo mohou samy o sobě tvořit datový sklad. Datová tržiště tedy mohou sloužit při tvorbě datového skladu k jeho lepšímu seskupení. Mohou však sloužit i po vytvoření datového skladu, jako určité mezičlánky, se kterými pracují jednotlivá odvětví společnosti. Pracovníci společnosti tedy nemusí hledat ve velkém množství databází uvnitř datového skladu, ale mohou využít své datové tržiště, kde není tak velké množství databází, které v daném odvětví ani nepotřebují. (Novotný et al. 2005)

- **Dočasná datová úložiště (Data Staging Area)** nejsou povinnou součástí Business Intelligence v řešení firmy. Dočasná úložiště však ocení společnosti, které mají zatěžovaný produkční systém velkým náporem extrahovaných dat nebo je u extrahovaných dat potřeba změnit formát před tím, než je bude možno nahrát do datového skladu (například u systémů generujících textové dokumenty). Dočasné úložiště tedy slouží jako první úložiště dat generovaných ze systémů a aplikací společnosti, před tím, než jsou nahrány do datového skladu či řešena nějakým BI řešením. Data z dočasného úložiště jsou zpracována, převedena a poté z dočasného úložiště odstraněna. Zapojení dočasného datového úložiště do struktury je zobrazeno na obrázku tři. (Novotný et al. 2005)

- **Operativní datové úložiště (Operational Data Source).** Stejně jako u dočasného datového úložiště nemusí být operativní úložiště dat součástí Business Intelligence uvnitř každé společnosti. Operativní datová úložiště lze definovat dvěma způsoby. První způsob pracuje s daty v reálném čase, a to přímo ze zdrojových systémů napojených přímo k Enterprise Application Integration (EAI). Druhý způsob pracuje s daty z datového skladu. V tomto případě tedy pracuje pouze s aktuálními snímky, ne

s celou historií, stejně jako dočasná datová úložiště. Obrázek 3 zobrazuje zapojení operativního úložiště. (Novotný et al. 2005)



Obrázek č. 3: *Operativní a dočasné úložiště dat*

Zdroj: GONZALES, Michael L., 2003. IBM data warehousing: with IBM business intelligence tools. New York: Wiley. ISBN 978-0-471-13305-6.

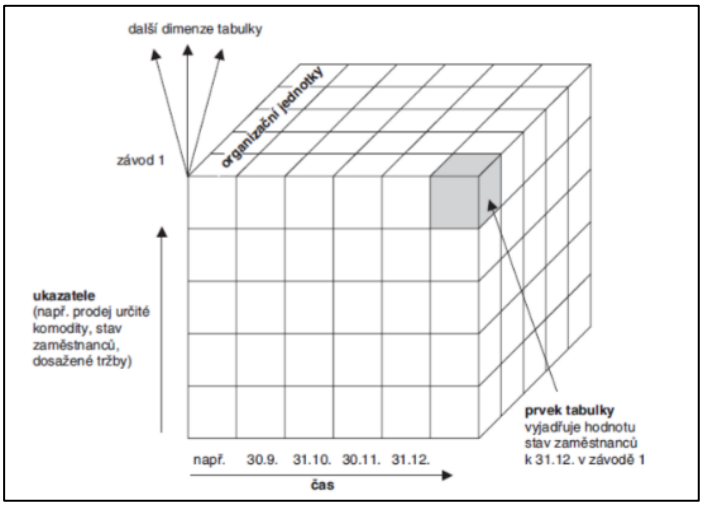
1.2.3 Vrstva pro analýzu dat

V této vrstvě Business Intelligence jsou již data přístupná pro jejich analýzu. Tato vrstva obsahuje tři základní pojmy: reporting dat, dolování z dat (Data mining) a systémy On-Line Analytical Processing (OLAP).

- **Reporting.** Datový reporting je proces dotazování se do jednotlivých databází datového skladu či data martu pomocí jazyka SQL (u relačních databází). Pokud je potřeba pracovat s nerelační databází, je možné využít například jazyka Python. Existují dva základní druhy datového reportingu, a to ad hoc reporting a standardní reporting. Ad hoc reporting se zabývá požadavky které vznikly nyní a je potřeba je rychle řešit. Na databáze jsou tedy formulovány jednorázové dotazy, které se často neopakují, jsou specifické. Standardní reporting se provádí v určitých časových intervalech. Na databáze jsou v tomto případě směřovány dotazy, které jsou například každé čtvrtletí stejné. Management může např. každé čtvrtletí potřebovat znát platební morálku největších klientů společnosti. Výstupy reportingu jsou tabulky, grafy a další vizualizace. (Novotný et al. 2005)

- **OLAP** neboli On-Line Analytical Processing je software, který umožňuje vytvářet online analýzy z datového skladu či tržiště rychle a ve velkém množství. OLAP je možné si představit jako kostku (viz obrázek 4), která obsahuje několik dimenzí (kategorií) a tvoří jádro systémů OLAP. Když jsou sledována například data o prodeji výrobku jsou s prodeji sledovány také dimenze, kterými mohou být datum prodeje, země/město

prodeje, kategorie výrobku a další. OLAP tedy dokáže zpracovávat více dimenzí zároveň rychleji a efektivněji než datové sklady. V datovém skladu jsou ukládány jednotlivé tabulky (databáze) a je zde možné organizovat data pouze do dvou dimenzí najednou. Kostka OLAP pracuje s jednou tabulkou, kterou rozšiřuje o další vrstvy, další dimenze. Při uvedení příkladu na prodeji výrobku lze tento prodej organizovat do jednotlivých vrstev kostky následovně, první vrstva může obsahovat dny/časy a roky prodejů, další vrstva jednotlivé země či města prodejů, další vrstva kategorie prodaných výrobků a tak dále. Další kostky je možné tvořit také v jednotlivých vrstvách. Toto vrstvení je možné také v datových skladech, má však dopad na výkon skladu. (Anon. [b.r.])



Obrázek č. 4: Multidimenzionální databáze

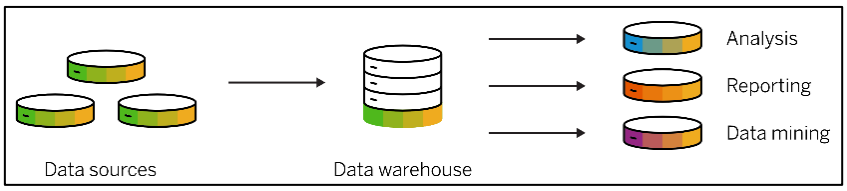
NOVOTNÝ, Ota, Jan POUR a David SLÁNSKÝ, 2005. *Business intelligence: jak využít bohatství ve vašich datech*. 1. vyd. Praha: Grada Publishing. ISBN 978-80-247-1094-5.

Vícerozměrný OLAP neboli **MOLAP** (Multidimensional OLAP) je OLAP pracující přímo s vícerozměrnou kostkou OLAP. To znamená, že se data ukládají přímo v kostkách OLAP. Dalším druhem kostky je **ROLAP** (Relational OLAP), který ukládá data v relačních databázích. ROLAP tedy pracuje napřímo s daty v relačních databázích, což zhoršuje výkon datového skladu. Následujícím typem je **HOLAP** (Hybrid OLAP), který kombinuje předchozí zmíněné OLAP. Kombinuje relační databáze s více rozměrnými v rámci jednoho OLAP systému. Umožňuje tvořit datové kostky skrz které je možné se dostat k relačním databázím což je flexibilní přístup pro zpracování dat. Relační tabulky však HOLAP zpomalují, též komplexnost HOLAP vyžaduje údržbu a časté aktualizace. (Anon. [b.r.]) Posledním typem je **DOLAP** (Descop OLAP), který je

ze zmíněných typů nejmladším. V rámci DOLAP je možné stáhnout určitou podmnožinu kostky a tu uložit na lokálním disku. (Novotný et al. 2005)

- **Data mining** neboli dolování z dat umožňuje získávat informace z velkoobjemových dat pomocí speciálních algoritmů. Více o data miningu v kapitole 1.3. (Novotný et al. 2005)

Obrázek 5 znázorňuje počáteční datové zdroje, které jsou transformovány do datového skladu a z datového skladu jsou data čerpána pro následné analýzy a další výstupy BI.



Obrázek č. 5: Vrstva pro analýzu dat

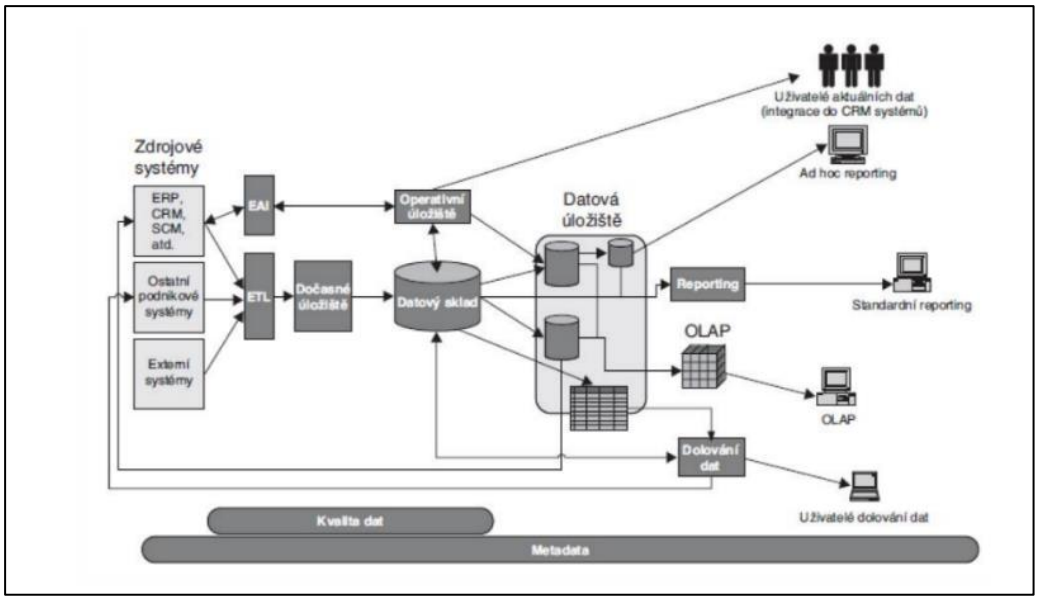
Anon., [b.r.]. Co je to datový sklad? | Definice, komponenty, architektura | SAP Insights. *SAP* [online] [vid. 2022e-12-23]. Dostupné z: https://www.sap.com/cz/insights/what-is-a-data-warehouse.html

1.2.4 Prezentační vrstva

Prezentační vrstva obsahuje nástroje pro koncové uživatele. Jedná se o různé webové aplikace, systémy či jiné analytické aplikace které propojí uživatele s dalšími vrstvami Business Intelligence a pomohou jim tak, dostat se k informacím, které potřebují. (Novotný et al. 2005)

1.2.5 Vrstva oborové znalosti (know-how)

Know-how je nedílnou součástí pro fungování Business Intelligence ve společnosti. V praxi se většinou jedná o znalost fungování prostředí, do kterého se BI implementuje. Toto know-how zajišťují zaměstnanci společnosti, která má o nasazení Business Intelligence zájem, tito pracovníci znají podnikové procesy a systémy které se ve firmě využívají. Druhou znalostí je technologická, kterou má firma dodávající a implementující BI technologii. Na obrázku 6 je možné vidět hlavní nástroje Business Intelligence, které byly v rámci jednotlivých vrstev rozebrány a jejich vzájemné vazby.



Obrázek č. 6: Hlavní nástroje BI a jejich vazby

Zdroj: NOVOTNÝ, Ota, Jan POUR a David SLÁNSKÝ, 2005. *Business intelligence: jak využít bohatství ve vašich datech*. 1. vyd. Praha: Grada Publishing. ISBN 978-80-247-1094-5.

1.3 Data mining

Pojem Data mining (DM) lze definovat několika způsoby, pro tuto práci byla zvolena definice z konference kybernetiky z roku 2006: *„Data mining je proces identifikace a interpretace vzorců v datech pro řešení konkrétního obchodního problému. Strategie vytěžování dat pro BI zahrnují klasifikaci, odhad, predikci, analýzu časových řad, asociační analýzu nebo analýzu nákupního koše.“* (Anon. [b.r.])

Data mining, neboli dolování dat, vytěžování dat či objevování znalostí z dat (KDD, Knowledge Discovery from Data) lze také definovat například jako extrakci dopředu neznámých informací, velmi často, z rozsáhlých databází. Definic je dostupných opravdu mnoho, základem všech definic je však to, že se jedná o proces, v rámci kterého jsou v datech hledány souvislosti. Metodou data miningu lze řešit různé typy úloh. Z hlediska typu je DM dělen na explorační data mining a prediktivní data mining. (Skalská 2010)

- **Explorační data mining.** Explorační neboli průzkumový data mining řeší úlohy popisného typu. Soustředí se tedy na popsání získaných dat, případně jejich následnou vizualizaci. Analýzy tohoto typu mohou být užity ke zjištění toho jaký statistický model je možné využít k popsání sledovaných veličin. Dále je exploračním data miningem možné hledat odlehlé hodnoty, které mohly vzniknout při samotném sběru dat, nebo se jedná opravdu o odlehlé hodnoty, díky kterým je možné odhalit anomálie v datech. Tímto způsobem lze odhalit neobvyklé chování spotřebitelů, které může být nelegálním počinem či situací o které by měla být daná společnost informována.

Shluková analýza například seskupuje data do shluků podle jejich vlastností, pokud jsou zobrazena data mimo velký shluk, znamená to, že mají odlišné vlastnosti. Podle vzdálenosti od hlavního shluku je možné zjistit v jaké míře jsou vlastnosti vzdálených prvků od těch s běžnými vlastnostmi rozdílné. (Skalská 2010)

Další využití exploračního data miningu může být při hledání příčin chybějících pozorování. Může se stát, že v celkové datové sadě, která je zpracovávaná chybí v některých sloupcích data. Tyto chybějící hodnoty mohou vzniknout například nedůvěrou respondentů v anonymitu dotazníku a mohou tak zkreslit celkový výsledek výzkumu. (Skalská 2010)

- **Prediktivní data mining** se zaměřuje na typy úloh, ve kterých lze na základě historických dat vytvořit model, který předpovídá určitý děj do budoucna (Skalská 2010). Tento typ úlohy je využit pro praktickou ukázku v této práci, kdy je sestaven model, který na základě vložených (historických dat) předpovídá, zda daný klient z banky odejde či nikoli. Informace o odchodech klientů je v historických datech uvedena, těmito daty se tedy model učí, následně po nahrání nových (aktuálních) dat model rozhodne s jakou pravděpodobností klient z banky odejde či nikoli.

1.3.1 Metodologie CRISP DM

Pro data mining existuje více metodologií vytvořených společnostmi zabývajícími se ICT. Tyto metodologie jsou zaměřeny na konkrétní software, nejsou tedy univerzální. Je možné jmenovat například postupy 5A či SEMMA. (Siobos [b.r.])

Nejvíce využívanou metodologií je CRISP-DM neboli Cross-Industry Standard Process for Data mining, která je softwarově univerzálním a obecným procesem či normou pro sestavení data miningového modelu. CRISP-DM je univerzální pro různé typy úloh, softwarů a je volně dostupná pro všechny. Impuls pro vytvoření metodologie CRISP-DM zadala Evropská komise, která poskytla grant na sestavení jednotné data miningové metodologie. Na vytvoření spolupracovalo více společností z různých odvětví, v čele tohoto seskupení stála společnost IBM (International Business Machines Corporation). Tato metodika bude dále rozebrána a je využita též v praktickém příkladu diplomové práce. (Anon. [b.r.])

Metodologie CRISP_DM obsahuje šest částí, podle kterých by se mělo postupovat během tvorby DM modelu (Olson 2018). Části postupu jsou následující.

1. **Porozumění problému.** V této fázi je potřeba porozumět řešenému problému a jeho cílům. V této fázi obvykle přichází požadavek na projekt ze strany managementu, který

zajišťuje konzultace dané problematiky, či zajistí ke konzultacím zaměstnance, který se v dané oblasti pohybuje. Data miningový model totiž může být sestavován na oblasti napříč společností, a to například na oblast logistiky, marketingu, financí či obchodu. Tvůrce modelu potřebuje jak analytické znalosti, tak znalost daného odvětví, na které model sestavuje. Dále se v této fázi stanovuje plán sestavení modelu, a to jakého data může management očekávat funkční implementovaný model. (Anon. [b.r.])

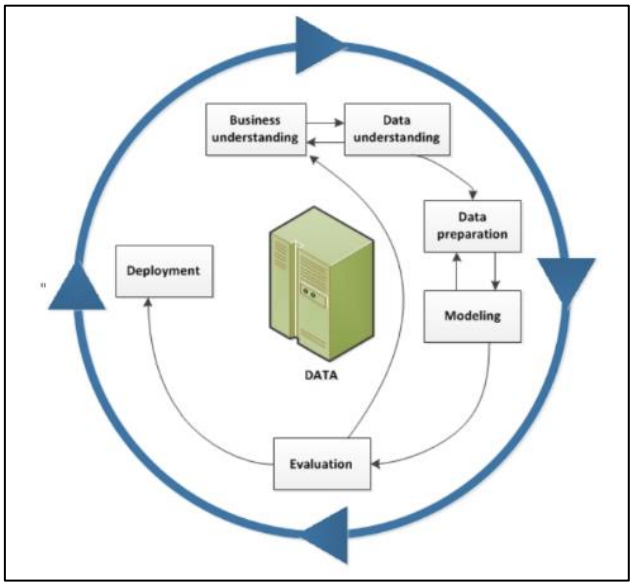
2. **Porozumění datům.** Ve fázi, kdy je jasný problém a cíl je potřeba porozumět datům. V této fázi je s daty realizováno několik kroků. Prvním krokem této fáze je **sběr vstupních dat**. V rámci sběru dat se definují zdroje, ze kterých jsou data získávána, případně problémy, které při získávání dat vznikly a návrhy, jak je řešit. Dalším krokem je **popis dat**. Popisem dat jsou zjišťovány základní informace o datové sadě či sadách, jako typ dat, kolik sada obsahuje záznamů či významy polí. Zjišťování významu polí je důležité i z toho důvodu, že je možné předpokládat, zda je možné realizovat projekt na základě získaných dat. Pokud je předpokládáno, že projekt nebude úspěšný, konzultuje se obsah datové sady se specialistou z daného odvětví. Následujícím krokem je **zkoumání dat** v rámci kterého jsou data vizualizována grafy a tabulkami a popisována. Zjišťovány jsou například počty opakujících se hodnot v rámci jednoho atributu, maxima, minima, či průměrné hodnoty. Posledním krokem v této fázi je **ověření kvality dat**. Pokud je zjišťována kvalita dat, zjišťuje se například, zda data obsahují prázdné hodnoty či zda data obsahují chyby. Pokud se chyby a prázdné hodnoty v datové sadě nacházejí, zjišťuje se jejich četnost a navrhuje se možnost jejich řešení. (Anon. [b.r.])

3. **Příprava dat.** Ve fázi porozumění datům byla data sebrána, popsána, zkoumána a byla ověřena jejich kvalita. Po ukončení této fáze je možné pokračovat samotnou přípravou dat v rámci které jsou data čištěna o případné nulové hodnoty, či jsou nulové hodnoty nahrazovány na základě jiných hodnot. Dále data **integrujeme**, tedy slučujeme stejné atributy z různých databází. Některé atributy mohou mít různé názvy, ale stejné hodnoty, tyto sloupce sloučíme. Poté jsou data **formátována** tak, aby splňovala určitou formu, která je k modelování vyžadována, například pořadí či specifické označení sloupců. Fáze přípravy dat je obvykle časově nejnáročnější, ale pro tvorbu modelu je stěžejní. Po ukončení fáze přípravy dat je k dispozici připravená datová sada k modelování. (Anon. [b.r.])

4. **<u>Modelování</u>.** V rámci fáze modelování jsou vybírány správné techniky řešení daného projektu (jako rozhodovací stromy, neuronové sítě atd.). V souvislosti se zvolenou technikou jsou vybírány také algoritmy, které budou aplikovány a hodnoceny. V této fázi může dojít k potížím způsobeným špatnou přípravou dat, v tomto případě je potřeba se vrátit k předchozímu kroku přípravy dat viz obrázek 7. Pokud problémy nenastanou je v této fázi zjištěn výsledek na základě práce vybraného algoritmu. (Anon. [b.r.])

5. **<u>Evaluace</u>** znamená vyhodnocení výsledků. V této fázi jsou výsledky projektu konzultovány s managementem případně pracovníky z daného odvětví. Po této konzultaci může být zjištěno, že model rozhoduje moc, nebo málo přísně, můžeme ho tedy za určitý typ chyb penalizovat, nebo můžeme pozměnit nastavené podmínky pro rozhodování. (Anon. [b.r.])

6. **<u>Nasazení modelu do praxe</u>.** V této fázi je model hotový, výsledky jsou vyhodnoceny a je na řadě rozhodnutí o tom, jak zjištěné výsledky využít. Výstupem této fáze může být například závěrečná zpráva o výsledcích dosažených modelováním, nebo může být data miningové řešení automatizováno. V rámci automatizace napojeno na potřebné databáze a využíváno pro monitoring. (Anon. [b.r.])



Obrázek č. 7: Schéma metodologie CRISP-DM

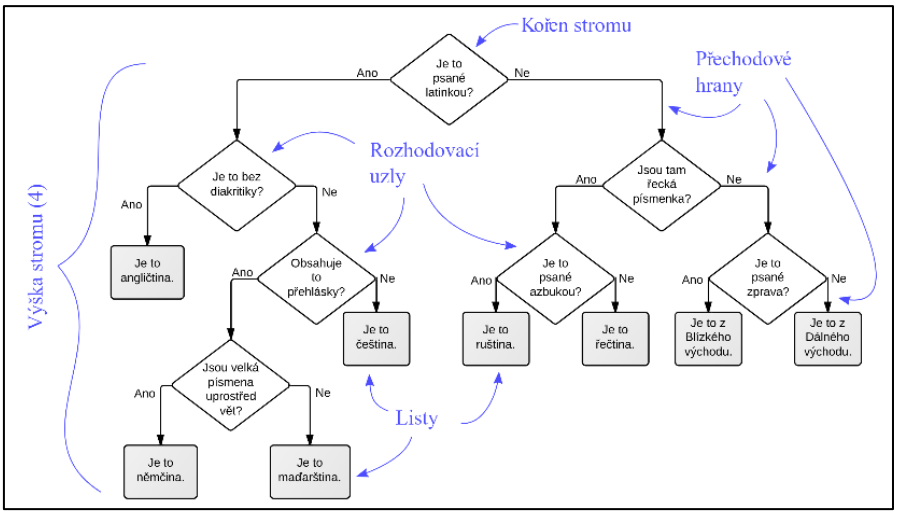
Zdroj: SIOBOS, Aneta, [b.r.]. *Lekce 3 - Data Mining - Metodologie procesu a používané techniky* [online] [vid. 2023-01-04]. Dostupné z: https://www.itnetwork.cz/metodologie-data-mining-procesu-a-pouzivane-techniky

1.3.2 Techniky data miningu

Data mining obsahuje množství technik vycházejících ze statistiky či strojového učení. Pro účely této práce je uveden pouze vzorek vybraných technik. Vybranými technikami jsou rozhodovací stromy, které jsou využity také při praktickém příkladu, dále clustering a klasifikace, neuronové sítě a genetické algoritmy. (Novotný et al. 2005)

- **Rozhodovací stromy.** Název „strom" je odvozen od vzhledu schématu modelu, ve kterém definujeme kořenový uzel, větve a listy viz obrázek 8. Rozhodovací stromy patří mezi nástroje strojového učení a řeší úlohy predikce i klasifikace. **Klasifikace** reprezentuje rozdělení položek do tříd/roztřídění podle podobných vlastností. V rámci schématu stromu je klasifikace dobře viditelná. Funguje tím způsobem, že k jednotlivým atributům (klient, produkt, země, věk, počet produktů) přiřadíme určitou cílovou proměnnou, například zda klient z banky odešel či nikoli, často uvádíme binárně (1 a 0 či ANO a NE). Rozhodovací strom poté začíná od kořenu, kterým může být odchod klienta z banky a postupuje takto od kořene přes větve k listům. Větev je například počet produktů, které klient má v dané instituci a strom zjišťuje, pokud má jeden produkt odešel (1), pokud má dva a více produktů neodešel (0), tudíž z relevantních větví pokračují listy a zkoumá další podmínky. Tímto postupem vyhledá rozhodovací strom atributy, které mají největší podíl na tom, že klient z instituce odešel či nikoli. (Trejbal 2014)

  Takto naučený strom je možné dále využít k predikci. **Predikce**, jak již bylo zmíněno znamená předpovídání určitých skutečností na základě historických dat. Například zda na základě zadaných vlastností klient z instituce odejde či nikoli. Tuto predikci může firma využít ve svůj prospěch například možností využití slev pro klienty u kterých je pravděpodobnost odchodu vysoká, či sestavením marketingové kampaně, která na dané klienty cílí. (Trejbal 2014)



Obrázek č. 8: Rozhodovací strom

Zdroj: Anon., [b.r.]. *Rozhodovací stromy a chytré otázky – Základy informatiky pro střední školy* [online] [vid. 2023j-01-01]. Dostupné z: https://popelka.ms.mff.cuni.cz/~lessner/mw/index.php/U%C4%8Debnice/Informace/Rozhodovac%C3%AD_stromy_a_chytr%C3%A9_ot%C3%A1zky

Rozhodovací stromy se dělí na obecné stromy (nebinární) a binární. **Obecné stromy** se mohou od kořenu větvit libovolným počtem větví z toho vyplývá, že mají běžně méně úrovní. Do obecných stromů je možné řadit algoritmy C5.0 či CHAID. **Binární stromy** lze rozeznat tak, že se větví pouze do dvou větví, na základě toho mají často více úrovní. Toto větvení pouze po dvou větvích zároveň způsobuje, že počítají rychleji. Mezi binární stromy patří například algoritmus C&RT (CART). (Anon. [b.r.])

K rozhodování využívají rozhodovací stromy různé druhy algoritmů, některé z nich byly zmíněny v předchozím odstavci. Pro účely této práce jsou vybrány algoritmy: C&RT (CART), C5.0 a CHAID. Algoritmus **CHAID** (Chi-Squared Automatic Interaction Detector) tvoří obecné stromy, tedy stromy s libovolným počtem větví a patří mezi nejstarší algoritmy. Algoritmus **C5.0** patří též mezi obecné algoritmy. Vytvořený model pomocí C5.0 je poměrně robustní a dokáže dobře řešit problémy vzniklé v případě chybějících dat či velkém počtu řádků. Modely vytvořené s C5.0 bývají poměrně rychlé a zároveň přehledné a srozumitelné. **C&RT** neboli (CART – Classification and Regression Tree) patří mezi algoritmy tvořící binární stromy, tedy stromy pouze se dvěma větvemi. (Anon. 2021a)

Rozhodovací stromy jsou v porovnání s jinými technikám data miningu poměrně dobře pochopitelné, proto je často využívají začínající specialisté v oboru strojového učení a data miningu. Rozhodovací stromy nepotřebují speciální přípravu dat před

sestavením modelu. Výstupem modelu je grafické zobrazení stromu, které lze popsat bez odborné terminologie. Je tedy vhodným modelem pro prezentaci výstupu například pro vedení firmy. Na schématu stromu je dobře viditelné, jak model pracoval a jak přesně docílil výsledku. U ostatních data miningových technik často nelze popsat proces, neboť fungují na základě složitého počítačového programu, do kterého uživatel nemůže vidět (black-boxy). Uživatel zná vstupy a výstupy, ale samotný proces většinou nelze sledovat. Další výhodou rozhodovacích stromů je, že dokáží samy identifikovat a vybrat relevantní entity/proměnné a ty nerelevantní z modelu vyloučit. (Trejbal 2014)

Každý predikční model předpovídá výsledek s určitou pravděpodobností a pokud není pravděpodobnost stoprocentní dochází k určitým chybám. Model může způsobit chyby prvního druhu (False Positive) a chyby druhého druhu (False Negative). Chyba prvního druhu **(False Positive)** znamená, že model vyhodnotil záznam jako pozitivní (binárně 1), ale to je špatně, správná odpověď je 0. Na příkladu klienta, který odejde či neodejde z dané instituce by to znamenalo, že model vyhodnotil – klient odejde, ale klient neodešel. Opakem, který může nastat je chyba druhého druhu **(False Negative)**, kdy model vyhodnotí negativní výsledek (0/NE), ale správně je (1/ANO). U chyby false negative by tedy model vyhodnotil, že klient neodejde, ale on by odešel. (Trejbal 2014)

Tento druh chyby je v daném příkladu ten horší, neboť když model vyhodnotí že klient neodejde a on odejde, nebude možné se na daný typ klienta zaměřit např. v rámci marketingu. Za tento typ chyby je tedy možné model penalizovat. Po nastavení penalizace dokáže model identifikovat jaký typ chyby je horší a v případě, kdy se nebude moci rozhodnout, přikloní se k chybě menší závažnosti. Penalizace je důležitá především při rozhodování týkajících se lidských životů, například při doporučování vhodné medikace pro pacienta či při rozhodování o jedlých a jedovatých houbách. (Trejbal 2014)

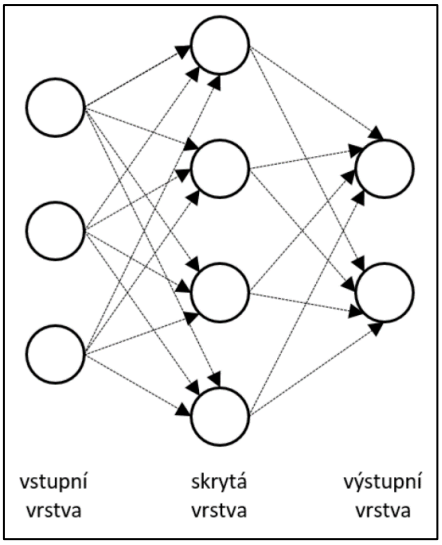
Při tvorbě data miningových modelů je dobré zmínit druhy učení modelu, a to s učitelem a bez učitele. **Učení s učitelem** (Supervised learning) je učení modelu při kterém jsou známá nejen vstupní data, ale i ta správná výstupní data. Na těchto datech se tedy model naučí, nebo si ověří, jak mají být výstupy správně a tento postup může poté aplikovat i na dalších datových sadách. Učení s učitelem je jako učení ve škole. Žáci se naučí například sestavovat rovnice při matematice a ověří si přitom správnost svých výpočtů se správným výsledkem. Takto se naučí počítat rovnice a poté to mohou

aplikovat i na další rovnice. Tímto způsobem se tedy model učí generovat přesnější výsledky. Nevýhodou metody učení je vyšší nákladovost. (Lacko [b.r.])

Učení bez učitele (Unsupervised learning) je situace, kdy jsou k dispozici pouze vstupní data bez výstupních. V tomto případě není známo, zda má úloha řešení, zda jsou mezi atributy souvislosti, či zda jsou v datech shluky. V praxi se často využívá kombinace těchto dvou metod zvaná **polo řízené učení** (Semisupervised learning). V tomto případě je k dispozici část dat se známým výstupem, na které se model natrénuje a poté jsou aplikována veškerá data, která je potřeba analyzovat. (Lacko [b.r.])

- **Neuronové sítě.** Daný název získaly neuronové sítě z důvodu podobnosti fungování modelu s nervovou soustavou v lidském mozku. V lidském mozku i v umělých neuronových sítích figurují neurony. Neurony přijímají signály z vnitřního i vnějšího prostředí, ty poté zpracovávají a následně na ně odpovídají. Vstupové signály mohou být přijímány z jiných neuronů (zevnitř) či zvnějšku, přičemž každý tento vstup má jinou váhu, která se v rámci sítě zohlední. (Anon. [b.r.])

  Úlohu je možné řešit s jedním a více neurony, které jsou vzájemně propojeny. Na obrázku 9 jsou tři vstupy obsahující informace vkládané do modelu, ty jsou spojeny s neurony ve skryté vrstvě (black-box) a výstupem jsou dva neurony. (Anon. [b.r.])

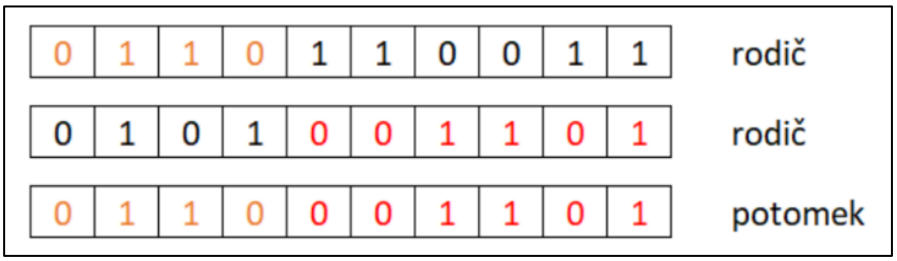


Obrázek č. 9: Neuronové sítě

Zdroj: Anon., [b.r.]. *Neuronové sítě a princip jejich fungování | NaPočítači.cz* [online] [vid. 2023i-01-01]. Dostupné z: https://www.napocitaci.cz/33/neuronove-site-a-princip-jejich-fungovani-uniqueidgOkE4NvrWuNY54vrLeM670eFNQh552VdDDulZX7UDBY/

Neuronové sítě mají široké pole využití, například při hledání poruch na strojích (využívá se v automobilovém průmyslu), k textové analýze či převodu mluvené řeči do textové formy, dále při rozeznávání různých předmětů či osob (hledání SPZ, rozeznávání obličejů, analýza rentgenových snímků, ...). Neuronové sítě se využívají také k automatizaci strojů v rámci například autopilota či automatického řízení strojů. Své uplatnění najde také v marketingu i financích např. při automatickém rozhodování o obchodu na burze či rozhodování o rizikovosti klientů. (Anon. [b.r.])

- **Genetické algoritmy** občas označované také jako **evoluční** jsou algoritmy napodobující princip biologické evoluce. Tento, z pohledu počítačové vědy, nevšední princip algoritmu napodobuje křížení druhů, pomocí kterého dochází k výměně genetických informací. Tyto algoritmy můžou být užity například při řešení logistické úlohy ve které je potřeba dostat nákladní vozy ze skladů do prodejen co nejefektivnější cestou. Plán této přepravy je zakódován do binární podoby viz obrázek 10. V počáteční fázi řešení je vygenerován náhodný počet řešení neboli náhodný počet jedinců (při biologické evoluci). Tato řešení (jedinci) jsou poté kříženi tak, že se vyberou dva náhodní (rodiče), jejichž vlastnosti se zkříží a vznikne nový prvek (potomek) viz obrázek 10. Takto vygenerovaná řešení (potomci) poté podléhají mutaci, která se projeví změnou bitu na opačný. Tato mutace je na obrázku 10 znázorněna červenou barvou, kdy se místo jednička změnila na nulu. (Anon. [b.r.])



Obrázek č. 10: Princip fungování evolučních algoritmů

Zdroj: Anon., [b.r.]. *Evoluční algoritmy a princip jejich fungování | NaPočítači.cz* [online] [vid. 2023i-01-02]. Dostupné z: https://www.napocitaci.cz/33/evolucni-algoritmy-a-princip-jejich-fungovani-uniqueidgOkE4NvrWuNY54vrLeM674MW0OH42R01Ag_rzFJ8D5c/?query=genetick%E9%20algoritmy&serp=1

Tato křížení i mutace tvoří nové možnosti řešení daného problému. Aby byl algoritmus úspěšný je potřeba zajistit vyšší množství křížení než mutací. Nicméně mutace zajišťují tvorbu nových řešení, je tedy důležité udržovat oba algoritmus v rovnováze. (Anon. [b.r.])

- **Clustering a klasifikace.** Clustering, neboli klastering či **shlukování** je metoda sdružování prvků se stejnými či podobnými vlastnostmi. Klasifikace představuje proces rozdělení prvků do kategorií. Tyto definice se mohou zdát stejné. Rozdíl mezi clusteringem a klasifikací je v tom, že klasifikace využívá trénovací sadu dat v rámci učení modelu takzvaně s učitelem. Clustering naopak trénovací sadu nevyužívá, je to tedy metoda učení modelu bez učitele, přímo a pouze se vstupními daty. (Anon. [b.r.])

1.3.3 Text mining, web mining a zpracování obrazu

Text mining, web mining i zpracování obrazu jsou podkategoriemi samotného data miningu. U text miningu dochází k dolování dat z textu. Tento text je v rámci text miningu transformován z nestrukturovaných dat na strukturovaná. Po transformaci jsou v textu hledány různé vzory, které jsou dále využity k různým účelům. Jednou z metod text miningu je například dolování na základě klíčových slov. (Anon. 2022b)

Web mining pracuje jak se strukturovanými, tak s nestrukturovanými či polo-strukturovanými daty. Jedná se o způsob dolování dat z webu. Toto dolování obsahuje automatické dolování a shromažďování dat z internetu. Důvodem pro web mining je lepší porozumění uživatelům daného webu například pomocí sledování vzorců vyhledávání. (Azad a Abhishek 2014)

Zpracování obrazu extrahuje jednotlivé atributy z obrazu. Jedná se o detekci předmětů či osob. Atributy při detekci obličejů jsou specifické rysy v obličeji, jako je šířka či hustota obočí, velikost očí, uší, tvar obličeje a další. Tyto tributy model rozpoznává na základě shluků v prostoru. Ke kvalitní detekci obličeje je potřeba kamera zohledňující vícerozměrnost obličeje (3D). Kamery na mobilních telefonech podporují často pouze 2D obraz, což může být pro uživatele nebezpečné, neboť bude možné mobilní telefon otevřít i fotografií. Tomuto nebezpečí se může uživatel vyhnout, například využitím mimiky při skenování obličeje. (Anon. [b.r.])

2 Využití data miningu v odvětví ekonomie a financí

2.1 Marketing

Marketing je odvětví, ve kterém byl a stále je data mining hojně využívaný. Podle českého ekonoma Jaroslava Světlíka je definice marketingu následující: *„Marketing je proces řízení, jehož výsledkem je poznání, předvídání, ovlivňování a v konečné fázi uspokojení potřeb a přání zákazníka efektivním a výhodným způsobem zajišťujícím splnění cílů organizace"*. Z definice vyplývá, že je potřeba plnit potřeby jak zákazníka, tak organizace a jaký může být lepší způsob než sestavení modelu předvídajícího chování zákazníka na základě historického chování jiných zákazníků. (Světlík 2005)

Zde je možné využít **analýzu nákupního košíku**, pomocí které je zákazníkům ke kupovanému předmětu A doporučen také předmět B, který si předešlí zákazníci kupovali s předmětem A. Analýza nákupního košíku tedy funguje na principu korelací mezi kupovaným zbožím. Konkrétním příkladem může být zákazník kupující si klávesnici, ke které mu systém před dokončením objednávky doporučí myš a podložku pod myš. Tímto doporučováním se zvedají tržby společnosti a zároveň je v některých situacích šetřen čas zákazníka, který by musel trávit čas hledáním. (Anon. 2015)

Další možností využití data miningu v marketingu je při segmentaci trhu. **Segmentace trhu** je metoda marketingového řízení, která je využita, pokud chce společnost vstoupit na nový trh. V rámci metody dochází k dělení zákazníků do skupin na základě kritérií. Nejčastějšími kritérii pro dělení je

- demografické kritérium – věk, pohlaví, náboženství, rodinný stav a další,
- geografické – bydliště zákazníka,
- socioekonomické – povolání, vzdělání, příjem a další,
- psychologické – životní hodnoty zákazníků, zájmy a další. (ManagementMania [b.r.])

2.2 Bankovnictví

Rozšířeným data minimgovým systémem v bankovnictví je automatický **scoring** klientů. Na scoring je možné se dívat z vícero úhlů. První využití scoringu klienta je při čerpání úvěru, ať už spotřebitelského, hypotečního či leasingu, je potřeba provést posouzení klienta, zda bude schopen daný úvěr splácet či nikoli. V současné době fungují ve většině bank scoringové modely, které automaticky posoudí, zda klient úvěr získá či nikoli. Model se rozhoduje

na základě stanovených podmínek zahrnujících minimální příjem klienta, maximální náklady či celkovou maximální zadluženost. Klienti, které model posoudit nedokáže, například z důvodu složitosti případu, jsou předány na posouzení pracovníkovi banky. (Anon. [b.r.])

Pokud již klient u banky úvěr má, je součástí systémů i databáze společnosti, u těchto klientů je tedy možné provádět scoring průběžně a automaticky, například každý den, měsíc či kvartál. Podle toho, jaký má klient scoring může společnost zjistit například jaká je jeho platební morálka. Na základě scoringu lze identifikovat klienty, kteří pravděpodobně odejdou v nejbližších měsících, či vyčíslit dlouhodobou hodnotu zákazníka neboli Customer Value. (Novotný et al. 2005)

Při sestavování scoringového modelu se běžně vychází z těchto datových okruhů:

- **Behaviorální data.** Jedná se o data reflektující chování zákazníků jako jsou u bankovních institucí například frekvence transakcí, výše realizovaných transakcí, četnost výběrů z bankomatů a další.
- **Demografická data** mohou u právnických osob zobrazovat například jejich zisky, segment ve kterém působí či právní formu. U fyzických osob může být uvedeno pohlaví, věk, adresa atd.
- **Produktová data** obsahují informace o produktech, které jednotliví klienti využívali a jak často je využívali.
- **Kontaktní data** mohou obsahovat například záznamy využití klientské linky společnosti klientem či reakce klienta na marketingové kampaně, například v aplikacích či na webu společnosti.
- **Externí data** jsou data o klientovi z veřejně dostupných externích zdrojů. (Novotný et al. 2005)

2.3 Pojišťovnictví

Pojišťovny, stejně jako banky využívají **scoringové modely** a modely pro **monitoring**. Monitoring má za úkol porovnávat aktuální sady s historickými a poukazovat na výkyvy mezi nimi. Častým využitím data miningu v pojišťovnictví jsou také modely pro **detekci pojistných podvodů**, pomocí kterých lze identifikovat podezřelé pojistné události na základě historických dat o podvodech (sdílených mezi pojišťovnami napříč trhem). Sdílení dat o podvodech napříč pojišťovnami umožnila Česká asociace pojišťoven (ČAP), která dala impuls pro vytvoření systému SVIPO, pomocí kterého si pojišťovny mohou data vzájemně sdílet. Systém SVIPO tedy pomáhá jak při detekci již vzniklé podezřelé pojistné události, tak při prevenci proti ještě nevzniklým podvodům. (oPojištění [b.r.])

Pojistné podvody vznikají jak u životního, tak u neživotního pojištění. Podle průzkumu ČAP z roku 2021, za rok 2020 docházelo častěji k podvodům u neživotního pojištění, a to primárně u pojištění vozidel viz Tabulka 2. Zároveň dochází ke kontinuálnímu meziročnímu růstu pojistných podvodů. Průměrná změna mezi roky 2019 a 2020 je 110,4 procenta, přičemž nejvyšší nárůst je zaznamenán u pojištění odpovědnosti a to až 170 %. Tyto hodnoty poukazují na důležitost řešení pojistných podvodů, při kterém může pomoci další vývoj nástrojů data miningu a strojového učení. (Anon. 2021b)

Tabulka 2: Pojistné podvody 2020

Všechny šetřené případy pojistných podvodů ve specifikovaných oborech pojištění v roce 2020				
obor pojištění	počet případů (v ks)	rozdělení	výše prokázané hodnoty (v tis. Kč)	meziroční změna
Pojištění vozidel	4 686	49%	359 739	105%
Pojištění majetku	1 858	19%	356 900	89%
Pojištění odpovědnosti	1 126	12%	384 490	170%
Pojištění osob	1 962	20%	130 225	79%
Celkem	9 632	100%	1 231 354	109%

Zdroj: Anon., 2021b. POJISTNÉ PODVODY 2020: Nejčastěji se lidé snaží o podvod v pojištění vozidel, stoupají podvody v pojištění odpovědnosti [online] [vid. 2023-02-19]. Dostupné z: https://www.cap.cz/tiskove-centrum/tiskove-zpravy/104794-pojistne-podvody-2020-nejcasteji-se-lide-snazi-o-podvod-v-pojisteni-vozidel-stoupaji-podvody-v-pojisteni-odpovednosti

Pojišťovnám při detekci podvodů pomáhají technologie, či například typy médií, která jsou schopna přenášet data. Příkladem mohou být fotografie pořízené například přímo při pojistné události. Fotografie obsahují metadata, ze kterých je viditelný čas pořízení fotografie. Čas je faktor na základě kterého dokáží modely či pojišťovny odhalit množství nejasností. Dalším nosičem pro pojišťovnu důležitých dat může být systém GPS, kterým je možné monitorovat jak polohu osob, tak objektů. Například u vozidel, u kterých podle průzkumu ČAP za rok 2020 dochází k největšímu množství podvodů, je instalován bezdrátový modul GPS. Tento modul může pojišťovně poskytnout informace o polohách, časech či ujetých vzdáleností vozidla. Na základě těchto charakteristik může pojišťovna také předvídat rizika. (Mayer-Schönberger a Cukier 2014)

V České republice je možné u některých pojišťoven platit pojistné na základě ujetých kilometrů, protože čím více klient najede, tím je větší pravděpodobnost nehody. V UK a USA si klienti mohou zařídit pojištění za cenu stanovenou podle toho kam a kdy reálně jezdí,

v tomto případě pojišťovny nestanovují cenu na základě věku či bodů. (Mayer-Schönberger a Cukier 2014)

Zajímavým příkladem z praxe u životního pojištění je prediktivní model vyvinutý společností Deloitte Consulting pro pojišťovnu Aviva, který nahradil krevní testy u lékaře. Model pracuje na základě stovek proměnných, obsahujících převážně otázky na životní styl, dobu sledování televize, náplň práce, koníčky a další. Model na základě těchto proměnných vyhodnotí, jak vysoké je u klienta riziko závažného onemocnění, jako je cukrovka, deprese či vysoký krevní tlak. Tato metoda je pro pojišťovnu méně nákladná než lékařské vyšetření a může být pro některé klienty komfortnější. Zároveň se model podle (Mayer-Schönberger a Cukier 2014) osvědčil.

3 Úloha zpracovaná v programu IBM SPSS Modeler na téma odchodů klientů z banky XY

Vstupní data set obsahuje 10 000 řádků reprezentujících klienty americké banky XY a informace o nich v rámci 14ti prediktorů (ve sloupcích). Úloha je zpracovávána v programu IBM SPSS Modeler vyvinutém americkou mezinárodní společností International Business Machines Corporation (IBM). Při sestavování modelu se autorka řídí obecnou volně dostupnou metodologií bez vlastníka **CRISP-DM**. Principem metodologie je nejprve porozumět problému (1. krok) poté porozumět samotným datům (2. krok), následuje příprava dat (3. krok), modelování (4. krok), evaluace (5. krok) a nasazení modelu do praxe (6. krok).

Cílem této úlohy je na základě dostupných (historických) dat predikovat, zda klient z banky XY odejde či nikoli. S tímto modelem se může banka více zaměřit na typy klientů, kteří pravděpodobně odejdou a zaujmout je například cílenou marketingovou kampaní.

3.1 Porozumění problému

Porozumění problému odchodů klientů z banky XY je složité bez asistence specialisty fungujícího uvnitř dané banky. Pokud je však na problematiku odchodů klientů z bank nahlíženo všeobecně, jedná se o přirozený jev, který je způsoben konkurenčním bojem mezi bankami. Banky si mohou konkurovat několika způsoby.

- **Snižování úrokových sazeb.** Klienti žádající o úvěr hledají co nejvýhodnější podmínky, tedy běžně co nejnižší úrok. Pokud je klient rizikovější, například kvůli nízkému věku či špatné platební morálce, nehledá převážně podle sazby, ale spíše podle kritérií banky. Každá banka se tedy může částečně zaměřovat na určitý typ klientů.

  Úrokové sazby komerčních bank se mění podle základní repo sazby České národní banky. Čím vyšší je základní repo sazba, tím vyšší jsou úrokové sazby komerčních bank. V roce 2022 zvedla Česká národní banka repo sazbu až na 7 % což je nejvyšší hodnota za posledních jedenáct let. Česká národní banka se snaží o zbrzdění ekonomiky a omezení přílivu peněz do ekonomiky ve snaze o snížení inflace, která se v roce 2022 dostala až na 18 %. (Anon. 2022a)

Tato situace je pro banky poměrně složitá, neboť chtějí poskytovat úvěry, ale klientů, kteří by měli zájem není tolik kvůli vysokým úrokovým sazbám. Banka tedy musí zvážit, kam až se vyplatí sazbu pro klienty snižovat, aby se to z obchodního hlediska pro banku vyplatilo. V této situaci se banky mohou více zaměřit na ostatní poskytované produkty jako je pojištění či investice a na jejich zdokonalování.

Co se týče hypotečních úvěrů v souvislosti se zpracovávaným tématem odchodů klientů z banky, je důležité zmínit fixační období hypotéky. Fixační období je období, po které má klient u dané banky na hypotéce stále stejný úrok. Toto období je smluvně podložené a nejčastěji trvá 5 let (případně 10 let). Vždy na konci fixačního období může klient doplatit hypotéku bezplatně, či přejít k jiné bance. Refinancování hypotečních úvěrů může být tedy častým důvodem k odchodu klientů z banky.

- **Zvyšování úroků na spořících účtech.** Situace během roku 2022, kdy se zvyšovaly úroky u úvěrů, se zvyšovaly také úroky na spořících účtech. Vyšší sazby pro klienty všeobecně nabízí spíše banky, které nově expandovaly na daný trh, protože potřebují získat klienty na úkor vyššího zisku. Banky, které mají dostatečné množství klientů, nejsou tolik tlačeny do přílišného zvyšování úroků na spořících účtech.

  Běžné a spořící účty jsou základními službami bank. V případě, že má klient u banky účet, je vysoká pravděpodobnost, že si u banky zařídí i jiné produkty, jako investice či úvěr, pokud bude tyto produkty vyhledávat. Proto je výhodné z pozice bank běžné a spořící účty nezanedbávat. Mezi benefity na běžných účtech, kterými si banky v současné době konkurují, patří například cashback, tedy vracení určitého procenta zpět z nákupu. Mezi další benefity patří vedení účtu zdarma, výběry ze všech bankomatů zdarma, slevové kupony v konkrétních prodejnách atd. U spořicího účtu je rozhodující úrok, výše vkladu do které úrok platí, možnost zřízení bez běžného účtu a další podmínky. Některé banky nabízí také jednorázovou odměnu za založení běžného či spořícího účtu. Odměna je často finanční nebo materiální v podobě produktu (například fitness náramek atd.).

- **Portfolio nabízených produktů.** To, kolik má banka klientů může záviset také na velikosti sortimentu, který banka nabízí. Mezi bankovní produkty v současné době patří
  - aktivní bankovní produkty jsou úvěrové produkty,

- pasivní bankovní produkty obsahují běžný a spořící účet, termínované vklady, stavební spoření a další (Syrovátková [b.r.]).

V současné době bankovní instituce ve snaze zaujmout co největší část trhu vytvářejí k sobě přidružené stejnojmenné pojišťovny v rámci kterých rozvíjejí životní i neživotní pojištění. Příkladem banky, která má již stejnojmennou pojišťovnu, poskytující poměrně nově i životní pojištění je ČSOB.

- **Image a marketing banky.** Na to, zda banka obstojí v konkurenčním boji má vliv také to, jak dlouho banka na trhu působí a jaké si vybudovala jméno. Banky působící na českém trhu delší dobu jsou pro klienty známé a mohou k nim mít větší důvěru, než k bankám vstupujícím na trh nově. Velký vliv má však dobře přizpůsobený marketing banky, její produktové řízení či dobrá strategie nastavení distribučních kanálů, díky kterému se mohou i menší banky dostat mezi leadery na trhu.

Všechny odchody klientů však nemusí být způsobeny působením konkurence na trhu, ale může se jednat o jiné důvody. Příkladem může být situace okolo Sberbank CZ. Sberbank je původem ruská banka, které se dostala do problémů kvůli rusko-ukrajinskému konfliktu. Klienti vystupující proti ruské agresi začali ve velkém vybírat své úspory z banky a rušit své účty a další produkty. U banky došlo k takovému odlivu vkladů, že se stala nedostatečně likvidní a Česká národní banka zahájila kroky k odebrání její bankovní licence.

V souvislosti s těmito kroky Česká národní banka odebrala bance Sberbank pravomoc nakládat s aktivy a pasivy. V tuto chvíli byly vklady klientů zmraženy a řešení této krizové situace bylo předáno na Garanční systém finančního trhu, který pomocí zvolené finanční instituce organizuje vyplácení pojištěných vkladů klientům. Pojištěný vklad mají u bank klienti do 100 000 EUR, cca. 2,5 milionů Kč. (Anon. [b.r.])

V rámci prvního bodu metodologie CRISP DM (Porozumění problému) je také stanovení cíle, který by měl model splnit. Cílem modelu je tedy identifikovat klienty, kteří z banky XY s určitou pravděpodobností odejdou.

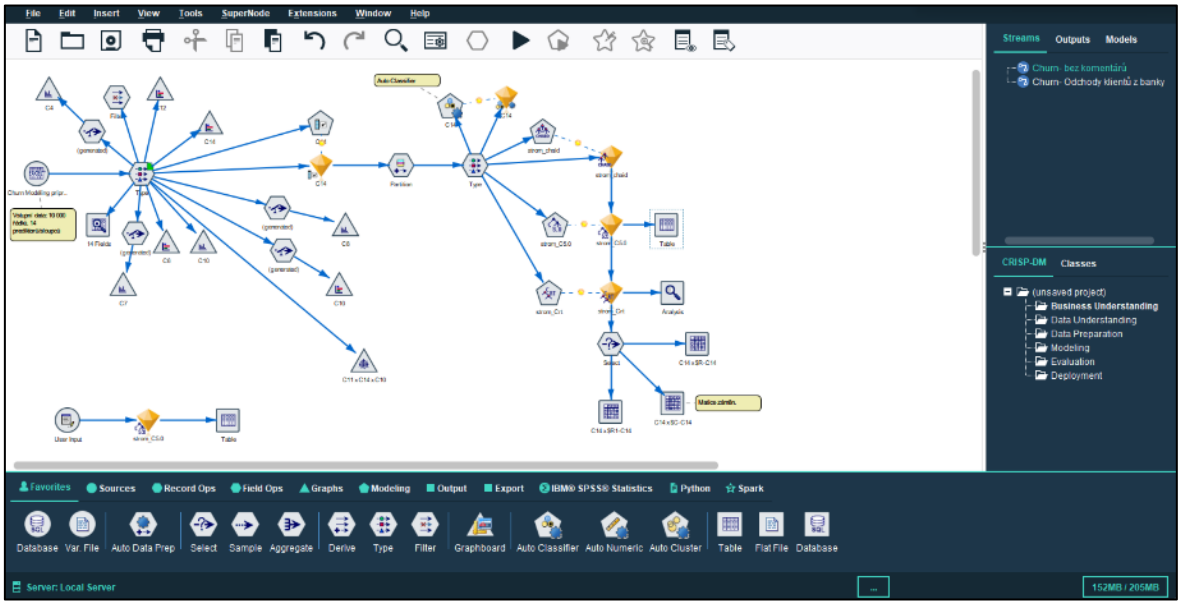
K popisu dat je již využit IBM SPSS Modeler, proto je nejprve potřebné porozumět jeho prostředí a terminologii, která se při práci s ním využívá.

3.1.1 Prostředí programu IBM SPSS Modeler

Na obrázku 11 je viditelné prostředí programu a zároveň celý stream řešené úlohy *„odchodů klientů z banky“.* V okně palety uzlů nacházející se na dolní liště jsou ikony, neboli **uzly**, každý uzel představuje již nadefinovaný proces, který bude po vložení uzlu proveden. Uzly lze také

vytvářet a to například pomocí jazyka Python. Uzly jsou na liště zdrojové a procesní. Zdrojovými uzly se nahrávají data do programu SPSS Modeler a je možné je najít ve složce Sources. Nahrávat lze například data z databáze, Excelu, SASu či SPSS Statisticsu. Procesními uzly jsou Record a Field Operations, mezi Record Ops. patří uzly pro práci s řádky, ve Field Ops. jsou uzly pro práci se sloupci (prediktory). Jednotlivé uzly se spojují a v datovém okně tvoří řetězec, neboli **datový stream.** Dále jsou na liště uzly pro grafy, modelování či Export dat.

V pravé horní části obrazovky, viz obrázek 11, se nachází okno správce výstupů, ve kterém jsou viditelné aktuálně spuštěné datové streamy a modely. V pravé dolní části, pod oknem správce výstupů, je projektové okno ve kterém se jednotlivé provedené činnosti řadí podle metodologie CRISP_DM.



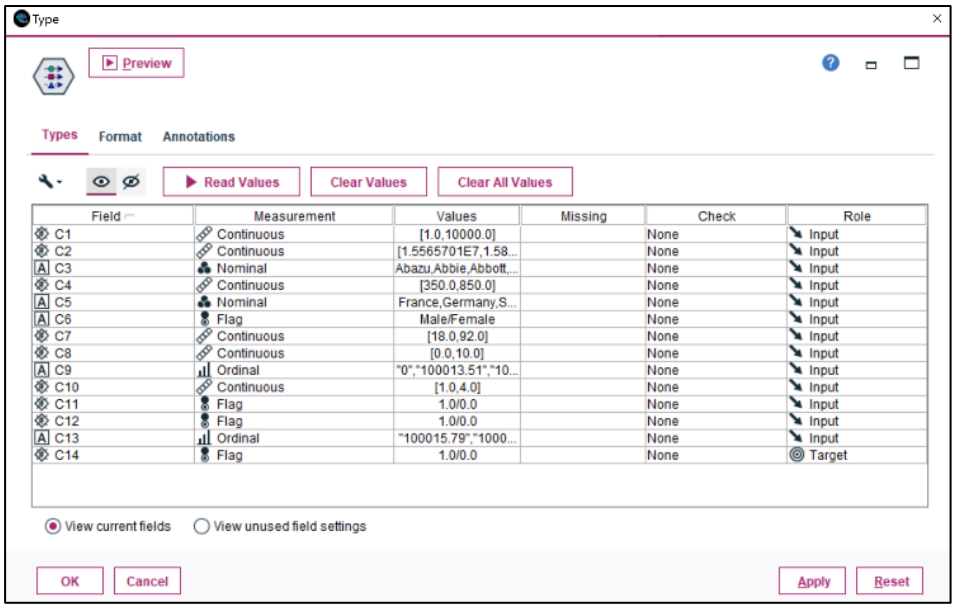
Obrázek č. 11: Kompletní stream řešené úlohy

Zdroj: Vlastní zpracování

3.2 Porozumění datům

V této fázi je již jasný problém (odchody klientů z banky) i cíl, ke kterému je směřováno. Nyní je potřeba porozumět datům. Zdrojem datové sady, která je k řešení úkolu k dispozici je platforma Kaggle, která obsahuje více než 50 000 volně dostupných datových sad k užití (Anon. [b.r.]). U jednotlivých datových sad jsou informace o jejich původu, velikosti či cíli, ke kterému je možné při modelování směřovat. Dalším krokem v rámci porozumění datům je **popis dat** neboli zjišťování základních informací o datové sadě. Než se však dostaneme k uzlu Type, ve kterém je viditelný náhled na data viz obrázek 12, je potřeba data nahrát do programu SPSS Modeler. Uzel pro nahrávání dat je hledán v záložce *„Sources“,* uzel Excel, protože vstupní data nahráváme z Excelu ve výchozím formátu .xlsx. Po nahrání dat je připojen uzel Type viz

obrázek 11, ve kterém je zkontrolováno, zda program ke všem atributům přiřadil správné datové typy a typy proměnných.



Obrázek č. 12: Uzel Type

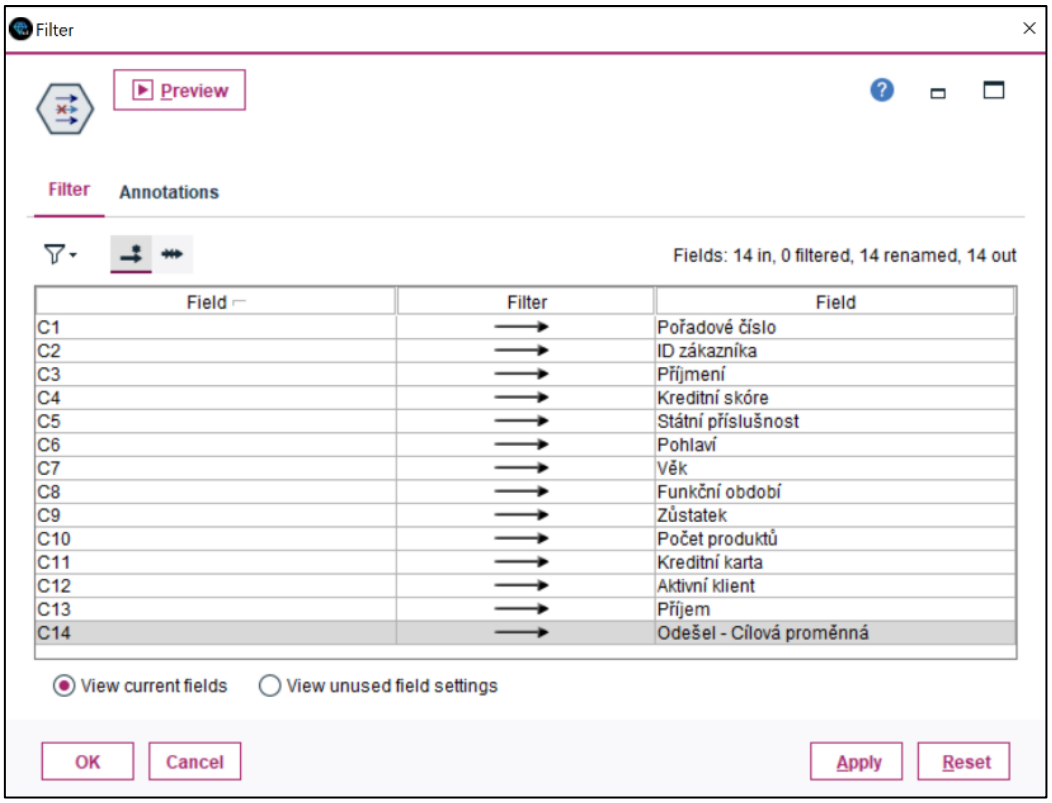
Zdroj: Vlastní zpracování

V levém sloupci uzlu **Type** s pojmenováním *„Field“* jsou datové typy a názvy atributů (C1, C2, …). Datovým typem je zde **string (A)**, neboli textový řetězec, v datové sadě se nachází buďto jako nominální proměnná, nebo flag. Flag je proměnná nabývající pouze dvou hodnot (např. 1, 0). Flag je například cílová proměnná zobrazující, zda klient odešel či nikoli (1, 0). Druhým datovým typem je zde **reálné číslo (R)**. To se v sadě nachází jako *„Continuous“* (=celé nebo desetinné číslo) a jako flag. Názvy atributů v datové sadě jsou značeny písmenem C a číslem. Reálné názvy polí jsou viditelné v uzlu **Filter**, kde je možné pole přejmenovat pro lepší přehlednost viz obrázek 13.

Ve třetím sloupci uzlu Type je náhled do sady a ve čtvrtém sloupci je možné vidět, zda se v sadě nachází prázdné hodnoty. V této sadě prázdné hodnoty nejsou. V posledním sloupci uzlu Type lze nastavit, proměnné co do modelu vstoupí (Input), případně vypnout nepotřebné, stanovuje se zde také cílová proměnná (Target). Input jsou tedy vstupní proměnné neboli nezávislé proměnné či **prediktory**. Target je výstupní proměnná neboli cílová, ta je na prediktorech závislá, neboť podle toho, co prediktory obsahují je vyhodnocen Target (zda klient odejde či nikoli).

Po kontrole dat v uzlu Type je možné vidět, že atributy C9 (průměrný zůstatek) a C13 (mzda) byly zařazeny do datového typu string. Jedná se však o reálná čísla, a ne textový řetězec, který atributům program přiřadil. Chyba je ve zdrojových datech, částky jsou zaznamenány podle

normy, která desetinné číslo odděluje tečkou místo čárky, (př. 10000.24). V tomto případě vnímá program číslo jako text, proto je potřeba ve zdrojových datech nahradit tečku čárkou.



Field	Filter	Field
C1	→	Pořadové číslo
C2	→	ID zákazníka
C3	→	Příjmení
C4	→	Kreditní skóre
C5	→	Státní příslušnost
C6	→	Pohlaví
C7	→	Věk
C8	→	Funkční období
C9	→	Zůstatek
C10	→	Počet produktů
C11	→	Kreditní karta
C12	→	Aktivní klient
C13	→	Příjem
C14	→	Odešel - Cílová proměnná

Obrázek č. 13: Uzel Filter

Zdroj: Vlastní zpracování

Důležitou součástí fáze *„porozumění datům“* je **zkoumání dat** v rámci kterého jsou data vizualizována pomocí grafů a tabulek. Zjišťovány jsou například počty opakujících se hodnot v rámci jednoho atributu, maxima, minima, či průměrné hodnoty.

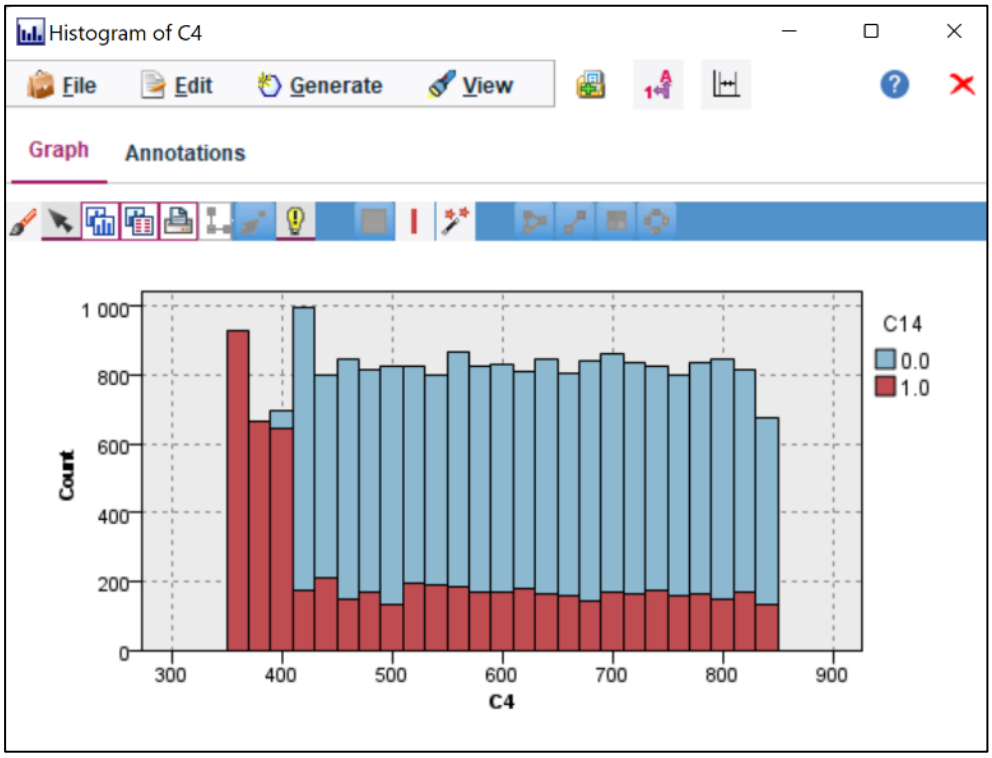
Již nyní lze předpokládat, že první tři prediktory (**C1, C2 a C3**) nebudou pro model důležité, proto s nimi ve vizualizacích a později ve tvorbě modelu není počítáno. Rychlý náhled na data lze provést pomocí uzlu Data Audit, který zobrazí grafy k jednotlivým atributům, tyto grafy jsou v diplomové práci vizualizovány jednotlivě k vybraným atributům.

C4 je **kreditní skóre**, které by již určitou váhu při rozhodování modelu mít mohlo. Kreditní skóre znázorňuje platební morálku klienta, pohybuje se mezi hodnotami 300-850. Čím nižší číslo, tím horší skóre. Dobré skóre se pohybuje okolo 670 a výše. (Anon. [b.r.])

Kreditní skóre je ve vztahu k cílové proměnné znázorněno grafem č. 2. Na grafu je vidět, že klienti s velmi špatným skóre odcházejí (červené sloupce), protože v bance pravděpodobně zkoušeli žádat o úvěr, který na základě kreditního skóre nezískali. Nemusí však jít pouze

o klienty, kteří žádali o úvěr. Může se jednat například o klienty s nižšími příjmy, kteří mají dlouhodobě problémy platit včas a změnili banku na základě lepší nabídky, či bonusu nabízeným jinou bankou.

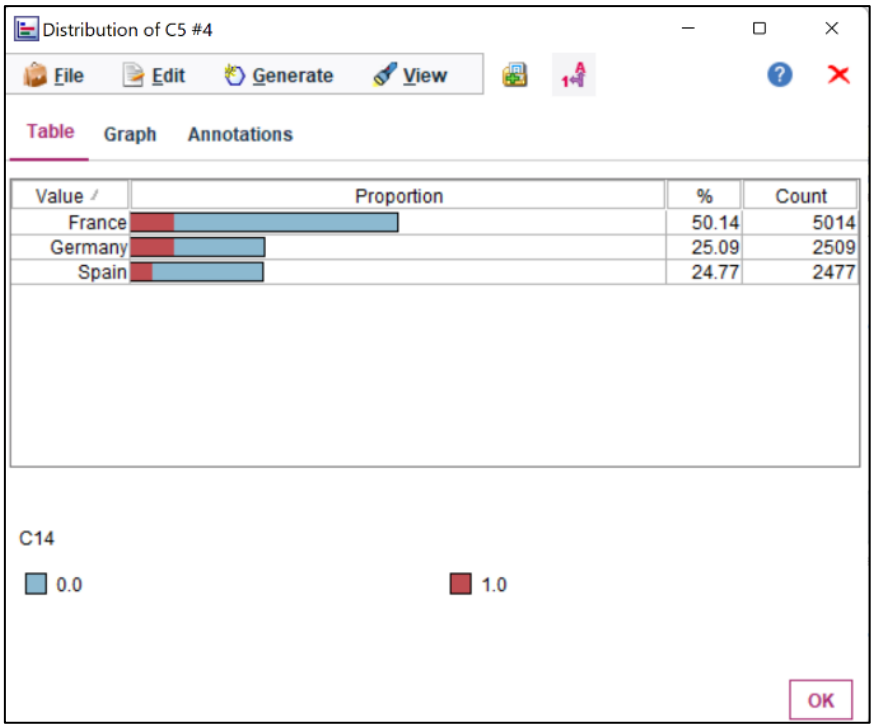
Zároveň je vidět, že klienti se skórem nad 400, které je stále velmi špatné v bance ve velké míře zůstali. Z toho vyplývá, že banka poskytuje úvěry i klientům s velmi špatným skóre, jedná se tedy nejspíše o benevolentnější banku v porovnání s konkurencí. Průměrné skóre u klientů, kteří z banky neodešli je 652 (dobré skóre), přičemž minimum je 405 (velmi špatné skóre).



Graf č. 2: Cílová proměnná v závislosti na kreditním skóre

Zdroj: Vlastní zpracování

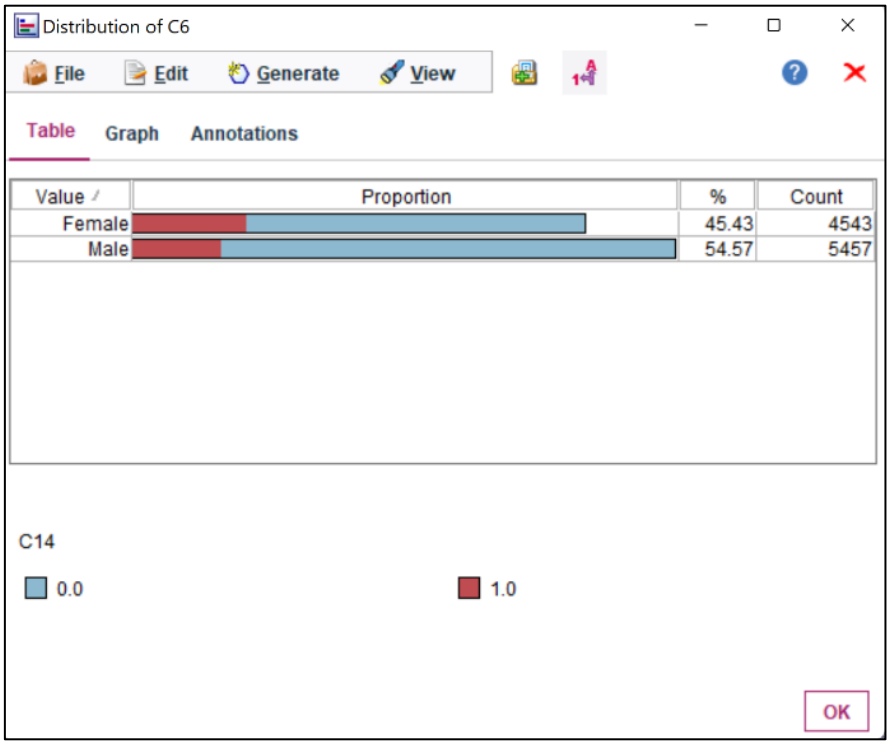
Dalším prediktorem je **C5 – státní příslušnost klienta**. Nejvíce klientů z datové sady pochází z Francie a jedná se přibližně o 50 % klientů, dalšími národnostmi jsou Němci a Španělé viz graf č. 3. Po balancování proměnné lze zjistit, že z banky nejvíce odcházejí Němci a to až 34 % z celkové množiny Němců.



Graf č. 3: Cílová proměnná v závislosti na státní příslušnosti

Zdroj: Vlastní zpracování

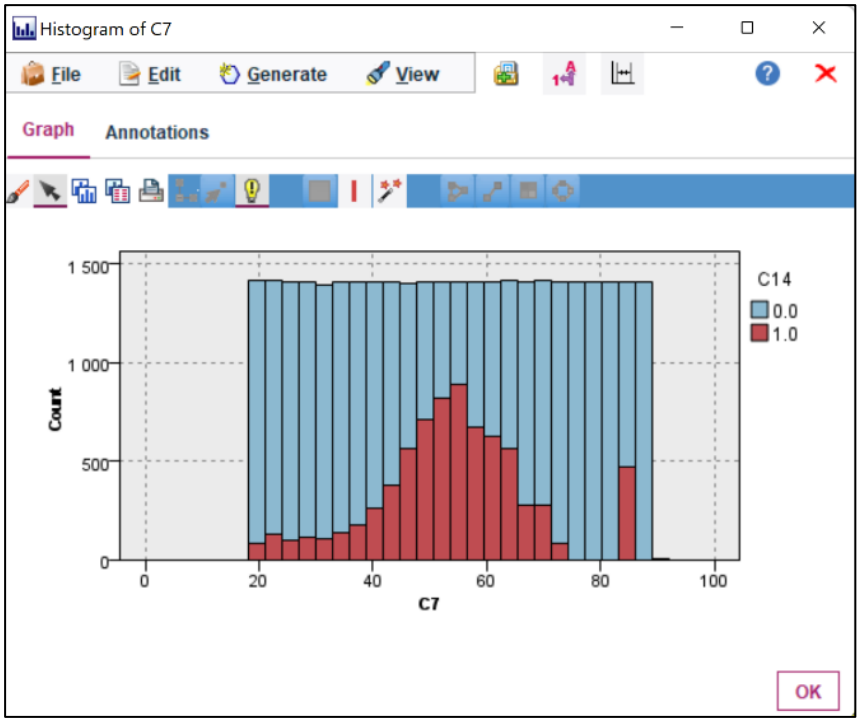
Prediktor **C6** je **pohlaví klienta** viz graf č. 4. Co se týče pohlaví nachází se v sadě více mužů, kterých je přes 54 %, nicméně z banky odcházejí více ženy.



Graf č. 4: Cílová proměnná v závislosti na pohlaví klienta

Zdroj: Vlastní zpracování

Prediktor **C7** znázorňuje **věk klienta** a v rozhodování modelu by mohl sehrát zásadnější roli než například pohlaví či národnost. V datové sadě se nachází klienti od 18 až do 92 let. Průměrný věk u klientů, kteří z banky neodešli je 37 let, naopak u klientů, kteří odešli je průměrný věk 45 let. Nejčastěji dle grafu č. 5 odcházejí klienti v 55 letech.



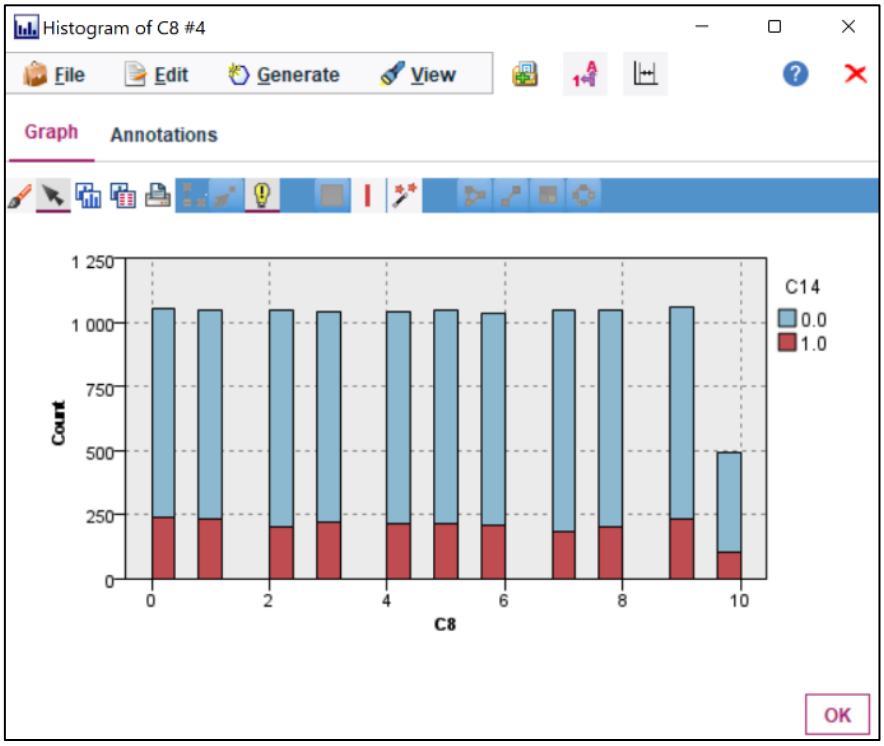
Graf č. 5: Cílová proměnná v závislosti na věku klienta

Zdroj: Vlastní zpracování

C8 znázorňuje **funkční období**. Tento prediktor určuje, kolik let klient strávil v bance XY k datu vytvoření datové sady, či k datu jeho odchodu. Maximální funkční období klientů je 10 let viz graf číslo 6. Objevují se zde také klienti, kteří v bance nestrávili ani 1 rok. Důvodem může být například odměna za založení účtu, kterou klienti vybrali a účet zrušili, převedení či splacení spotřebitelského úvěru do jednoho roka, či využití speciálního produktu, kterým může být cestovní pojištění, pojištění karty atd. Z grafu 6 je zřejmé, že mezi funkčními obdobími 0-10 nejsou přílišné výkyvy mezi počty klientů, ať už odcházejících či těch kteří zůstávají.

Odchyluje se pouze desetileté období, a to celkově nižším počtem klientů, než mají zbylá funkční období. To je nejspíše způsobeno tím, že klienti chtějí změnu z důvodu dlouhého setrvávání u jedné instituce. Pokud má klient u banky produkt 10 let, je tento produkt již zastaralý a v jiné bance dokáží klientovi s největší pravděpodobností nabídnout lepší

podmínky. Tyto změny k výhodnějším podmínkám pro klienta vznikají působením konkurence na trhu bankovních produktů.



Graf č. 6: Cílová proměnná v závislosti na funkčním období klienta

Zdroj: Vlastní zpracování

Prediktorem **C10** je **počet produktů** klienta viz graf číslo 7. Klienti z analyzované datové sady využívají jeden až čtyři produkty. Jeden produkt je využíván více jak polovinu klientů, konkrétně se jedná o 5 084 klientů z 10 000. Dva produkty využívá téměř polovina klientů, konkrétně 4 590. Tři nebo čtyři produkty využívá minimum klientů, dohromady se jedná o 327 klientů.

Pokud je analyzována závislost cílové proměnné na konkrétním prediktoru viz graf 7, je zjištěno, že klienti, kteří mají dva produkty odcházejí méně než klienti s jedním produktem. Tato skutečnost není příliš překvapivá. Klienti vlastnící jeden produkt, využívají s nejvyšší pravděpodobností běžný účet. Pokud má klient pouze běžný účet je pro něj jednoduché banku kdykoli změnit, protože nemá jiné produkty, které by mohly jeho rozhodování ovlivnit. Do této množiny mohou patřit také klienti, co využili jednorázovou, výhodnou nabídku banky, například založení účtu za odměnu a po výběru odměny účet zrušili.

Klienti vlastnící dva produkty, nejspíše vlastní běžný a spořící účet, či běžný účet a úvěr, kreditní kartu, či určitý typ pojištění. Tito klienti se rozhodli využít dalšího produktu banky, nejspíše z důvodu důvěry v danou banku, či z důvodu získání výhodného produktu v porovnání s nabídkou bank na trhu. Z tohoto důvodu je logické, že klienti neodcházejí v takové míře.

Pozastavení však přichází nad hojnými odchody zákazníků, kteří vlastní tři nebo čtyři produkty. U těchto dvou případů dochází k odchodům zákazníků z 86 %. Přitom by se dalo tvrdit, že čím má klient u banky více produktů, tím spíš z ní neodejde, protože má v banků důvěru, proto má tolik produktů. Po detailnější analýze je však zjištěno, že průměrné funkční období klientů, kteří z banky odešli a vlastnili 3 nebo 4 produkty je 5, 09 roku a modus je 5. Pět let je podle MONETA Money Bank, a. s. zároveň standardní doba pro fixaci hypotéky (Anon. [b.r.]).

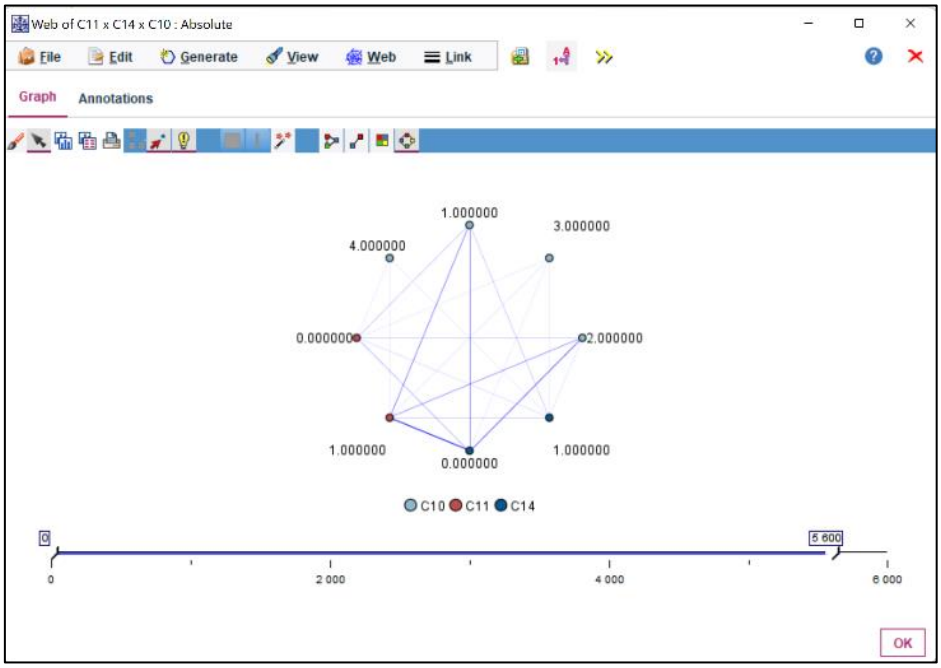
Na základě těchto zjištění autorka usuzuje, že klienti se třemi či čtyřmi produkty mohou u banky pouze využívat hypoteční úvěr, se kterým se pojí další produkty, jako založení běžného účtu, pojištění nemovitosti, či pojištění schopnosti splácet. Po skončení doby fixace poté klient z banky odchází za lepší úrokovou sazbou.

Graf č. 7: Cílová proměnná v závislosti na počtu produktů klienta

Zdroj: Vlastní zpracování

Prediktor **C11** znázorňuje, zda klient vlastní (1) či nevlastní (0) **kreditní kartu**. Ke zkoumání dat je využit pavučinový graf (Web), který dokáže porovnat více atributů. Konkrétně jsou porovnávány atributy: C10 (počet produktů), C11 (vlastní/nevlastní kreditní kartu) a C14 (odešel/neodešel z banky). Tloušťka čáry určuje větší počet klientů viz graf číslo 8.

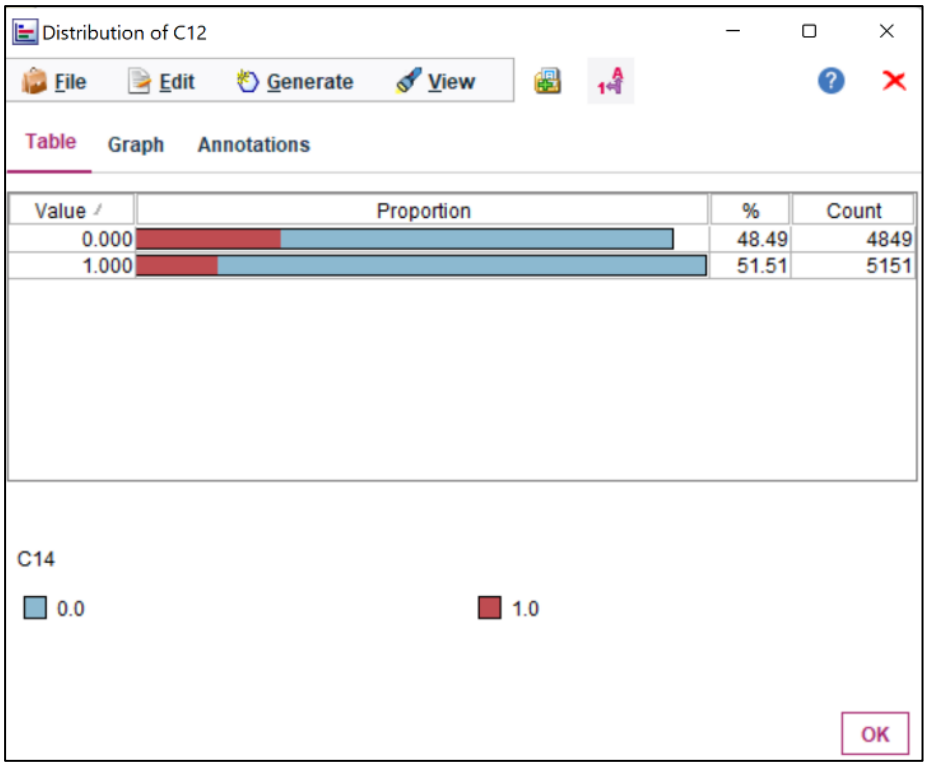
Banka XY má celkově více klientů, kteří vlastní kreditní kartu, 7 055 klientů kartu vlastní, 2 945 kartu nevlastní. S přihlédnutím na cílovou proměnou odchází spíše klienti, kteří kreditní kartu nevlastní. Na grafu je viditelné že mezi klienty, kteří neodešli (C14=0) a mají kreditní kartu C11=1) je silnější přímka, tedy je zde více klientů. Co se týče počtu produktů, je z grafu viditelné, že nejvíce klientů má 1 či 2 produkty. Dále je vidět, že více klientů s 1 a 2 produkty (C10=1, 2) má kreditní kartu (C11=1).



Graf č. 8: Cílová proměnná v závislosti na vlastnictví kreditní karty a počtu produktů

Zdroj: Vlastní zpracování

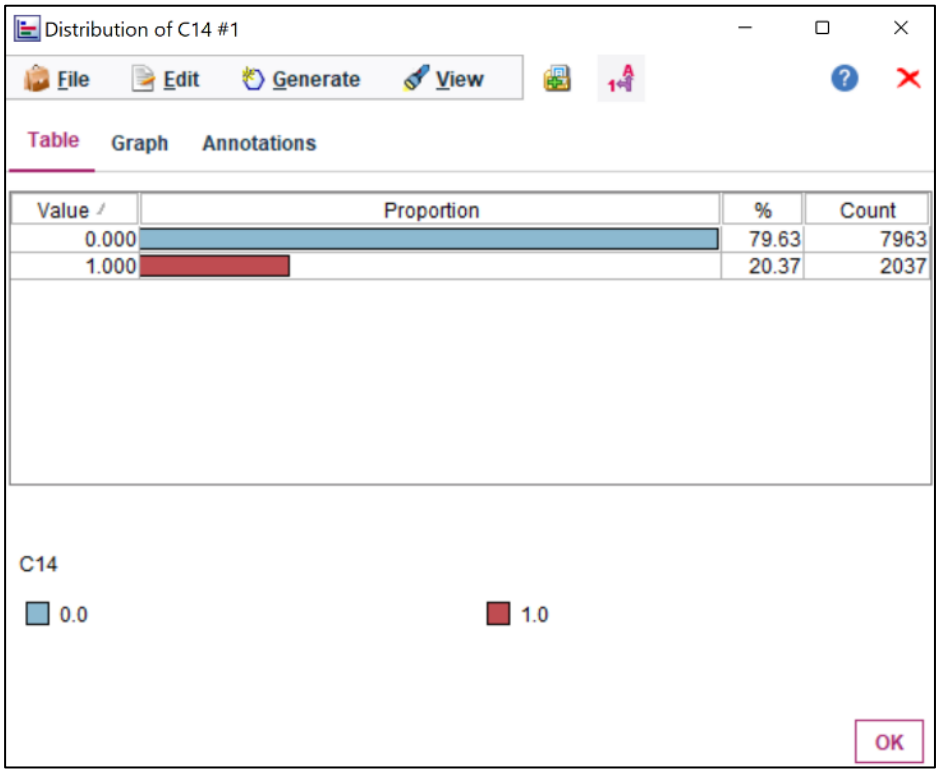
C12 znázorňuje **aktivního klienta** banky. Aby byl klient aktivní, je potřeba aby splňoval určité podmínky aktivity, jako počet ověření zůstatku na účtu během určitého časového období, počet převodů v rámci banky, převody mimo banku, zrušení trvalého příkazu atd. (Anon. 2002). Z grafu číslo 9 je viditelné, že banka XY má více aktivních klientů (téměř 52 %), než neaktivních a zároveň neaktivní klienti více odcházejí.



Graf č. 9: Cílová proměnná v závislosti na aktivitě klienta

Zdroj: Vlastní zpracování

C14 je **cílovou proměnnou** určující, zda **klient z banky XY odešel (1) či neodešel (0)**. Z grafu č. 10 je viditelné, že má banka fluktuaci okolo 20 %.

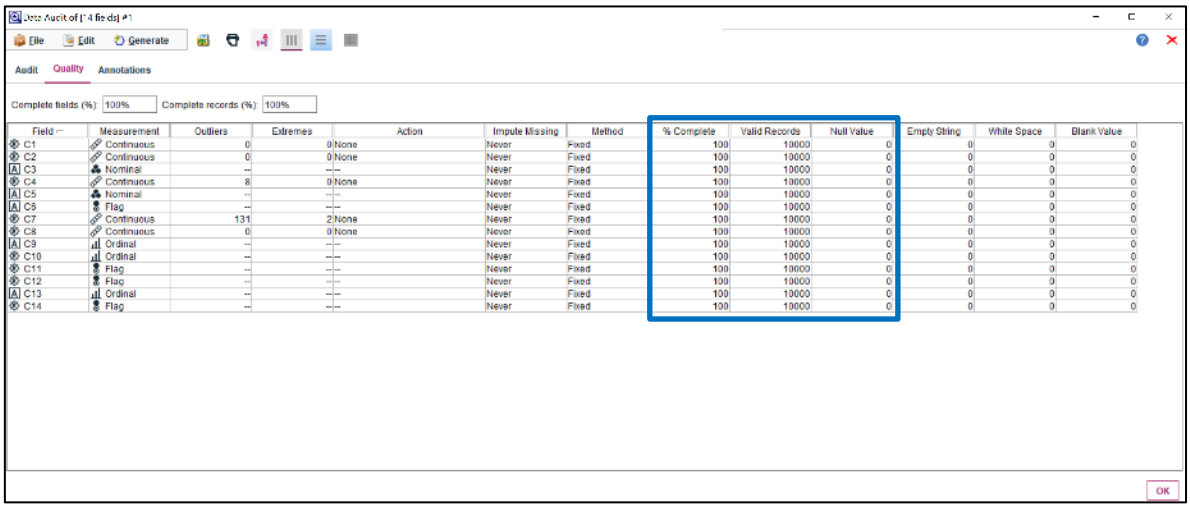


Graf č. 10: Cílová proměnná

Zdroj: Vlastní zpracování

Posledním krokem v této fázi je ověření kvality dat. Pokud je zjišťována kvalita dat, zjišťuje se například, zda data obsahují prázdné hodnoty či zda data obsahují chyby. Prázdné hodnoty již bylo možné vidět v uzlu Type viz graf číslo 12. V uzlu Data audit je však podrobnější přehled viz obrázek číslo 14.

V uzlu Data Audit v záložce Quality lze zjistit nulové hodnoty, procentuálně vyjádřenou kompletnost záznamů či počet validních záznamů. Pokud jsou nulové hodnoty nalezeny, je potřeba je vhodným způsobem ošetřit, a to buďto doplněním (například průměrem) nebo odstraněním. V datové sadě odchodů klientů z banky XY nejsou nalezeny žádné nulové hodnoty a všechny hodnoty jsou kompletní, data jsou tedy v dobré kvalitě.



Obrázek č. 14: Data Audit

Zdroj: Vlastní zpracování

3.3 Příprava dat

Ve fázi porozumění datům byla data získána, popsána, zkoumána a byla ověřena jejich kvalita. Po ukončení této fáze je možné pokračovat samotnou přípravou dat v rámci které jsou data čištěna o případné nulové hodnoty, či jsou nulové hodnoty nahrazovány na základě jiných hodnot. Poté jsou data **formátována** tak, aby splňovala určitou formu, která je k modelování vyžadována, například pořadí či specifické označení sloupců.

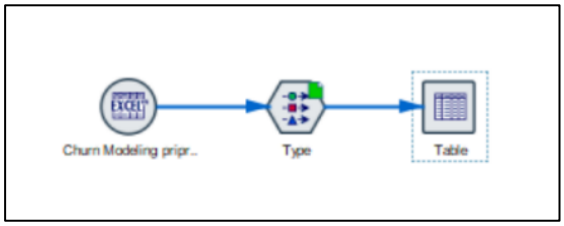
V případě úlohy *„odchodů klientů z banky XY“*, nebyly rozpoznány nulové hodnoty. Co se týče formátování byly zvoleny univerzální názvy atributů (C1, C2, C3 atd.) pro bezproblémový chod modelu. Reálné názvy jsou víceslovné a objevuje se v nich diakritika, což by mohlo při spuštění modelu dělat problémy. Dále byla změněna norma uvedení čísel u atributů C9 (průměrný

zůstatek) a C13 (mzda klienta). Konkrétně byl změněn oddělovač desetinných míst z tečky na čárku (př. 130 244.00 => 130 244,00). V praxi zabere příprava dat mnohdy nejvíce času, často je čerpáno z více zdrojů, proto je také potřeba jednotlivé atributy sjednotit a poté data čistit a hledat prázdné či chybné hodnoty. Tato úloha vychází z již předpřipravených dat, proto tato fáze nezabrala tolik času.

3.4 Modelování

Fáze modelování již obsahuje hledání a využití vhodných algoritmů, pomocí kterých model vyhodnotí důležitost jednotlivých atributů a vyhodnotí cíl. Tato datová sada obsahuje cílovou proměnnou, kde je již vyplněno, zda daný klient odešel či nikoli, využíváme tedy metodu *„učení s učitelem“* viz kapitola 1.3.

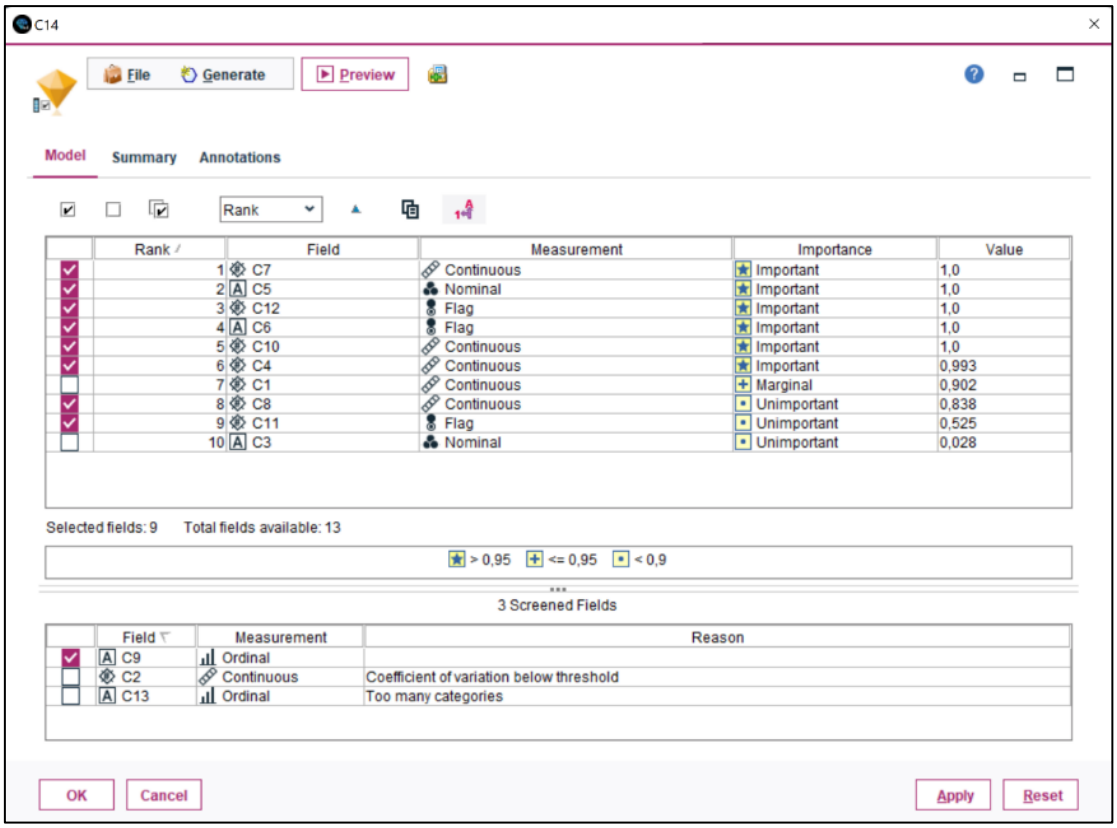
Před tím, než budou hledány vhodné algoritmy, je k uzlu Type vložena takzvaná *„cache“* (zelený čtvereček na uzlu viz obrázek číslo 15). Cache slouží ke zrychlení modelu neboli k jeho optimalizaci. Je napojena na uzel Type, protože v tomto uzlu se načítají data, se kterými se poté modeluje. Pokud by model pracoval pomalu již ve fázi přípravy dat, mohla být cache zapojena již dříve před tvorbou grafů.



Obrázek č. 15: Cache

Zdroj: Vlastní zpracování

Prvním modelovacím uzlem je Feature Selection. Po vložení uzlu a spuštění streamu se vygeneruje model (diamant). Tento model rozhodne o důležitosti jednotlivých atributů. Důležité, mezní a nedůležité atributy ponechává model v základní (horní tabulce), vyřazené atributy přesouvá do spodní tabulky i s důvodem vyřazení (viz obrázek číslo 16). To, zda atribut vstoupí do dalšího modelování určí zaškrtnutí pole po levé straně dialogového okna. Autorka tedy přizpůsobí zařazení atributů dle vlastního uvážení o důležitosti, s přihlédnutím na vyhodnocení atributů modelem.

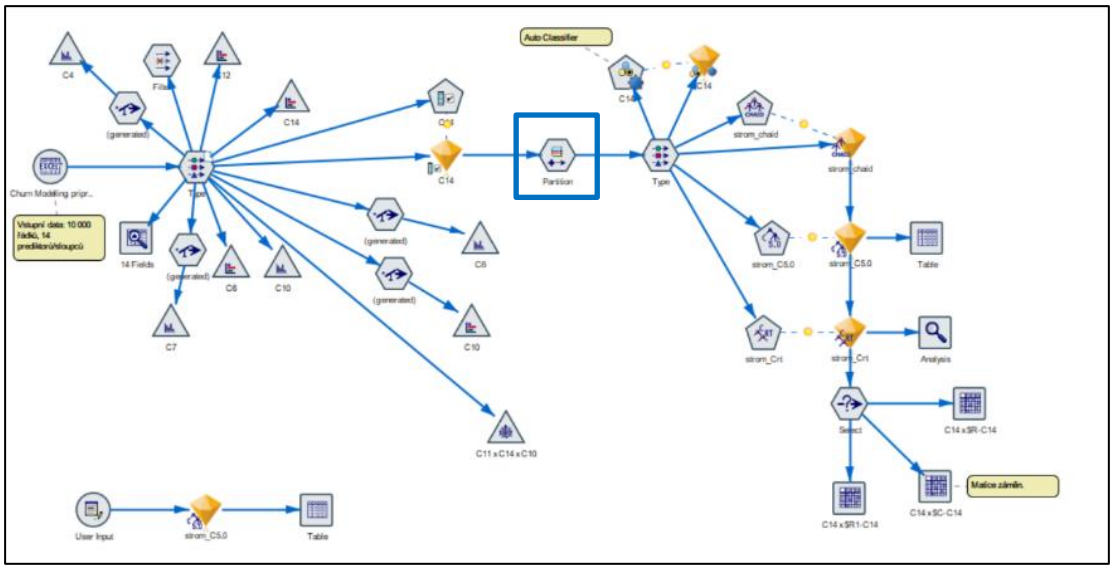


Obrázek č. 16: Model Feature Selection

Zdroj: Vlastní zpracování

V tuto chvíli je potřeba rozdělit datovou sadu na trénovací a testovací množinu. V rámci trénovací množiny zjišťuje model vztahy mezi cílovou proměnnou a relevantními prediktory. Takto získané znalosti poté trénuje na testovací množině, u které se pokusí stanovit cílovou proměnnou. Tuto stanovenou cílovou proměnnou poté porovná s reálnou (v datové sadě uvedenou) a vypočítá pravděpodobnost, se kterou dokáže vyhodnotit, zda klient odejde či nikoli.

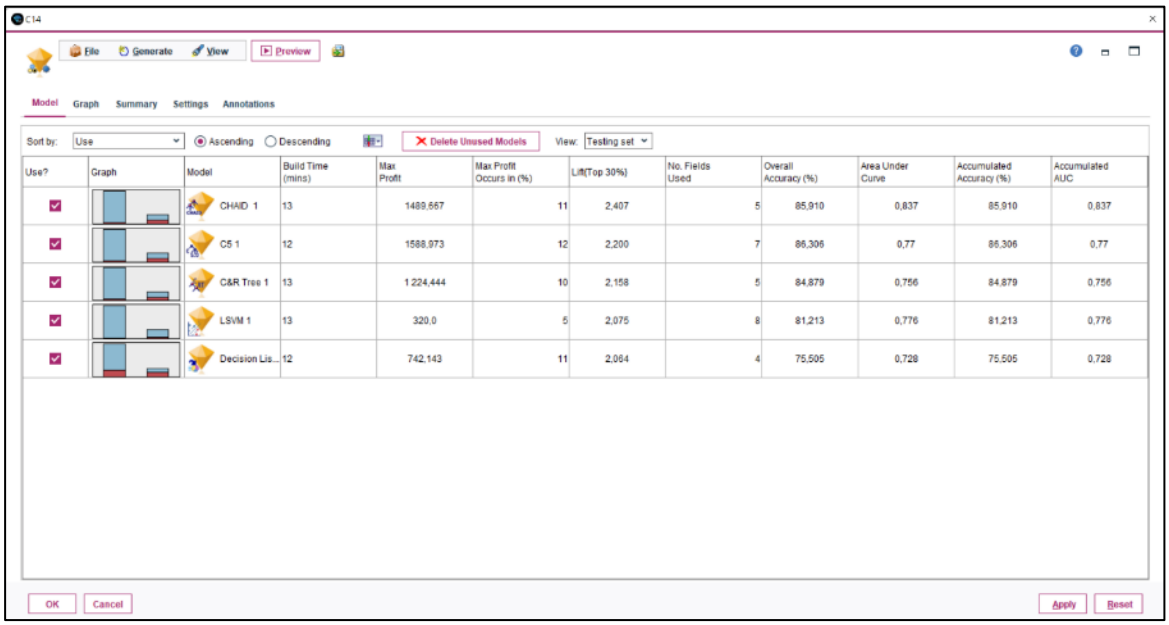
Datovou sadu lze rozdělit pomocí uzlu Partition viz obrázek číslo 17. Uvnitř uzlu je volen poměr rozdělení, zvolen je poměr 50 na 50. V uzlu Partition je možné datovou sadu rozdělit na celky trénovací, testovací a validační. Pro tuto úlohu, v této fázi je potřeba pouze rozdělit datovou sadu na training a testing, validace tedy není zahrnuta.



Obrázek č. 17: Uzel Partition

Zdroj: Vlastní zpracování

V této fázi jsou již stanovovány algoritmy, pomocí kterých bude možné danou úlohu řešit. Při tomto rozhodování je možné využít uzel Auto Classifier. Po spuštění streamu s tímto koncovým uzlem je vygenerován model (diamant), který obsahuje doporučené algoritmy viz obrázek číslo 18. Na prvních pozicích jsou vygenerovány algoritmy rozhodovacích stromů, konkrétně CHAID, C5 a C&R Tree. Stromy C5 a CHAID jsou obecné, mohou se tedy větvit do vícero jak dvou větví, C&R Tree (CART) je binární, větví se tedy pouze do větví dvou. Pro modelování jsou tedy zvoleny tyto tři rozhodovací stromy, z nichž bude vybrán ten s nejnižší chybovostí.

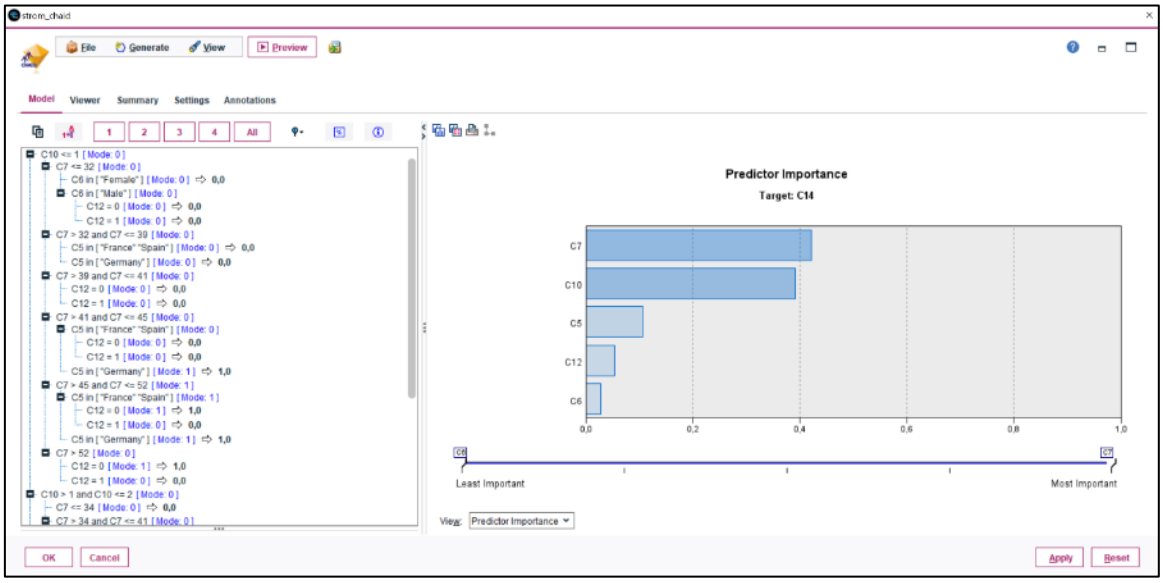


Obrázek č. 18: Model Auto Classifier

Zdroj: Vlastní zpracování

Prvním modelovaným stromem je CHAID. Ten je možné nalézt v záložce *„Modeling“*, uzel CHAID. Po zapojení uzlu do streamu a jeho spuštění je vygenerován samotný model rozhodovacího stromu se vzhledem diamantu. Po rozkliknutí modelu se zobrazí dialogové okno viz obrázek číslo 19. Ve sloupcovém grafu (na ose y) uvnitř modelu jsou prediktory, které model shledal důležité pro vyhodnocení cílové proměnné. Prediktory jsou řazené od nejdůležitějšího po nejméně důležitý. Poměr důležitosti zobrazují hodnoty na ose x.

Nejdůležitějšími prediktory podle algoritmu CHAID je C7, tedy věk klienty, v těsném závěsu C10 což je počet produktů. Ostatní prediktory nemají dle modelu tak silný vliv jako věk a počet produktů, nicméně do rozhodování modelu zasahují. Mezi tyto prediktory patří C5 (státní příslušnost), C12 (aktivní člen) a na posledním místě C6 znázorňující pohlaví klienta.



Obrázek č. 19: Rozhodovací strom CHAID

Zdroj: Vlastní zpracování

Na levé straně dialogového okna v modelu CHAID viz obrázek číslo 19 je vidět větvení stromu. Je možné nastavit fáze větvení 1-4, kořenem stromu je cílová proměnná C14 od které se strom postupně větví. V prví fázi jsou zobrazeny základní větve reprezentované prediktorem C10 (počet produktů), ten je větven do tří větví, C10<=1, C10>1 and C10<=2, C10>2. Z těchto větví poté *„rostou“* další větve a listy jako C7 (věk). C7 je druhou fází větvení a větví se pro každou skupinu C10 tímto způsobem: C7 (věk) <=32, C7>32 and C7<=39, C7>39 and C7<=41, C7>41 and C7<=45, C7>45 and C7<=52, C7>52. Ve třetí fázi větvení jsou pod jednotlivými kategoriemi větveny prediktory C6 (věk), C5 (státní příslušnost) a C12 (aktivní člen). V poslední fázi jsou mezi sebou větveny poslední tři prediktory.

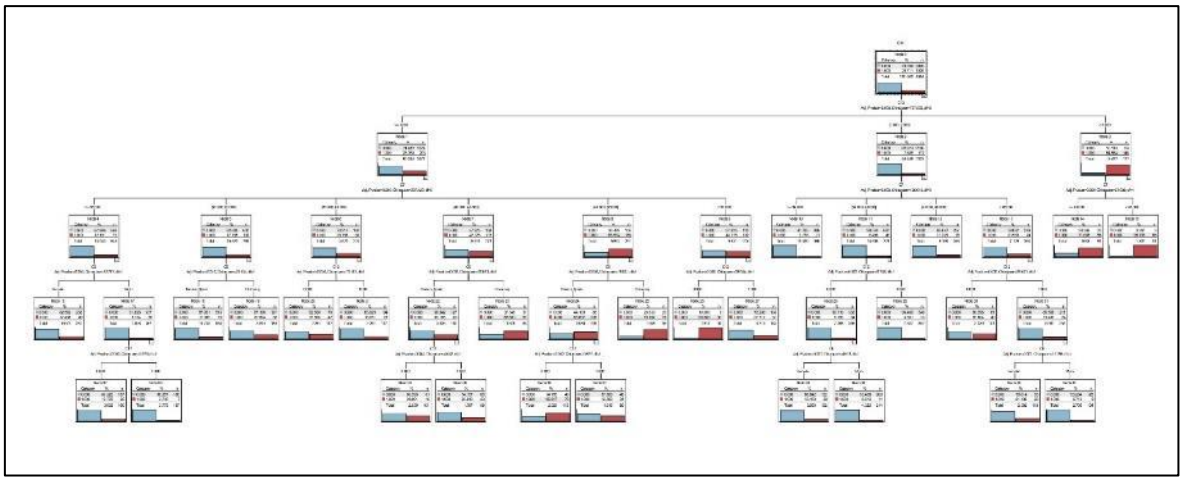
Pokud je větvení v obrázku číslo 19 rozebráno podrobně lze ho interpretovat následujícím způsobem: Pokud klient vlastní 1 či méně produktů a je mu 32 let nebo více a je to žena => NEODEJDE, pokud je to muž a je to aktivní nebo neaktivní člen => NEODEJDE. Pokud klient vlastní 1 či méně produktů a je starší 32 let nebo <= 39 let a je to Francouz, Španěl nebo Němec => NEODEJDE. Pokud klient vlastní 1 či méně produktů a je starší 39 let nebo <= 41 let a je aktivní nebo neaktivní klient => NEODEJDE. Pokud klient vlastní 1 či méně produktů a je starší 41 let nebo <= 45 let a je to Francouz nebo Španěl a je aktivní nebo neaktivní klient => NEODEJDE, pokud je to Němec => **ODEJDE**. Pokud klient vlastní 1 či méně produktů a je starší 45 let nebo <= 52 let a je to Francouz nebo Španěl a je neaktivní člen => **ODEJDE**, pokud je to aktivní člen => NEODEJDE, pokud je to Němec, nehledě na aktivitě => **ODEJDE**. Pokud klient vlastní 1 či méně produktů a je starší 52 let a zároveň je to neaktivní člen => **ODEJDE**, pokud je aktivním členem => NEODEJDE.

Pokud má klient více jak 1 produkt nebo <= 2 produkty a je mladší nebo starý 34 let => NEODEJDE. Pokud má více jak 1 produkt nebo <= 2 produkty a je starší 34 let nebo <=41 let zároveň je neaktivní člen => žena ani muž NEODEJDE, aktivní klient též neodejde. Klient s 1 produktem nebo <= 2 produkty starší 41 let nebo <=45 => NEODEJDE. Klient s 1 produktem nebo <= 2 produkty starší 45 let => NEODEJDE nehledě na tom, zda je aktivní/neaktivní člen, žena či muž.

Klient s více jak 2 produkty mladší nebo starý 41 let => **ODEJDE**. Klient s více jak 2 produkty starší než 41 let => **ODEJDE**.

Na základě podrobného rozboru větvení stromu CHAID je vidět, že často odchází Němci, dále neaktivní klienti starší 45 let s jedním produktem a klienti nad 41 let s více jak 2 produkty.

Pokud je v dialogovém okně rozhodovacího stromu CHAID viz obrázek číslo 19 zvolena záložka *„Viewer“* je zobrazeno stromové schéma reprezentované dendrogramem viz obrázek číslo 20. Jedná se o znázornění rozboru schématu (if-then), viz podmínky rozepsané výše, pouze v přehlednější vizuální formě. Na dendrogramu je vidět, že je CHAID stromem obecným, větví se do libovolného počtu větví.



Obrázek č. 20: Dendrogram stromu CHAID

Zdroj: Vlastní zpracování

Dalším aplikovaným algoritmem je strom C5.0. Náhled do modelu viz obrázek číslo 21. Tento algoritmus vyhodnotil důležitost dat lehce odlišně, a to jak poměry, tak počtem zohledněných prediktorů. Jako nejdůležitější vyhodnotil algoritmus atribut C10 (počet produktů) a to s 50% zastoupením. Na druhém místě je C7 (věk) pohybující se lehce nad 20 %. Další atributy jsou zastoupeny v menší míře, ale do rozhodování modelu jsou zařazeny a je jich více než u algoritmu CHAID. Jedná se o C5 (státní příslušnost), C6 (pohlaví), C12 (aktivní člen), **C11** (kreditní karta) a **C4** (kreditní skóre).

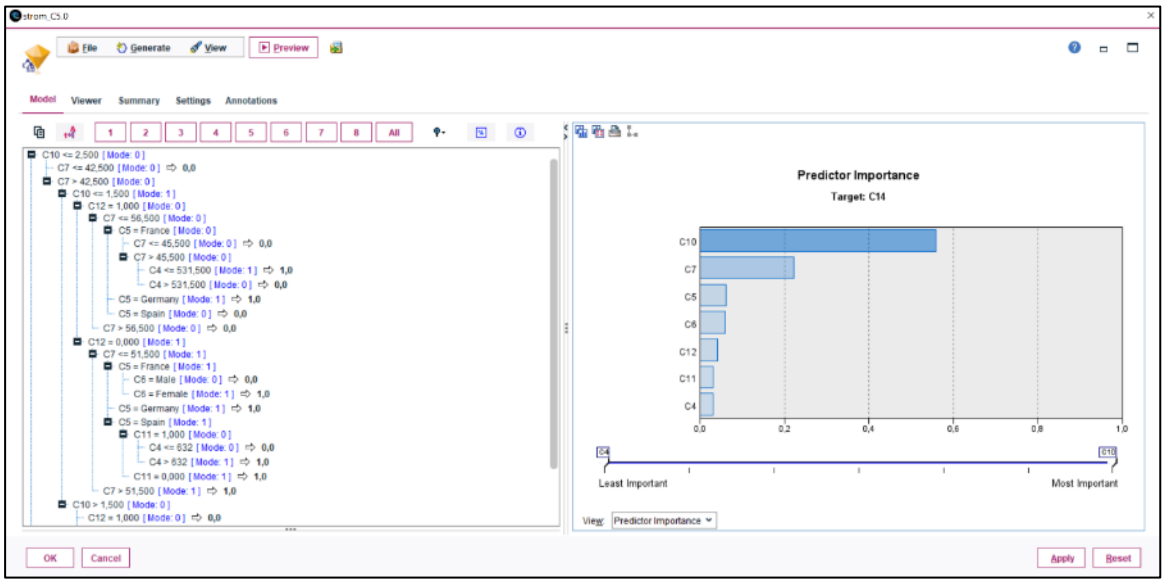
Již na první pohled je na levé straně dialogového okna viditelné rozdílné větvení stromu. C5.0 je stejně jako CHAID obecným stromem, větveným do libovolného počtu větví. Na obrázku číslo 22 je však viditelné, že dochází primárně k větvení po 2 větvích. Následkem toho není strom tak robustní (široký), je naopak delší, tato struktura zapříčiňuje to, že je model rychlejší.

U stromu C5.0 se autorka nezaměřuje na kompletní rozbor větvení, ale pouze na části, ve kterých algoritmus shledal, že klient odejde. Strom C5.0 má více fází, konkrétně 8, protože se větvil primárně po 2 větvích. Začíná u kořene C14, větví se přes C10 (počet produktů), C7 (věk), C12 (aktivní člen) a dále. Konkrétně je viditelné, že odcházejí tito klienti:

- Francouz s <=1,5 produkty ve věku <=45,5 let, aktivní člen a kreditním skóre <=531,5.
- Němec s <=1,5 produkty ve věku <=45,5 let, aktivní člen.
- Žena z Francie s <=1,5 produkty ve věku <=51,5 let, neaktivní člen.
- Němec s <=1,5 produkty ve věku <=51,5 let, neaktivní člen.
- Španěl vlastnící kreditní kartu s <=1,5 produkty ve věku <=51,5 let s kreditním skóre >632 a Španěl bez kreditní karty.
- Klient s <=1,5 produkty ve věku >51,5 let bez kreditní karty.

- Klient s více jak 2,5 produkty.

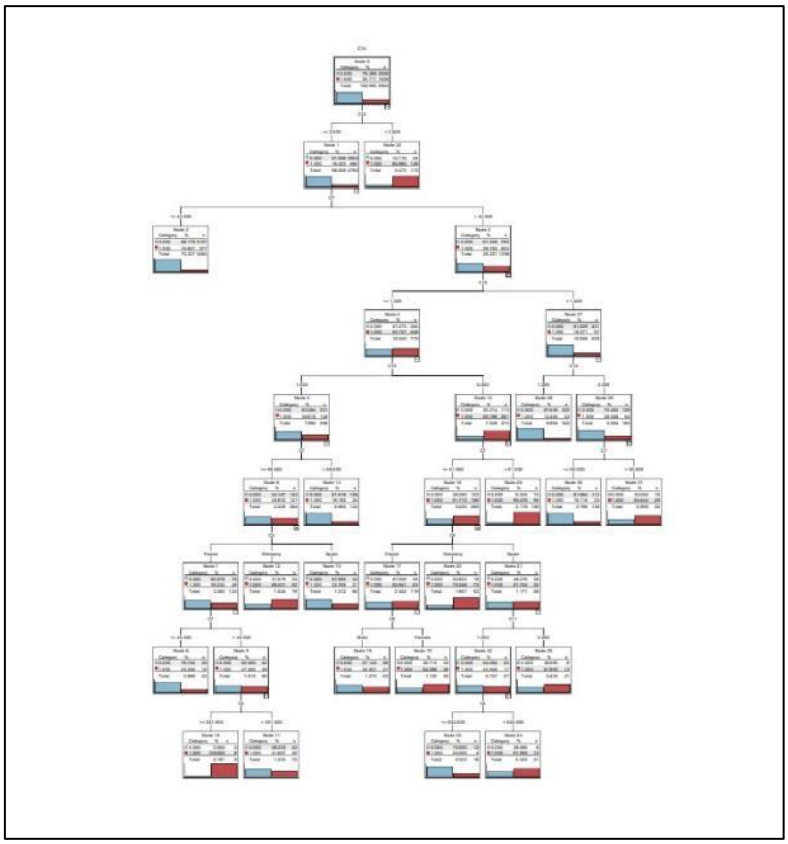
Z tohoto souhrnu je viditelné, že odcházejí hodně Němci tak jako u algoritmu CHAID, dále je však možné pozorovat další aspekty odchodů, díky dalším vstupujícím prediktorům, jako je například kreditní skóre. Je známé, že v průměru odcházejí z banky XY klienti ve 45 letech, pokud však přihlédneme na kreditní skóre, tak dle algoritmu C5.0 odcházejí klienti z Francie pod 45,5 let s horším kreditním skóre. To může poukazovat na mladší klienty, případně na klienty v horší finanční situaci. Francouzi běžně z banky neodcházejí v takové míře, nejvíce jsou jejich odchody dle algoritmu spojeny s kreditním skóre, to může být způsobeno různými faktory ovlivňujícími finanční situaci obyvatel ve Francii.



Obrázek č. 21: Rozhodovací strom C5.0

Zdroj: Vlastní zpracování

Dendrogram obecného stromu C5.0 je zobrazen na obrázku číslo 22. Modrou barvou jsou ve schématu zobrazeni klienti kteří neodešli, červeně ti co odešli.



Obrázek č. 22: Dendrogram stromu C5.0

Zdroj: Vlastní zpracování

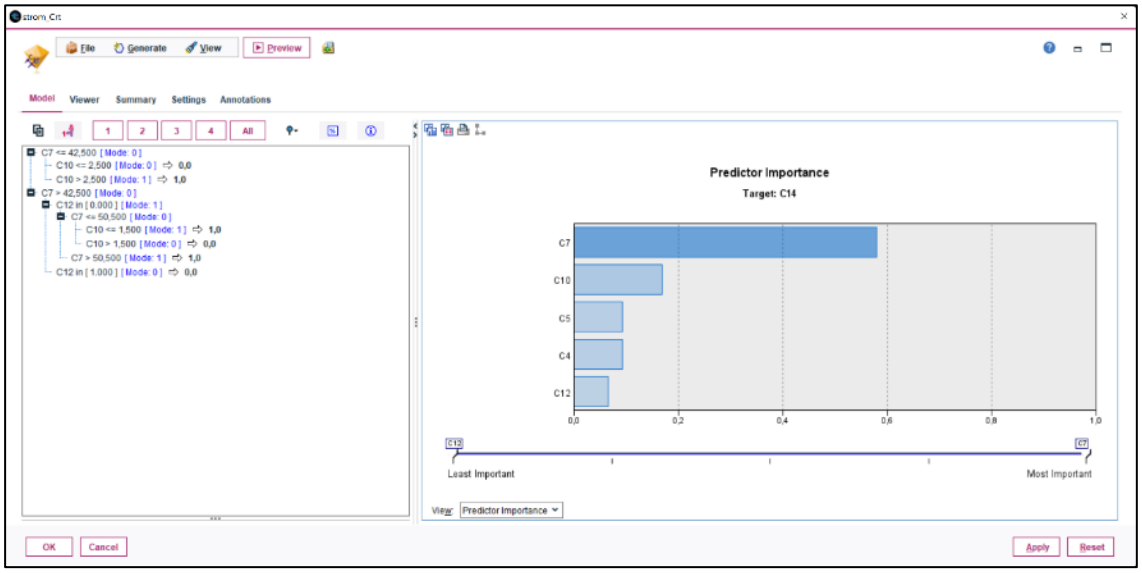
Posledním aplikovaným algoritmem strojového učení je C&RT (CART – Classification and Regression Tree), který jako jediný z aplikovaných algoritmů patří mezi binární stromy, tedy stromy větvící se maximálně do dvou větví. Po rozkliknutí vygenerovaného modelu (viz obrázek číslo 23) se zobrazí dialogové okno, s prediktory, které model vybral pro vyhodnocení výsledku. Pro algoritmus C&RT je nejdůležitějším prediktorem C7 (věk klienta) a to téměř z 60 %. Nižší zastoupení, pod 20 % mají C10 (počet produktů), C5 (státní příslušnost), C4 (kreditní skóre) a C12 (aktivní člen).

Na levé straně dialogového okna je vidět, že je větvení stromu méně rozsáhlé než u předchozích dvou a to do šířky i do výšky. Větvení má zde čtyři úrovně, první větve (v první fázi) jsou C7 (věk klienta) rozdělující klienty banky na <=42,5 a na starší jak 42,5. V druhé fázi se větví C10 (počet produktů) a C12 (aktivní člen). Ve třetí fázi se poté tyto tři atributy větví navzájem. Prediktory C5 (státní příslušnost) a C4 (kreditní skóre) nakonec ve schématu stromu zahrnuty nejsou.

Podrobný rozbor výsledků vyhodnocených algoritmem C&RT je následující: klient ve věku <= 42,5 let s <= 2,5 produkty => NEODEJDE, klient s > 2,5 produkty => **ODEJDE**. Neaktivní klient ve věku <= 50,5 let a s <= 1,5 produkty => **ODEJDE** a klient s > 1,5 produkty =>

NEODEJDE. Neaktivní klient ve věku vyšším jak 50,5 let => **ODEJDE.** Aktivní klient starší 42,5 let => NEODEJDE.

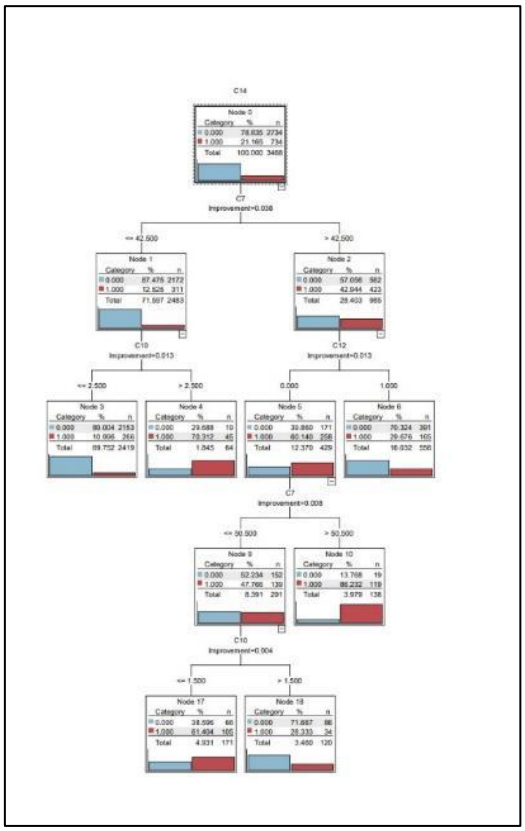
Pokud je tento rozbor shrnut, vychází z něj, že odchází klienti do 42,5 let co mají více jak 2,5 produktů. Naopak, podle algoritmu C&RT, klienti do 50,5 let odchází, pokud mají méně produktů, a to konkrétně do 1,5 produktu, ale musí jít o neaktivního klienta. Do této skupiny budou nejspíš patřit primárně klienti, co si zařídili produkt s bonusem a dále produkt nevyužívají, protože mají jinou banku, kde jsou aktivní. Dále z banky celkově odchází ve větší míře neaktivní klienti nad 50,5 let, nehledě na to, zda jsou aktivní či nikoli.



Obrázek č. 23: Rozhodovací strom C&RT

Zdroj: Vlastní zpracování

Na dendrogramu stromu C&RT (viz obrázek číslo 24) lze vidět jednoduchost modelu oproti předchozím dvěma a zároveň větvení maximálně do dvou větví (binární strom). Algoritmus C&RT ke svému rozhodnutí využil nejnižší počet prediktorů. I přes to, že vyhodnotil pět prediktorů za důležité, při tvorbě podmínek využil prediktory pouze tři.



Obrázek č. 24: Dendrogram stromu C&RT

Zdroj: Vlastní zpracování

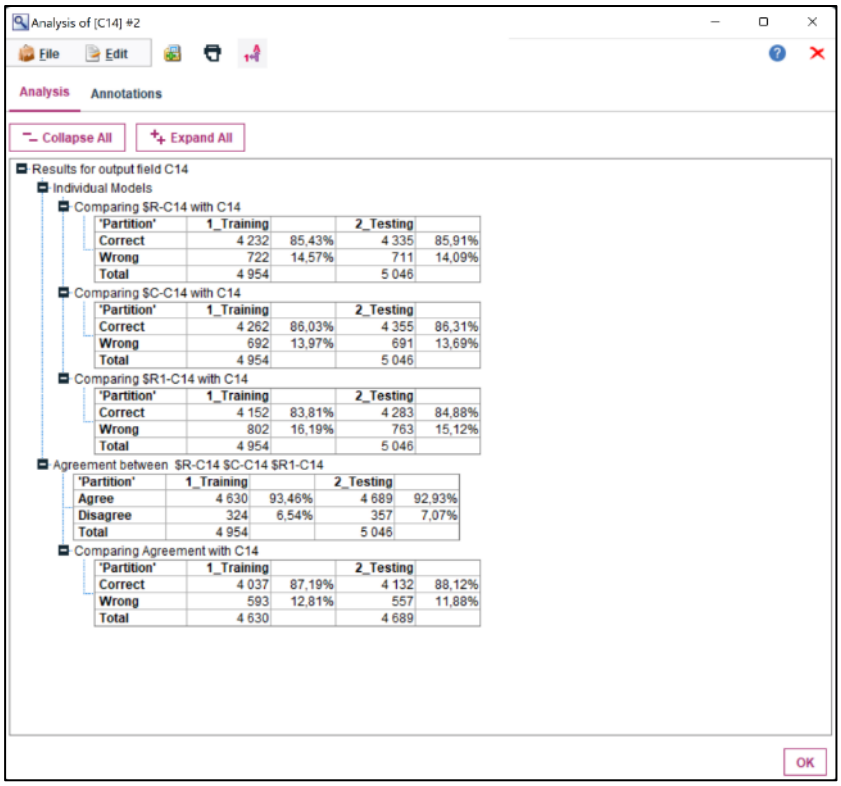
V tuto chvíli jsou sestaveny všechny tři modely, které je nyní potřeba vyhodnotit. U vyhodnocování modelů se sleduje několik parametrů. Konkrétně se sleduje to, kolik udělal chyb a kolik výsledků vyhodnotil správně a zároveň s jakou procentní pravděpodobností je schopný výsledek vyhodnotit správně.

3.5 Evaluace

Pokud by se jednalo o úlohu řešenou v praxi obsahovala by evaluace také konzultaci s managementem. V této fázi může být zjištěno, že je potřeba upravit některé podmínky, či model penalizovat za určitý typ chyby, aby se tak tyto chyby minimalizovali.

Nyní přichází srovnání již vytvořených modelů, ke kterému bude využit uzel *„Analysis“* v záložce *„Output“*. V tomto uzlu je možné vidět, jak jednotlivé algoritmy chybovaly v rámci trénovací a testovací množiny (viz obrázek číslo 25). Je zde konkrétně uveden počet chyb a počet správných odpovědí, obě hodnoty jsou vyjádřeny také procentuálně. První tabulka patří algoritmu CHAID, druhá C5.0 a třetí C&RT. Z tabulek je vidět, že nejhůře dopadl algoritmus C&RT s největším počtem chyb a nejnižším počtem správných odpovědí jak na trénovací, tak na testovací množině. Naopak nejlépe dopadl algoritmus **C5.0 s přesností**

86,31 % na testovací množině. Nicméně všechny algoritmy rozhodují s úspěšností vyšší jak 80 % a rozdíly mezi nimi nejsou příliš velké.



Obrázek č. 25: Uzel Analysis

Zdroj: Vlastní zpracování

Dalším uzlem pro vyhodnocení kvality modelu jsou matice záměn neboli *„Matrix“*, který porovná reálnou hodnotu cílové proměnné s hodnotou, kterou určil model. Matice záměn tedy zobrazí počet případů, kdy model rozhodl, že klient z banky odejde a on neodešel a naopak (viz tabulka číslo 3 pro CHAID, 4 pro C5.0 a 5 pro C&RT).

V matici záměn však již stačí zobrazit data z testovací množiny, k výběru těchto dat je využit uzel *„Select“* ze záložky *„Record Ops“*. Podmínka pro zobrazení dat je následující (Partition = "2_Testing"), autorka totiž chce, aby uzel Partition, který vytvořil v datové matici nový sloupec s hodnotami 1_Treining a 2_Testing nyní obsahoval pouze hodnoty 2_Testing. Po vytvoření uzlu Select jsou tedy napojeny tři matice záměn viz tabulky 3, 4 a 5.

Tabulka 3: Matice záměn pro CHAID

C14	0.0	1.0
0.0	3886	149
1.0	562	449

Zdroj: Vlastní zpracování

Tabulka 4: Matice záměn pro C5.0

C14	0.0	1.0
0.0	3879	156
1.0	535	476

Zdroj: Vlastní zpracování

Tabulka 5: Matice záměn pro C&RT

C14	0.0	1.0
0.0	3871	164
1.0	599	412

Zdroj: Vlastní zpracování

Algoritmus CHAID, viz tabulka 3, vyhodnotil správně odchod 449 klientů a 149 vyhodnotil špatně. Co se týče zůstávajících klientů, vyhodnotil správně, že zůstane 3 886 klientů a špatně vyhodnotil, že zůstane 562.

Algoritmus C5.0, viz tabulka 4, vyhodnotil správně odchod 476 klientů a 156 vyhodnotil špatně. Co se týče zůstávajících klientů, vyhodnotil správně, že zůstane 3 879 klientů a špatně vyhodnotil, že zůstane 535.

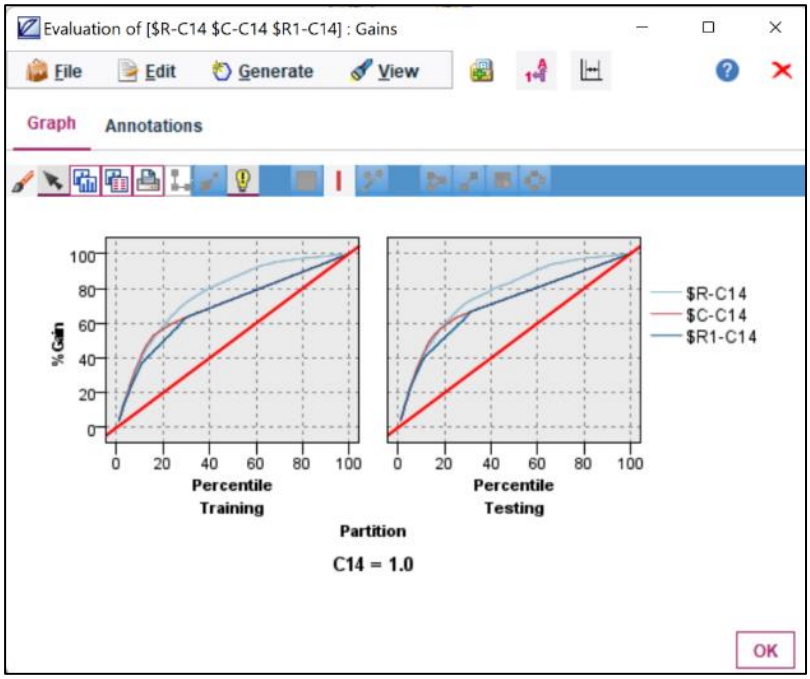
Algoritmus C&RT, viz tabulka 5, vyhodnotil správně odchod 412 klientů a 164 vyhodnotil špatně. Co se týče zůstávajících klientů, vyhodnotil správně, že zůstane 3 871 klientů a špatně vyhodnotil, že zůstane 562.

Nejvíce správných zařazení klientů, kteří **neodejdou** tedy uskutečnil algoritmus CHAID (3 886 klientů), ale nejméně chyb zde udělal C5.0 (535 chybně zařazených klientů). Pokud budou porovnány správné a chybné odpovědi u obou algoritmů, bude zjištěno, že CHAID pracoval se 14% chybovostí a C5.0 se 13% chybovostí. Lépe tedy vyhodnocoval **C5.0**.

Při zařazování klientů, kteří **odejdou** byl nejúspěšnější C5.0 s 476 správně zařazenými ale na počet chybných zařazení, byl naopak lepší CHAID se 149 chybami. Algoritmus C5.0 tedy rozhodl správně 476krát a udělal 156 chyb, CHAID rozhodl 449krát správně a 149krát špatně. V poměru správných a špatných hodnot, tedy lépe pracoval algoritmus **C5.0**. Podle matic záměn (Matrix) i uzlu Analysis by tedy autorka zvolila algoritmus C5.0, který vychází nejlépe.

Další alternativou pro hodnocení modelu je uzel *„Evaluation" (viz obrázek 26),* který lze nalézt mezi grafy. Graf *„Evaluation"* znázorňuje, jak vysokou důvěru má model ve své predikce (Anon. 2021a). Světle modrá křivka znázorňuje algoritmus CHAID, fialová algoritmus C5.0 a poslední tmavě modrá algoritmus C&RT přičemž čím dál je křivka od červené přímky, tím má model ve

své predikce důvěru vyšší. Z grafu je tedy viditelné, že míra důvěry je u všech modelů poměrně podobná, a to jak na trénovací, tak na testovací množině. Nejvyšší důvěru ve své predikce má však algoritmus **CHAID**, na druhé příčce je C5.0 a pouze s malou odchylkou od C5.0 je na poslední příčce C&RT.



Obrázek č. 26: Uzel Evaluation

Zdroj: Vlastní zpracování

Na základě vyhodnocených výsledků v rámci Evaluace zvolila autorka model C5.0 pro nasazení do praxe.

3.6 Nasazení modelu do praxe

V této fázi je model již hotový a zbývá rozhodnout, jak bude využit. Jednou z možností je sepsání závěrečné zprávy, ve které budou shrnuty výsledky, tedy faktory ovlivňující odchody klientů. V této zprávě by byli klienti rozděleni do skupin pode jednotlivých atributů, důležitých pro danou úlohu (věk, počet produktů, státní příslušnost atd.). Na základě této zprávy by bylo možné zacílit, například reklamou či zvýhodněným produktem pro klienty, kteří patří do skupin s pravděpodobným odchodem.

Další možností, jak model využít je jeho automatizace ve které by byl využit také prediktivní potenciál modelu. V praxi by byl model nejspíše napojen na databázi. První uzel na obrázku číslo 27 *„Database“ ze záložky „Sources“* určuje odkud jsou data čerpána. Po rozkliknutí uzlu „Database“ je dohledána cesta ke konkrétní databázi, ve které se zdroj aktualizuje. Do modelu jsou tedy nahrávána stále aktuální data. Vstupní data jsou napojena na již trénovaný

a testovaný model C5.0 (viz obrázek číslo 21), který byl v kapitole 3.5 vyhodnocen, jako nejlepší (fungující s nejvyšší pravděpodobností). Následně jsou výsledky vyhodnocené modelem zapisovány do další databáze. Závěrečný uzel *„Database“* se nachází v záložce *„Export“* a po jeho rozkliknutí je též potřeba nadefinovat cestu přímo do databáze, kam chce autor modelu data ukládat.

Obrázek č. 27: Nasazení modelu do praxe

Zdroj: Vlastní zpracování

Takto shromažďované výstupy v databázi je poté možné dále analyzovat a spojovat s dalšími databázemi. Z nashromážděných dat je například možné vizualizovat vývoj odchodů klientů, během jednotlivých let či jiných časových období a analyzovat tak chování zákazníků. Z těchto behaviorálních modelů/analýz mohou vznikat další poznatky, které mohou být do modelu C5.0 implementovány a mohou tak zvýšit jeho přesnost a užitečnost. Na základě těchto poznatků by banka XY mohla například sestavit model pro personalizaci obsahu. Pokud banka bude znát preference klientů a důvody kvůli kterým klienti odcházejí a zároveň klienty rozdělí do skupin se shodnými vlastnostmi, může například v rámci reklam v mobilním bankovnictví, doporučovat klientům produkty přímo jim na míru. Tyto činnosti však musí banka provádět pouze do určité míry, s přihlédnutím na ochranu osobních údajů klientů.

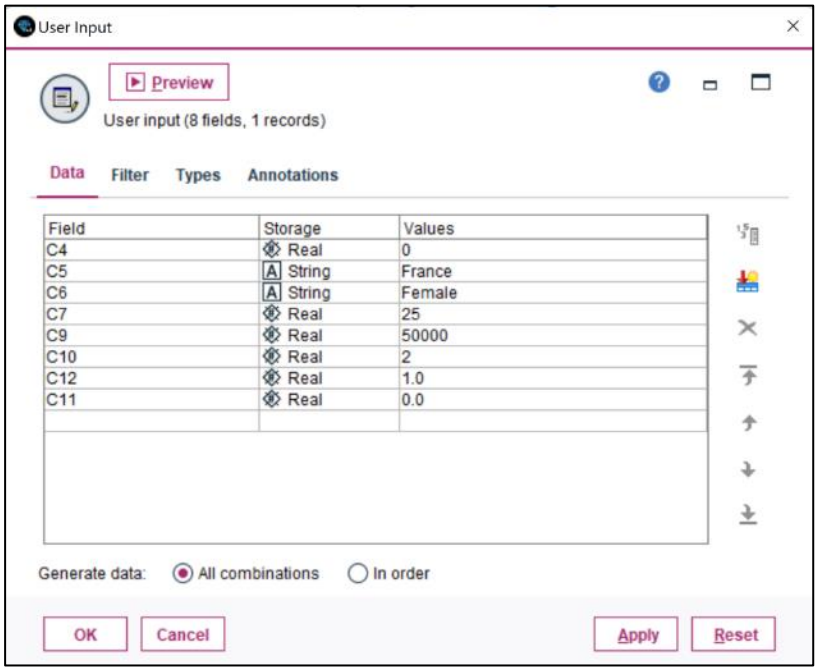
Při řešení úlohy odchodů klientů z banky však autorka nenahrávala data přes databázi. V programu IBM SPSS Modeler existuje uzel, který umožňuje ruční vepsání vlastností klienta na základě, nichž model vyhodnotí výsledek. Tento uzel se nachází v záložce *„Sources“* a nazývá se *„User Input“*. Za *„User Input“* je napojen model C5.0 a za něj uzel reprezentující formu výstupu viz obrázek číslo 28.

Obrázek č. 28: Stream naučeného modelu

Zdroj: Vlastní zpracování

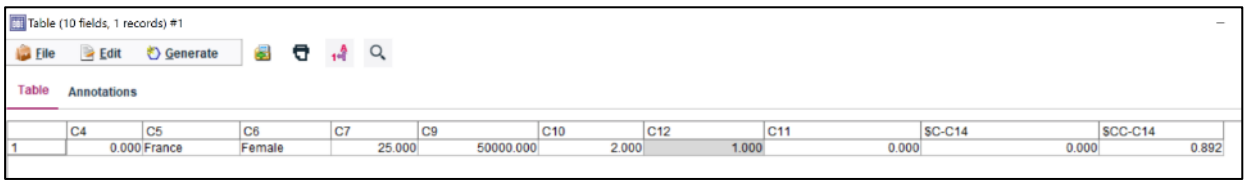
Po rozkliknutí uzlu *„User Input“* se zobrazí dialogové okno viz obrázek číslo 29. V tomto okně jsou ve sloupci *„Field“* zobrazeny prediktory vstupující do modelu C5.0, ve druhém sloupci jsou

jejich datové typy a do posledního sloupce *„Values“* je vpisován konkrétní klient. Pro tento příklad byla vybrána žena z Francie ve věku 25 let s kreditním skóre nula, zůstatkem na účtu 50 000, se dvěma produkty, aktivní klient bez kreditní karty. Po vyplnění hodnot je potvrzen výběr. Jako uzel pro export je zvolen uzel *„Table“*. Jedná se o datovou matici, která zobrazí vložená (námi specifikovaná data) a výsledek uvnitř programu SPSS Modeler. Tento výsledek totiž není potřeba nikam exportovat, je vyžadován pouze náhled viz obrázek číslo 30. Výsledkem, který model C5.0 vyhodnotil je, že by daná klientky z banky 89,2% pravděpodobností neodešla.



Obrázek č. 29: Uzel User Input

Zdroj: Vlastní zpracování



Obrázek č. 30: Uzel Table

Zdroj: Vlastní zpracování

4 Vyhodnocení poznatků z praktické části

Praktická část byla sestavována na základě metodologie CRISP-DM, která určovala postup při tvorbě celého streamu v programu IBM SPSS Modeler. První částí metodologie bylo porozumění samotnému problému. Tato fáze byla obtížnější bez konzultací specialisty či managementu uvnitř podniku. Úlohou managementu v této fázi je totiž stanovit cíle ke kterým je směřováno a zároveň popsat, na jaké faktory je dobré se při tvorbě modelu zaměřit. Model má primárně vyhodnotit klienty kteří odejdou, ale k čemu tato vyhodnocení povedou a jaký budou mít pro danou společnost přesah, to je možné se v tomto případě pouze domnívat.

Druhou fází v metodologii bylo porozumění datům. V datové sadě se nacházelo 14 atributů, včetně cílové proměnné, a jednalo se pouze o jednu datovou sadu. V porozumění samotným datům tedy nevznikl výrazný problém nicméně, pokud by tato úloha byla řešena v praxi, bylo by k dispozici více datových sad obsahujících třeba i desítky atributů. V tom případě by bylo porozumění o dost složitější a byly by znovu potřeba konzultace se specialisty na dané odvětví či managementu. Na druhou stranu by s nejvyšší pravděpodobností bylo možné z těchto dat vybrat i relevantnější atributy, než byly obsaženy v této použité sadě a tím by bylo možné zvýšit přesnost modelu i jeho důvěru v učiněná rozhodnutí.

Třetí fáze, příprava dat. Tato fáze by v praxi zabrala možná nejvíce času. Je předpokládáno, že by se vycházelo z několika datových sad uložených na různých místech datového skladu či jezera. V tomto případě by bylo potřeba data najít a poté jednotlivé databáze propojit a provést nad nimi takové operace, které by umožnily vygenerování požadované datové sady. Přičemž data by bylo potřeba vyčistit o duplicity, doplnit prázdné hodnoty či odhalit a redukovat chyby. V datové sadě přichystané pro řešení praktické části diplomové práce nebylo potřeba tyto záležitosti řešit, sada byla předpřipravena, a to bez prázdných hodnot i duplicit. Jediným zádrhelem ve zpracování byl špatný oddělovač desetinných míst, který bylo potřeba ve zdrojových datech nahradit čárkou místo tečky.

Následující fází bylo samotné modelování a evaluace. Modelování se soustředilo primárně na práci se třemi vybranými algoritmy strojového učení, rozhodovacími stromy. Konkrétně se jednalo o rozhodovací stromy CHAID, C5.0 a C&RT. Na vizualizacích stromů bylo viditelné, že CHAID a C5.0 jsou oproti C&RT robustnější a obsahují více větvení (tedy zohledňují více prediktorů). Rozhodovací strom C&RT v konečném výpočtu využil pouze tři prediktory, což se odrazilo na kvalitě jeho výsledků (oproti zbývajícím dvěma algoritmům). Algoritmy CHAID a C5.0 se v kvalitě výstupů poměrně střídaly. CHAID získal převahu na základě vyhodnocení grafu *„Evaluation“*, který shledal CHAID jako algoritmus, který si je svými zařazeními nejjistější.

Algoritmu C5.0 zase vyšly lepší výsledky při vyhodnocení matic záměn a při celkové pravděpodobnosti správného rozhodnutí.

Pokud budou rozebrány všechny relevantní prediktory, tak nejčastěji odcházeli klienti z Německa (v porovnání s ostatními státními příslušnostmi), klienti s kreditním skóre pod 405, více odcházely ženy než muži a věk ve kterém klienti nejčastěji odcházeli je 55 let. Co se týče funkčního období klientů, odcházeli nejčastěji klienti působící v bance 10 let a více, dále více odchází klienti bez kreditní karty, neaktivní klienti a klienti se 3 nebo 4 produkty (nejspíše po ukončení fixace hypotéky).

Posledním bodem v postupu, určovaným metodologií CRISP-DM, bylo nasazení modelu do praxe. V rámci této úlohy je nasazením do praxe myšleno vytvoření nového streamu s již naučeným modelem. Pro tento stream byl užit uzel *„User Input“*, který umožňuje vložit informace o klientovi, který je poté vyhodnocen naučeným modelem, jako odchozí či zůstávající. Uzlem *„User Input“* lze tedy do modelu ručně nahrávat nová data. V praxi byl model napojen přímo na firemní databázi.

5 Závěr

V současné době jsou již BI nástroje, včetně data miningu, ve středních a velkých podnicích standardem. Je však stále velké množství firem, které BI nástroje nevyužívají úplně naplno a stále se spoléhají na ověřené postupy, které však nemusí být tak efektivní. Pokud je na trhu konkurence, je potřeba udržovat tempo s dobou a implementovat BI řešení ve firmě stále komplexněji. Data mining je metodou, která může firmě pomoci s lepším cílením na zákazníky například v rámci analýzy nákupního koše či při segmentaci trhu. Může však firmu také ochránit před nechtěnými zákazníky, kteří například nebudou schopni splácet své závazky, a to pomocí skóringových modelů. U současných klientů může pomoci při detekci podvodů, například pomocí analýzy shluků.

Co je ale nejdůležitější, může napomoci k posunu lidstva například v rámci detekce závažných onemocnění na základě určitých příznaků, či při doporučování potřebné léčby pro pacienta. V současné době jsme na počátku nového věku informačních technologií, a to díky naučeným modelům, které pohání umělou inteligenci. V této době nastává boom umělé inteligence, která zvládá velké množství činností, se kterými se člověk ve svém životě setkává. Rozvoj těchto modelů však s sebou nese velké množství úskalí, která bude potřeba v průběhu nadcházejících let řešit. Ať už je to využívání umění vázaného duševním vlastnictvím do koláží umělé inteligence či naučený model, který za studenta napíše semestrální práci. I přes tato úskalí jsou modely data miningu a strojového učení něčím, co posune fungování lidstva o několik kroků vpřed.

Cílem diplomové práce bylo sestavit model, který vyhodnotí, jací klienti z banky XY s určitou pravděpodobností odejdou a jací nikoli.

Praktická část byla sestavována na principu metodologie CRISP-DM, která určovala postup řešení případu. V prvních fázích bylo nutné nejprve porozumět řešenému problému a datům. Třetí fáze zahrnovala přípravu dat, tedy například formátování a poté již bylo možné přejít k samotné modelaci. V první fázi modelace byly odebrány nepotřebné prediktory, pro rychlejší práci modelu, poté byla množina dat rozdělena na trénovací a testovací a poté byly zvoleny algoritmy pro řešení úlohy. Autorka zvolila tři algoritmy strojového učení, rozhodovací stromy C5.0, CHAID a C&RT. Tyto algoritmy byly napojeny, spuštěny a v rámci čtvrtého kroku CRISP-DM ohodnoceny. Nejlépe pracovaly algoritmy CHAID a C5.0, nakonec byl zvolen algoritmus C5.0 kvůli vyšší úspěšnosti při rozhodování. Tento algoritmus byl v rámci posledního kroku metodologie zasazen do nového streamu. Po načtení nového vstupu do streamu s naučeným modelem dokázal model vyhodnotit, zda zadaný klient odejde či nikoli a s jakou pravděpodobností se tak stane.

Seznam použité literatury

Anon., 2002. Aktivní klient ocení jinou banku nežli občasný uživatel. *iDNES.cz* [online] [vid. 2023-03-17]. Dostupné z: https://www.idnes.cz/finance/financni-radce/aktivni-klient-oceni-jinou-banku-nezli-obcasny-uzivatel.A020318_150827_fi_blind_mir

Anon., 2015. Data mining při analýze nákupního košíku. *Kurzy, konzultace, návody* [online]. [vid. 2023-02-12]. Dostupné z: https://exceltown.com/navody/postupy-a-spinave-triky/data-mining/data-mining-pri-analyze-nakupniho-kosiku/

Anon., 2021a. *IBM Documentation* [online] [vid. 2023-01-01]. Dostupné z: https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/es/spss-modeler/18.0.0?topic=trees-c50-node

Anon., 2021b. *POJISTNÉ PODVODY 2020: Nejčastěji se lidé snaží o podvod v pojištění vozidel, stoupají podvody v pojištění odpovědnosti* [online] [vid. 2023-02-19]. Dostupné z: https://www.cap.cz/tiskove-centrum/tiskove-zpravy/104794-pojistne-podvody-2020-nejcasteji-se-lide-snazi-o-podvod-v-pojisteni-vozidel-stoupaji-podvody-v-pojisteni-odpovednosti

Anon., 2022a. *Aktuální (2023) i historický vývoj úrokových sazeb hypoték | hyponamiru.cz* [online] [vid. 2023-02-22]. Dostupné z: https://www.hyponamiru.cz/aktualni-i-historicky-vyvoj-urokovych-sazeb-hypotek/

Anon., 2022b. *Data Mining vs Text Mining vs Web Mining: 3 kritické rozdíly - Naučte se | Hevo* [online]. [vid. 2023-04-14]. Dostupné z: https://hevodata.com/learn/data-mining-vs-text-mining-vs-web-mining/

Anon., [b.r.]. *03_rozhodovaci_stromy.pdf* [online]. [vid. 2023a-01-01]. Dostupné z: https://is.muni.cz/el/1431/jaro2008/Bi7490/um/03_rozhodovaci_stromy.pdf

Anon., [b.r.]. A history and timeline of big data. *WhatIs.com* [online] [vid. 2022b-11-24]. Dostupné z: https://www.techtarget.com/whatis/feature/A-history-and-timeline-of-big-data

Anon., [b.r.]. *Co je business intelligence | Microsoft Power BI* [online] [vid. 2022c-12-11]. Dostupné z: https://powerbi.microsoft.com/cs-cz/what-is-business-intelligence/

Anon., [b.r.]. *Co je datové jezero? | Microsoft Azure* [online] [vid. 2022d-12-26]. Dostupné z: https://azure.microsoft.com/cs-cz/resources/cloud-computing-dictionary/what-is-a-data-lake/

Anon., [b.r.]. Co je integrace podnikových aplikací (eai)? - definice z techopedie - Podnik 2022. *Icy Science* [online] [vid. 2022e-12-23]. Dostupné z: https://cs.theastrologypage.com/enterprise-application-integration

Anon., [b.r.]. Co je to datový sklad? | Definice, komponenty, architektura | SAP Insights. *SAP* [online] [vid. 2022f-12-23]. Dostupné z: https://www.sap.com/cz/insights/what-is-a-data-warehouse.html

Anon., [b.r.]. *Co jsou big data? | Oracle Česká Republika* [online] [vid. 2022g-11-12]. Dostupné z: https://www.oracle.com/cz/big-data/what-is-big-data/

Anon., [b.r.]. *ČNB zahájila kroky k odejmutí licence Sberbank CZ - Česká národní banka* [online] [vid. 2023h-03-23]. Dostupné z: https://www.cnb.cz/cs/cnb-news/tiskove-zpravy/CNB-zahajila-kroky-k-odejmuti-licence-Sberbank-CZ/

Anon., [b.r.]. *Data mining v bankách* [online] [vid. 2023i-02-12]. Dostupné z: https://www.systemonline.cz/business-intelligence/data-mining-v-bankach.htm

Anon., [b.r.]. *Evoluční algoritmy a princip jejich fungování | NaPočítači.cz* [online] [vid. 2023j-01-02]. Dostupné z: https://www.napocitaci.cz/33/evolucni-algoritmy-a-princip-jejich-fungovani-uniqueidgOkE4NvrWuNY54vrLeM674MW0OH42R01Ag_rzFJ8D5c/?query=genetick%E9%20algoritmy&serp=1

Anon., [b.r.]. Jaká je standardní doba fixace u hypotéky. *MONETA Money Bank* [online] [vid. 2023k-03-12]. Dostupné z: https://www.moneta.cz/detail-otazky-a-odpovedi

Anon., [b.r.]. *Kaggle: Your Machine Learning and Data Science Community* [online] [vid. 2023l-03-02]. Dostupné z: https://www.kaggle.com/

Anon., [b.r.]. *Kreditní skóre: Vše, co potřebujete vědět* [online]. [vid. 2023m-03-02]. Dostupné z: https://technoglitz.com/czechrepublic/kreditni-skore-vse-co-potrebujete-vedet/

Anon., [b.r.]. *M505 : Metodika CRISP–DM* [online] [vid. 2023n-01-04]. Dostupné z: https://mbi.vse.cz/public/cs/obj/METHOD-113

Anon., [b.r.]. *Neuronové sítě a princip jejich fungování | NaPočítači.cz* [online] [vid. 2023o-01-01]. Dostupné z: https://www.napocitaci.cz/33/neuronove-site-a-princip-jejich-fungovani-uniqueidgOkE4NvrWuNY54vrLeM670eFNQh552VdDDulZX7UDBY/

Anon., [b.r.]. *[PDF] Download Techniques, Process, and Enterprise Solutions of Business - Free Download PDF* [online] [vid. 2022p-12-30]. Dostupné z: https://hugepdf.com/download/download-techniques-process-and-enterprise-solutions-of-business_pdf

Anon., [b.r.]. Rozdíl mezi seskupováním a klasifikací Porovnejte rozdíl mezi podobnými podmínkami - Technologie - 2023. *strephonsays* [online] [vid. 2023q-01-04]. Dostupné z: https://cs.strephonsays.com/clustering-and-vs-classification-13022

Anon., [b.r.]. *Rozhodovací stromy a chytré otázky – Základy informatiky pro střední školy* [online] [vid. 2023r-01-01]. Dostupné z: https://popelka.ms.mff.cuni.cz/~lessner/mw/index.php/U%C4%8Debnice/Informace/Rozhodovac%C3%AD_stromy_a_chytr%C3%A9_ot%C3%A1zky

Anon., [b.r.]. *Stručná historie systémů pro podporu rozhodování* [online] [vid. 2022s-12-11]. Dostupné z: https://dssresources.com/history/dsshistory.html

Anon., [b.r.]. Stručný návod k ovládání IBM SPSS Statistics a IBM SPSS Modeler.

Anon., [b.r.]. Tajemství biometrie 2: Rozpoznávání obličeje. *Ábíčko.cz* [online] [vid. 2023u-04-14]. Dostupné z: https://www.abicko.cz/clanek/precti-si-technika/23285/tajemstvi-biometrie-2-rozpoznavani-obliceje.html

Anon., [b.r.]. *What is a data warehouse? The source of business intelligence - ProQuest* [online] [vid. 2022v-12-26]. Dostupné

z: https://www.proquest.com/docview/2563937034/3297200449174CBAPQ/3?accountid=17116

Anon., [b.r.]. *What is OLAP? | IBM* [online] [vid. 2022w-12-30]. Dostupné z: https://www.ibm.com/topics/olap

AZAD, Hiteshwar Kumar a Kumar ABHISHEK, 2014. Semantic-Synaptic Web Mining: A Novel Model for Improving the Web Mining. *The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings*. 454–457.

GONZALES, Michael L., 2003. *IBM data warehousing: with IBM business intelligence tools*. New York: Wiley. ISBN 978-0-471-13305-6.

HENDL, Jan, 2021. *Big data: věda o datech - základy a aplikace*. První vydání. Praha: Grada Publishing. ISBN 978-80-271-3031-3.

KROENKE, David a David J. AUER, 2015. *Databáze*. 1. vyd. Brno: Computer Press. ISBN 978-80-251-4352-0.

KŘÍŽ, RNDr Oldřich, Mgr Jiří NEUBAUER a Mgr Marek SEDLAČÍK, [b.r.]. POPISNÁ STATISTIKA A VÝBĚROVÁ ŠETŘENÍ. 176.

LACKO, Ľuboslav, [b.r.]. Strojové učení: S učitelem i bez něj. *CIO Business World* [online] [vid. 2023-01-01]. Dostupné z: https://www.cio.cz/clanky/strojove-uceni-s-ucitelem-i-bez-nej/

MANAGEMENTMANIA, [b.r.]. Segmentace trhu a zákazníků (Market Segmentation). *ManagementMania.com* [online] [vid. 2023-02-12]. Dostupné z: https://managementmania.com/cs/segmentace-trhu

MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER, 2014. *Big Data*. 1. vyd. Brno: Computer Press. ISBN 978-80-251-4119-9.

NOVOTNÝ, Ota, Jan POUR a David SLÁNSKÝ, 2005. *Business intelligence: jak využít bohatství ve vašich datech*. 1. vyd. Praha: Grada Publishing. ISBN 978-80-247-1094-5.

OPOJIŠTĚNÍ, [b.r.]. SVIPO II nově v oblasti pojištění osob. *oPojištění* [online] [vid. 2023-02-19]. Dostupné z: https://www.opojisteni.cz/spektrum/svipo-ii-nove-v-oblasti-pojisteni-osob/c:10625/

SCHILLER, Martin, [b.r.]. *Co se skrývá pod zkratkou ETL?* [online] [vid. 2022-12-23]. Dostupné z: https://www.systemonline.cz/clanky/co-se-skryva-pod-zkratkou-etl.htm

SIOBOS, Aneta, [b.r.]. *Lekce 3 - Data Mining - Metodologie procesu a používané techniky* [online] [vid. 2023-01-04]. Dostupné z: https://www.itnetwork.cz/metodologie-data-mining-procesu-a-pouzivane-techniky

SKALSKÁ, Hana, 2010. *Data mining a klasifikační modely*. Vyd. 1. Hradec Králové: Gaudeamus. ISBN 978-80-7435-088-7.

SVĚTLÍK, Jaroslav, 2005. *Marketing - cesta k trhu*. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk. ISBN 978-80-86898-48-3.

SYROVÁTKOVÁ, Ing Jaroslava, [b.r.]. Bankovní produkty – 1. část.

TREJBAL, Pavel, 2014. Jak na rozhodovací stromy. *Optimics* [online]. [vid. 2023-01-01]. Dostupné z: https://www.optimics.cz/jak-na-rozhodovaci-stromy/