TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Aplikace dolování dat pro podporu rozhodování v moderních podnicích

Bakalářská práce

Studijní program: B6209 Systémové inženýrství a informatika
Studijní obor: Manažerská informatika

Autor práce: **Petr Hollmann**
Vedoucí práce: Ing. Athanasios Podaras, Ph.D.
Katedra informatiky

Liberec 2022

TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Zadání bakalářské práce

Aplikace dolování dat pro podporu rozhodování v moderních podnicích

Jméno a příjmení: **Petr Hollmann**
Osobní číslo: E19000217
Studijní program: B6209 Systémové inženýrství a informatika
Studijní obor: Manažerská informatika
Zadávající katedra: Katedra informatiky
Akademický rok: **2021/2022**

Zásady pro vypracování:

1. Big data jako cenný vstup při rozhodovacím procesu
2. Data mining nástroje a techniky pro manipulaci s big data
3. Aplikování specifických data mining nástrojů ve vybranné společnosti
4. Analýza a prezentace vyvozených výsledků a diskuse
5. Závěry a budoucí implementace

Rozsah grafických prací:
Rozsah pracovní zprávy: 30 normostran
Forma zpracování práce: tištěná/elektronická
Jazyk práce: Čeština

Seznam odborné literatury:

- HAN, Jiawei, Micheline KAMBER a Jian REI, , 2012. *Data mining – concepts and techniques. Third Edition*. Burlington, MA: Elsevier. ISBN 978-0-12-381479-1.
- GROSSMANN, Wilfried a Stefanie RINDERLE-MA., 2012. *Fundamentals of Business Intelligence.* Berlin Heidelberg:Springer-Verlag. ISBN 978-3-662-46530-1.
- RAHLF, Thomas, 2017. *Data visualisation with R-100 examples.* Cham: Springer International Publishing. ISBN 978-3-319-49751-8.
- PROVOST, Foster a Tom FAWCETT, 2015. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.* Beijing Köln: O'Reilly. ISBN 978-1-449-36132-7.
- SLÁNSKÝ, David, 2018. *Data a analytika pro 21. století.* Praha: Professional Publishing. ISBN 978-80-88260-25-7.
- PROQUEST. 2021. *Databáze článků ProQuest* [online]. Ann Arbor, MI, USA: ProQuest. [cit. 2021-09-18]. Dostupné z: http://knihovna.tul.cz/

KONZULTANT: Šimon Kolář, FRM, KPMG Czech Republic, s.r.o., Risk Consulting.

Vedoucí práce: Ing. Athanasios Podaras, Ph.D.
Katedra informatiky

Datum zadání práce: 1. listopadu 2021
Předpokládaný termín odevzdání: 31. srpna 2023

L.S.

doc. Ing. Aleš Kocourek, Ph.D.
děkan

Ing. Petr Weinlich, Ph.D.
vedoucí katedry

V Liberci dne 1. listopadu 2021

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědom toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

6. května 2022 Petr Hollmann

Anotace

Tato bakalářská práce se zabývá obecně problematikou datové analýzy, konkrétně aplikací Data Miningových technik pro podporu a automatizaci rozhodování v moderních podnicích. V práci je nejprve uvedena definice, historie a budoucnost problematiky Big Dat. Dále se práce zaměřuje na techniky Data Miningu s důrazem na jejich použití v bankovním sektoru. V poslední části je vytvořen vícenásobný regresní model pro stanovení koeficientu Forward Looking Information (FLI), který slouží pro úpravu parametru Probability of Default (PD) v modelu pro výpočet očekávaných úvěrových ztrát (Expected Credit Loss, ECL), který je konceptuálně ukotven v mezinárodním standardu finančního výkaznictví (IFRS), konkrétně v části o finančních nástrojích (IFRS 9).

Klíčová slova

bankovnictví, Big Data, Data Mining, Forward Looking Information, International Financial Reporting Standard, lineární regrese, očekávané úvěrové ztráty, rozhodování na základě dat

Annotation

The bachelor thesis deals generally with the issue of data analysis, specifically the application of Data Mining techniques for decision support and automation in modern enterprises. The thesis first presents the definition, history and future of Big Data. Furthermore, the thesis focuses on Data Mining techniques with emphasis on application in the banking sector. In the last part, a multiple regression model is developed to determine the coefficient of Forward Looking Information (FLI), which is used to adjust the Probability of Default (PD) parameter in the Expected Credit Loss (ECL), which is conceptually anchored in the International Financial Reporting Standard (IFRS), specifically in the section on financial instruments (IFRS 9).

Keywords

banking, Big Data, Data Mining, Forward Looking Information, International Financial Reporting Standard, linear regression, Expected Credit Loss, data-driven decision making

Poděkování

Rád bych poděkoval vedoucímu práce panu Ing. Athanasiosu Podarasovi, Ph.D. a spolupracovníkům v oddělení Risk Consulting ve společnosti KPMG Česká republika, s.r.o. za cenné připomínky a skvělý přístup v průběhu tvorby bakalářské práce.

Obsah

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

Zkratka	Celý název	Popis
3V	Volume, Velocity, Variety	Základní metriky Big Dat – Rozsah, Rychlost, Různorodost.
CRISP-DM	Cross Industry Standard Process for Data Mining	celosvětově uznávaná standardizovaná metodologie pro Data Mining.
CRM	Customer Relationship Management	Řízení vztahů se zákazníky.
CRR	Capital Requirements Regulation	Nařízení o obezřetnostních požadavcích na úvěrové instituce a investiční podniky ze dne 26. června 2013.
DDD	Data Driven Decision Making	Rozhodování na základě dat.
EAD	Exposure At Default	Odhad celkové expozice banky, která je nezajištěná vůči protistraně v případě že dojde k selhání protistrany.
ECL	Expected Credit Loss	Očekáváná úvěrová ztráta. Vážený průměr úvěrových ztrát, kde vahami jsou příslušná rizika neplnění závazků.
EIR	Effective Interest Rate	Úroková míra, která přesně diskontuje odhadované budoucí peněžní platby nebo příjmy po očekávanou dobu trvání finančního aktiva nebo finančního závazku na hrubou účetní hodnotu finančního aktiva nebo na naběhlou hodnotu finančního závazku.
FLI	Forward Looking Information	Očekávané budoucí ekonomické podmínky.
HDFS	Hadoop Distributed File Systém	Distribuovaný souborový systém napsaný v jazyce JAVA.

HDP	GDP – Gross demand product	Hrubý domácí produkt.
IAS	Internation Accounting Standards	Mezinárodní účetní standard.
IFRS 9	International Financial Reporting Standard, Financial Instruments	Mezinárodní standard finančního výkaznictví – Finanční nástroje.
IoT	Internet of Things	Internet věcí – jedná se o síť fyzických zařízeních se senzory a síťovou konektivitou, která umožňuje propojení a výměnu dat.
KDD	Knowledge Discovery from Data	Proces získávání znalostí z dat.
LGD	Loss Given Default	Poměř ztráty z expozice při selhání protistrany k dlužné částce přesně v momentě selhání.
PD	Propability of default	Pravděpodobnost selhání expozice
SICR	Significant Increase in Credit Risk	Významný nárůst úvěrového rizika.
Stage	Stage	Stupeň znehodnocení expozice.
Stage I	Stage I	Expozice, ve kterých nedošlo k signifikantnímu nárůstu úvěrového rizika od počátečního zaúčtování, odhad ročních očekávaných úvěrových ztrát.
Stage II	Stage II	Expozice, ve kterých došlo k signifikantnímu nárůstu úvěrového rizika od počátečního zaúčtování, odhad celoživotních očekávaných úvěrových ztrát.
Stage III	Stage III	Expozice v defaultu, odhad celoživotních očekávaných úvěrových ztrát.
YARN	Yet Another Resource Negotiator	Systém řízení zdrojů pro Hadoop.

Úvod

V dnešní digitální době je lidstvo zahlceno datami všeho druhu od vědeckých, zdravotních, demografických, finančních až po marketingové. Data se v průběhu času stala velice cenným aktivem společností. Moderní společnosti se na základě dat rozhodují při činnostech jako je zlepšení služeb, zvýšení konkurenceschopnosti a produktivity. Nicméně dat na světě vzniká tolik, že je potřeba neustále hledat cesty, jak data analyzovat na automatické bázi. Jedná se o jednu z nejvíce aktivních oblastí databázového výzkumu. Výzkumní pracovníci v oblastech statistiky, vizualizace dat, umělé inteligence nebo strojového učení udělali v posledních desetiletích mimořádný pokrok. (Provost a Fawcett 2013)

V moderních podnicích po celém světě jsou denně generovány obrovské soubory dat, například prodejní transakce, záznamy o obchodování s akciemi, popisy výrobků a mnoho jiných údajů. Jejich byznysové rozhodnutí jsou z velké části ovlivněny datovou analýzou. Ukazuje se, že rozhodnutí na základě dat v dlouhodobém měřítku podniku přináší větší výnos a lepší postavení na trhu než rozhodování na základě osobních zkušeností nebo jiných subjektivních faktorů. Nicméně je vždy moudré zasadit informace získané z dat do kontextu situace a posuzovat data pouze jako jeden ze vstupních faktorů při rozhodování. Potřeba data efektivně analyzovat vedlo ke zrodu Data Miningu. Tento obor je mladý, dynamický a slibný. Data Mining udělal a nadále bude dělat velké pokroky v následujících desetiletích. (Provost a Fawcett 2013)

Teoretickou část práce tvoří souhrnný popis problematik Big Dat, Data Miningu s důrazem na využití těchto technologií v bankovním sektoru a také popis konceptuálního modelu pro výpočet očekávaných úvěrových ztrát (Expected Credit Loss, dále jen ECL). ECL jsou ukotvené v mezinárodním účetním standardu pro finanční nástroje (International Financial Reporting Standard – Financial Instruments, dále jen IFRS 9).

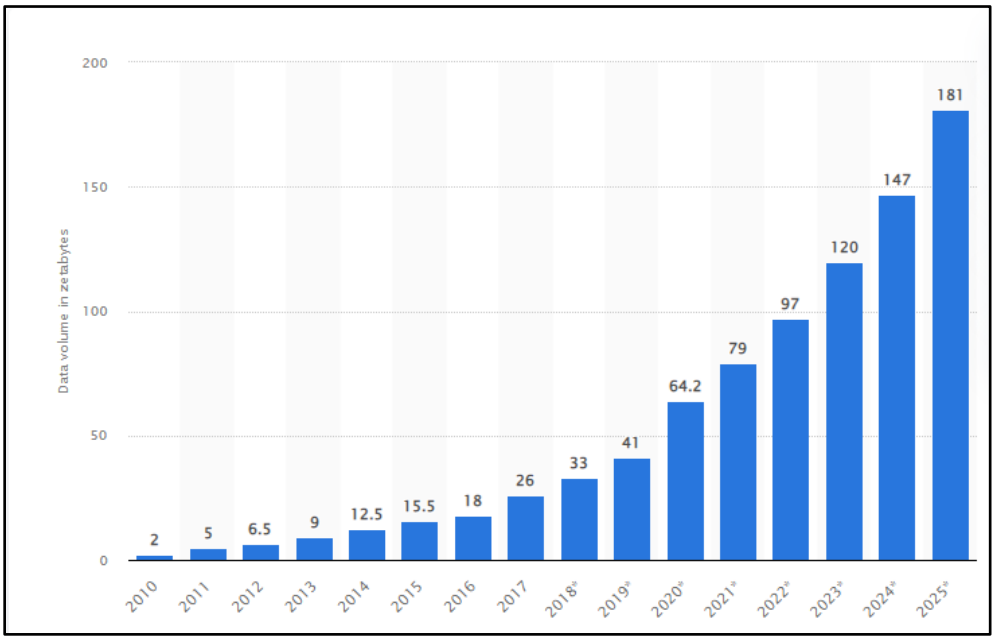
Primárním cílem bakalářské práce je za pomoci technik Data Miningu a standardu IFRS 9 vytvořit vícenásobný regresní model založený na makroekonomických veličinách inflace, nezaměstnanosti, hrubého domácího produktu a pravděpodobnosti tržního selhání. Výsledné koeficienty z modelu jsou dále použity pro předpověď očekávaných budoucích ekonomických podmínek (Forward Looking Information, dále jen FLI). Finální FLI jsou vyjádřeny koeficientem, který slouží pro úpravu pravděpodobnosti selhání (Probability of Default, dále jen PD) při výpočtu ECL.

1 Big Data

Tato kapitola je zaměřena na problematiku Big Dat, do českého jazyka se tento termín často překládá jako velká data nebo veledata, nicméně v rámci práce bude autor používat originální anglické označení Big Data.

Postupně se v kapitole popíše, co jsou to data, informace, znalosti, a jaký je mezi těmito termíny rozdíl. Dále se definují Big Data a jejich vlastnosti. Stručně je v kapitole nastíněna i historie, vývoj nástrojů a požadavků na zpracování dat. V poslední části je nastíněno, jaký by mohl být v budoucnosti vývoj v oblasti Big Dat.

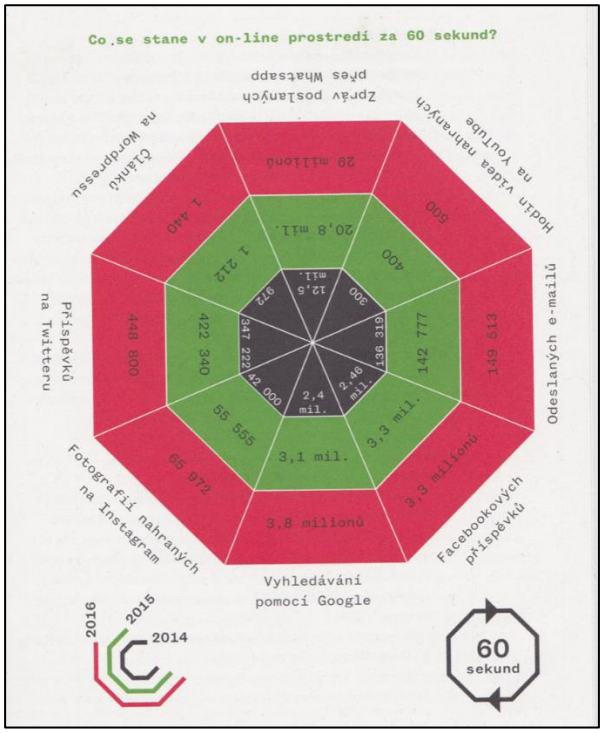
Růst objemu dat je v posledních třech desetiletí exponenciální, především díky náhlému vzestupu v oblasti informačních technologií. Jak ukazuje Obrázek 1, tak objem dat se dle výzkumu společnosti Statista každé dva až tři roky zdvojnásobí. Nárůst dat v posledních letech je větší než se předpokládalo, hlavně z důvodu celosvětové pandemie covid-19, která urychlila proces digitalizace. Nicméně jen malé procento těchto nově vytvořených dat se uchová. Pouze 2 % dat vytvořených v roce 2020 bylo uchováno i v roce 2021. (Statista Research Department 2022)



Obrázek 1: Objem vytvořených dat (a predikce od roku od 2018) na světě v letech 2010 až 2025 v zettabajtech

zdroj: Statista Research Department 2022

Pro lepší představu je na Obrázku 2 zobrazen objem vytvořených dat v online prostředí za 60 sekund (tedy za jednu minutu). Výzkum společnosti KPMG z roku 2017 uvádí, že každých 60 sekund je vytvořeno více multimediálního obsahu, než je jeden člověk schopen zkonzumovat za celý život. Tento trend se bude i nadále prohlubovat a množství dat bude růst exponenciálně. (Slánský 2018)



Obrázek 2: Tempo růstu datového objemu v online prostředí v letech 2014-2016
zdroj: Slánský 2018

1.1 Data, informace a znalosti

Pojmy data, informace a znalosti se mohou na první pohled zdát jako synonyma, ale ve skutečnosti jsou to odlišné termíny. Jejich vztah je zobrazen na Obrázku 3.

1.1.1 Data

Data reprezentují objektivní skutečnosti o událostech. Mohou to být čísla, písmena nebo jiné symboly. Data se získávají měřením, pozorováním nebo šetřením. Sami o sobě nedávají data smysl a nepřináší příjemci žádnou přidanou hodnotu. Po jejich získání se je příjemce snaží pochopit, přiřadit jim význam a následně je na základě znalostí interpretovat. (Cejpek a Univerzita Karlova 2005)

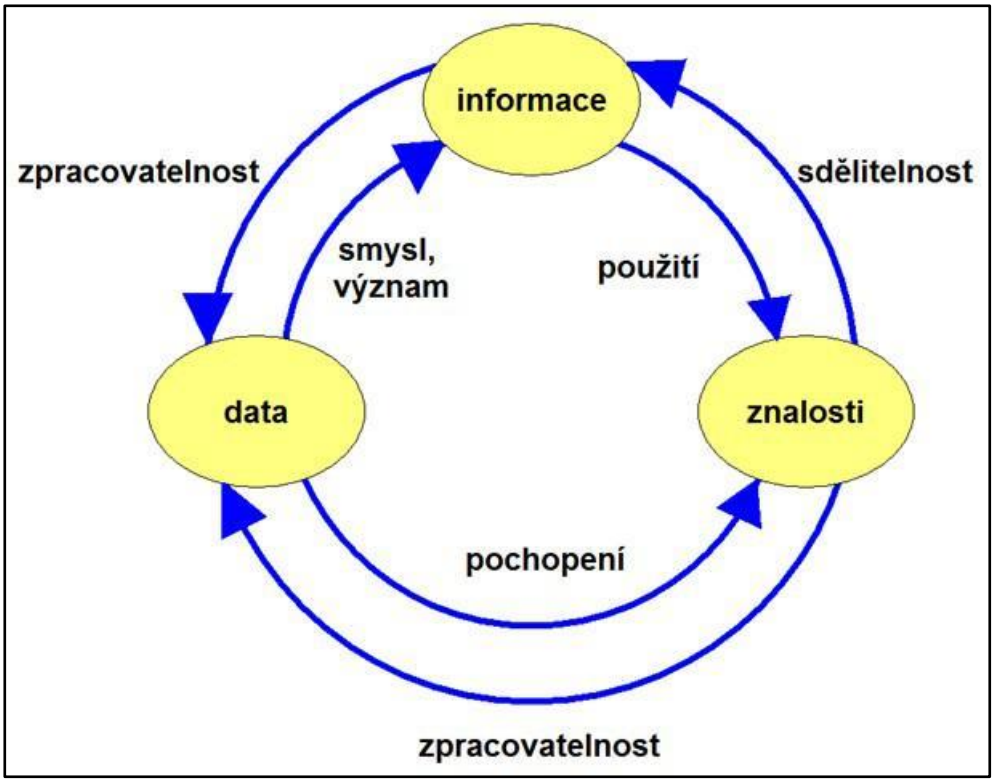
Data se dělí na **strukturovaná**, ty zachycují výhradně atributy, objekty a fakta. Nejčastěji je možné se setkat se strukturovanými daty v relačních databázích. Druhým typem jsou **nestrukturovaná** data, která lze vyjádřit jako určitý tok bitů. Mezi nestrukturovaná data se řadí například obrázky, videozáznamy nebo nestrukturované dokumenty, jako je volně psaný text. Již diskutovaná Big Data jsou z drtivé většiny nestrukturovaná, a proto musí být před analýzou upravována do adekvátní struktury. (Cejpek a Univerzita Karlova 2005)

1.1.2 Informace

Přiřazením významu datům na základě znalostí, zkušeností a vědomostí příjemce vzniká informace, která slouží pro snížení entropie neboli neurčitosti s ohledem na příjemcovy potřeby a požadavky na další zpracování, skladování nebo přenášení dat. Například číslice 5 patří mezi data, pokud tuto číslici ale vysvětlíme jako ranní hodnotu na venkovním teploměru, získáme z ní na základě předchozích zkušeností užitečnou informaci o tom, že je venku chladno, a tudíž je potřeba adekvátně se obléct. (Sklenák 2001)

1.1.3 Znalosti

Znalosti představují zobecněné poznání reality. Vznikají především ze zkušeností a studia. Dále jsou závislé na inteligenčních schopnostech a na mentální kapacitě daného jedince. Znalosti jsou dlouhodobé principy ukotvené v lidském mozku a jejich ukotvení trvá určitý čas. (Sklenák 2001)



Obrázek 3: Vztah mezi daty, informacemi a znalostmi

zdroj: Novotný nedatováno

1.2 Definice Big Dat

Big Data jsou soubory dat, které jsou příliš velké na to, aby je za rozumný čas dokázal zpracovat běžný hardware a software. Většinou se jedná o objemy dat v řádech terabytů a petabytů, které jsou nestrukturované a ukládané v datových skladech. (ECOVIS ježek team 2019)

Technologický pokrok 21. století (náhlý vzestup webů, chytrých zařízení nebo IoT) zapříčinil, že zpracování dat, které bylo historicky poměrně statickou úlohou, začalo být více dynamické. Data již nejsou centralizovaná, strukturovaná v relačních databázích a snadno zvládnutelná běžnými metodami.

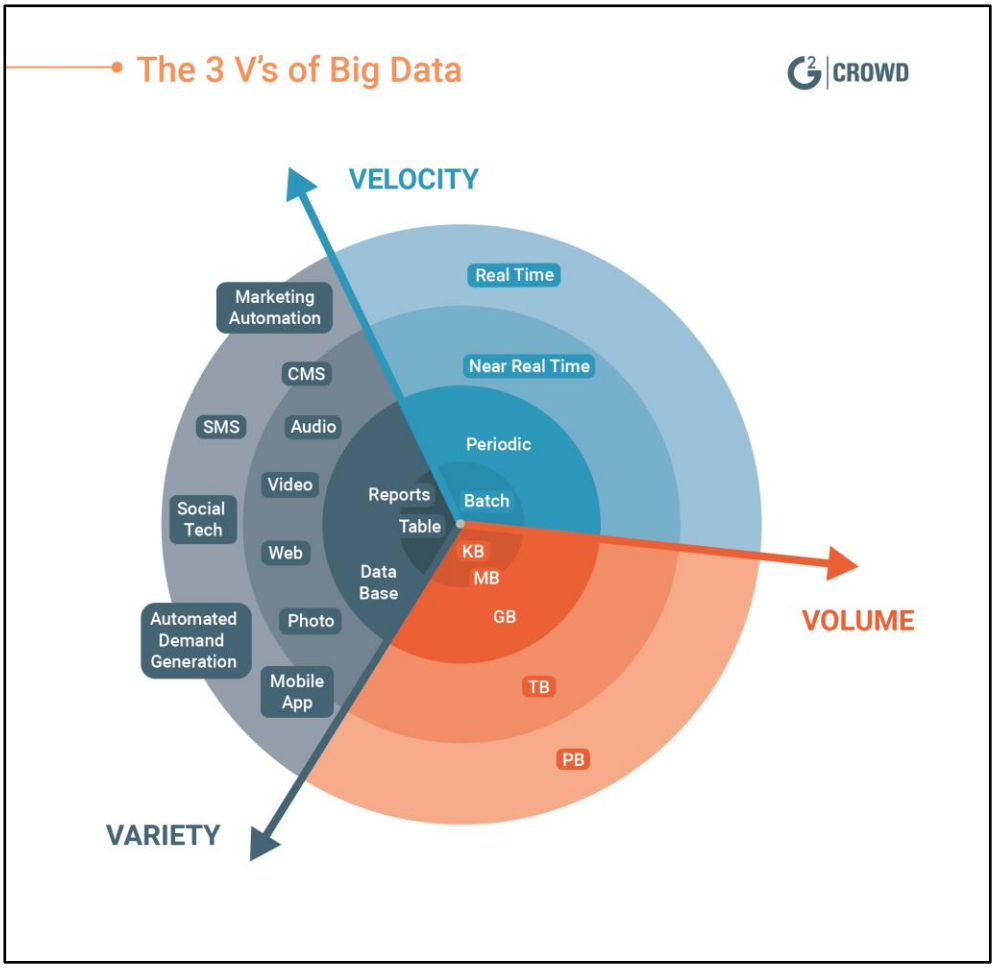
Pro lepší uchopení problematiky Big Data se zavedla metrika tzv. trojrozměrnosti dat. První rozměr je rozsah (Volume), druhý rozměr je rychlost (Velocity) a třetí rozměr je různorodost (Variety). Metrika je zobrazena na Obrázku 4 (často se v literatuře používá zkratka 3V podle prvních písmen anglických ekvivalentů, respektive 3R v literatuře české) (Dolák 2011)

1.2.1 Rozsah, rychlost a různorodost

Rozsah (Volume) neboli objem je ukazatel množství vygenerovaných a uložených dat. Velikost dat určuje jejich hodnotu, potenciální využitelnost a zda je lze vůbec považovat za velká neboli „Big“. Data se považují za Big Data pokud dosahují objemu většího nebo rovnu terabytům. (Smari a Annual IEEE Computer Conference 2013)

Rychlost (Velocity) ukazuje, s jakou rychlostí jsou data generována a zpracována. Big Data jsou často k dispozici v reálném čase. Ve srovnání s běžnými daty jsou Big Data vytvářena ve větší frekvenci. Rychlost se rozděluje na dva druhy: rychlost generování a rychlost zpracování, zaznamenání a zveřejňování. (McArdle a Kitchin 2016)

Různorodost (Variety) znamená množství typů dat, které jsou k dispozici. Klasické datové typy jsou strukturované a často dokonale zapadají do relační databáze. S tím jak rostou Big Data na významu, se objevují ve více nestrukturované podobě, například jako obrázky, videa nebo volný text. Často vyžadují dodatečné předzpracování, aby bylo možné odvodit význam a doplnit metadata. (Oracle Team 2022)



Obrázek 4: Základní metriky Big Data

zdroj: Pickell 2018

1.2.2 Věrohodnost a složitost

V literatuře se dále objevuje i čtvrtá a pátá metrika, které byly přidány k původnímu rozsahu, rychlosti a různorodosti. Jedná se o věrohodnost a složitost.

Společnost IBM definovala čtvrtou metriku jako tzv. **věrohodnost (Veracity)**. Big Data jsou často nekonzistentní, neúplná nebo nevěrohodná. Při analyzování je tedy potřeba s určitou mírou nevěrohodnosti vždy počítat. (Diáz Alba 2020)

Společnost SAS přidala pátou metriku, a tou je **složitost (Complexity).** Dnešní data pocházejí z mnoha zdrojů a jejich sjednocení je stále složitější. Právě sjednocení a určení vztahů mezi daty je pro kvalitní analýzu klíčové, jinak nad svými daty společnosti rychle ztratí kontrolu a hodnota získaných dat rapidně klesá. (SAS Team 2015)

1.3 Historie Big Dat

Pojem Big Data se používá od počátku 90. let. Ačkoliv není jasné, kdo tento termín použil jako první, jeho zpopularizování se připisuje Johnu R. Masheyovi. (Big Data Framework Team 2019)

1.3.1 První problém

Počátky problému zahlcení velkým objemem dat sahá až do druhé poloviny 19. století. V roce 1880 Americký úřad pro sčítání lidu procházel těžkým obdobím, především kvůli velkému nárůstu obyvatel. Zpracování výsledků sčítání trvalo úřadu 8 let. V roce 1890 statistik Herman Hollerith vynalezl počítací stroj, který pomocí děrných štítků umožňoval hromadné zpracování dat. Sčítání lidu z roku 1890 trvalo vyhodnotit 3 měsíce a Hollerith tak ušetřil úřadu na nákladech tohoto sčítání přes 5 milionů dolarů. V průběhu 20. století se data vyvíjela vysokou rychlosti a v roce 1965 vybudovala americká vláda první datové centrum se záměrem uchovávat miliony sad otisků prstů a daňových přiznání. (Adilin 2021)

1.3.2 První fáze (1970 – 2000)

V roce 1970 publikovaly výzkumné laboratoře IBM první článek o relačních databázích. Relační databáze umožnily efektivnější způsoby vyhledávání a zpracování dat ve velkých databázích. (Pickell 2018)

Základem první fáze byla správa databází a datové sklady. Z dnešního pohledu se jednalo o běžné techniky ukládání, vytěžování databázovými dotazy a optimalizaci, které jsou standardní u relačních databází. (Big Data Framework Team 2019)

1.3.3 Druhá fáze (2000 – 2010)

Od počátku nového století začal internet a jeho webové stránky nabízet jedinečné možnosti sběru dat a jejich následné analýzy. Šlo především o společnosti jako Yahoo, Amazon nebo eBay, které začaly analyzovat chování zákazníků.

Největší průlom druhé fáze ale přišel s příchodem sociálních sítí (zpočátku Facebook a Twitter), což vedlo k šíření ještě většího množství dat. Tyto události zvýšily potřebu nástrojů, technologií a analytických technik, které by z těchto dat dokázaly získat smysluplné informace. (Big Data Framework Team 2019)

V roce 2005 byl vyvinut open-sourcový nástroj Hadoop, který byl vytvořen speciálně pro ukládání a analýzu velkých datových sad.

Nástroj Hadoop vyvinula společnost Apache Software Foundation a do dnešního dne stále vycházejí aktualizované verze. Pro zpracování Big Dat je zásadní, jelikož díky němu je zpracování velkých objemů dat snazší a jejich ukládání výrazně levnější. Hadoop slouží jako báze, na které jsou postavené mnohé komerčně dodávané řešení pro společnosti jako je Amazon, Apple, Google nebo Netflix. (Oracle Team 2022)

Hadoop se skládá ze tří komponent, které byly speciálně navrženy pro práci s Big Daty:

1. První komponentou je distribuovaný souborový systém HDFS (Hadoop Distributed File System). Skladování Big Dat na jednom zařízení je neudržitelné, a proto jsou data ukládaná na více zařízeních v podobě datových bloků.
2. Druhou komponentou je tzv. MapReduce, ve které se data rozdělí do několika bloků. Každý blok se následně zpracovává souběžně napříč uzly. Na konci procesu se výsledky zpracování spojí a ušetří značné množství času.
3. Třetí komponentou je tzv. YARN (Yet Another Resource Negotiator), která má na starost systém řízení zdrojů. Zpracovává požadavky na úlohy a efektivně spravuje prostředky jednotlivých uzlů. (Apache Hadoop Team 2022)

1.3.4 Třetí fáze (2010 – současnost)

Hlavním cílem analýzy většiny organizací jsou stále nestrukturovaná data z webových stránek. V současné době se objevují nové možnosti získávání cenných informací z mobilních zařízení a tzv. internetu věcí (Internet of Things, dále jen IoT).

Mobilní zařízení umožňují analyzovat behaviorální data[1] (například kliknutí na webové stránce nebo vyhledávání dotazů na Google), a také je možné sledovat pohyb nebo údaje týkající se zdraví (počet kroků, tep). Tato data poskytují nepřeberné množství nových možností, jak je využít. Pro příklad je možné na základě dat upravit zdravotní péči nebo optimalizovat dopravu ve městech. (Big Data Framework Team 2019)

V roce 2014 se začal rozšiřovat fenomén IoT. Jednotlivé zařízení generují velké množství dat každý den. Dle společnosti Statista v roce 2019 existovalo více než 10 miliard aktivních IoT zařízení. Předpokládá se, že v roce 2025 překročí počet aktivních zařízení 30 miliard. (Vailshery 2021)

Díky propojení světa s internetem se moderní podniky rozhodly zvýšit výdaje na analýzu dat tak, aby objevily možnosti, jak snížit provozní náklady, zvýšit efektivitu a vyvinout nové produkty a služby na základě datové analýzy.

Tento trend je patrný i ve státní sféře, po světě vznikají tzv. „smart cities“, která využívají data v reálnem čase ke sledování spotřeby elektřiny, dopravy nebo aktuálního znečištění vzduchu a spousty dalších parametrů. (Pickell 2018)

1.4 Budoucnost Big Dat

Celé odvětví Big Data je stále v počátcích. Dle společnosti Allied Market Research trh s Big Data a podnikovou analytikou dosáhl v roce 2019 hodnoty 193,14 miliardy dolarů a odhaduje, že do roku 2027 vzroste na 420,95 miliardy dolarů. Tomu odpovídá meziroční nárůst o 10,9 %. (Phillips 2021)

[1] Behaviorální data popisují, co uživatelé dělají na webových stránkách, v mobilních aplikacích, e-mailech, chatbotech, nositelných zařízeních (například chytrých hodinkách) nebo zařízeních IoT.

Big Data jsou často spojována s dalšími pojmy jako je Machine Learning (strojové učení), Data Science (datová věda), Artifical Inteligence (umělá inteligence), Deep Learning (hluboké učení) a další. Všechny pojmy mají jeden stejný jmenovatel, a tím jsou data. Data budou hrát čím dál větší roli při zlepšování současných produktů, služeb a umožní pokrok ve výzkumu. Například každé auto společnosti Tesla, které má autonomní řízení, odesílá data, díky kterým se trénuje model umělé inteligence a s každou chybou jej neustále zlepšuje. (Neo 2020)

2 Data Mining

Druhá kapitola je zaměřena na problematiku Data Miningu, do českého jazyka se tento termín často překládá jako dolování nebo vytěžování dat, nicméně v rámci práce bude autor používat originální anglické označení Data Mining.

V rámci kapitoly je prvně definován pojem Data Mining a krátce popsaná jeho historie. Dále jsou rozděleny základní typy úloh, které se pomocí Data Miningu řeší, jaké mohou nastat problémy při tvoření modelů a jak tyto problémy efektivně řešit. Poslední část se týká Data Miningu v bankovním sektoru, kde je rozhodování na základě dat rychle se vyvíjející trend.

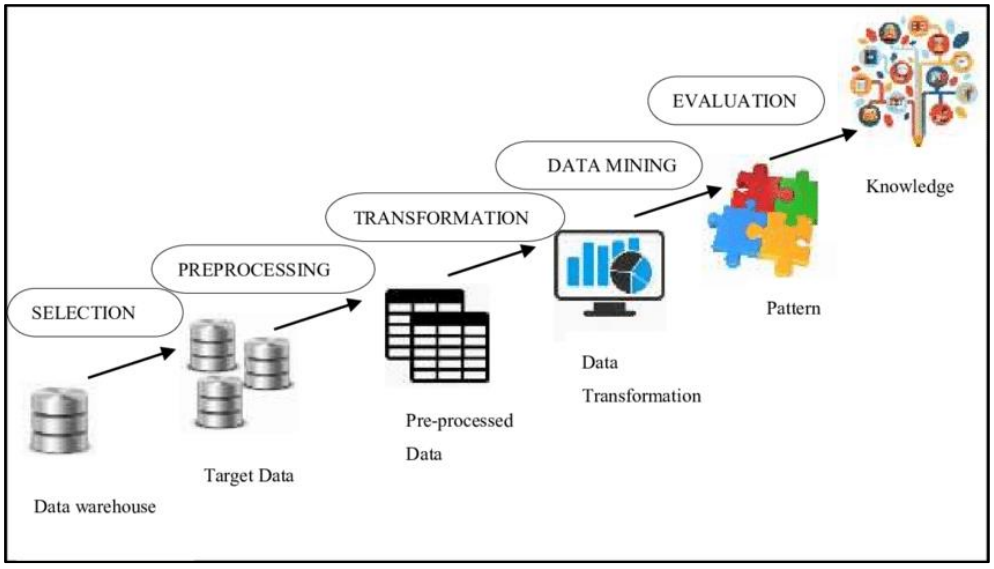
Data Mining je analytickým krokem v procesu dobývání znalostí z databází (Knowledge Discovery in Databases, dále jen KDD). Termín Data Mining v sobě ukrývá hned několik disciplín, především statistiku, strojové učení a databázové systémy. (Provost a Fawcett 2013)

Přínosy rozhodování založené na datech (Data Driven Decision Making, dále jen DDD) byly přesvědčivě prokázány. Ekonom Erik Brynjolfsson a jeho kolegové z MIT a Penn's Wharton School provedli studii o tom, jak DDD ovlivňuje výkonnost firmy. Vyvinuli měřítko DDD, které hodnotí firmy podle toho, jak silně využívají data k rozhodování v rámci celé společnosti. Došli k závěru, že čím více se firma řídí daty, tím je produktivnější. Posun o jednu směrodatnou odchylku na stupnici DDD je spojen se 4 až 6% nárůstem produktivity. DDD také koreluje s vyšší rentabilitou aktiv, rentabilitou vlastního kapitálu, využitím aktiv a tržní výkonností. Závěrem je, že vztah mezi DDD a výše jmenovanými ukazateli je kauzální. (Provost a Fawcett 2013)

2.1 Definice Data Miningu

Často se pojmy Data Mining a KDD mylně považují za synonyma. Ve skutečnosti je Data Mining pouze jeden z kroků KDD. Tento proces se dělí do několika kroků, které se mohou opakovat do té doby, než je dosaženo kýženého výsledku. Tyto kroky ukazuje Obrázek 5:

1. **Čištění dat a selekce (Selection)** – Slouží především pro odstranění nekonzistentních a odlehlých dat (Outliers) a následné selekce relevantních dat pro danou úlohu.
2. **Datová integrace (Preprocessing)** – Některé úlohy vyžadují spojení více datových zdrojů, často se první a druhý krok souhrnně označují jako předzpracování (Preprocessing).
3. **Datová transformace (Transformation)** – Data je nutné sjednotit do konsolidované formy a připravit pro účely technik Data Miningu.
4. **Data Mining** – Proces objevování skrytých vzorů v datech a znalostí za pomocí metod Data Miningu.
5. **Zhodnocení vzorů (Evaluation)** – Identifikace skutečně důležitých vzorů, které vyjadřují vydolované znalosti.
6. **Vizualizace výsledků (Knowledge) –** Finální krok, ve kterém jsou výsledné znalosti vizualizovány uživatelům. Častými spotřebiteli těchto dat je vyšší management nebo představenstvo společnosti. (Han a Kamber 2012)



Obrázek 5: Proces získávání znalostí z databází (KDD)

zdroj: (Ily amalina ahmad sabri 2019)

2.2 Historie Data Miningu

Kořeny Data Miningu sahají do druhé poloviny 18. století, kdy byl publikován článek Thomase Bayese o teorému pro vztah aktuální pravděpodobnosti a předchozí pravděpodobnosti, který se nazývá Bayesova věta. Ta má zásadní význam pro Data Mining a pravděpodobnost, protože umožňuje pochopení složitých skutečností na základě odhadovaných pravděpodobností.

V roce 1805 A. M. Legendre a C. F. Gauss použili regresi k určení drah těles kolem Slunce. Regrese je jedním z klíčových nástrojů při získávání dat.

První polovina 20. století byla začátkem éry počítačů. V publikaci On Computable Numbers z roku 1936 pojednává Alan Turing o myšlence univerzálního stroje schopného provádět výpočty jako naše současné počítače. Moderní počítače jsou postaveny na koncepcích, jejichž průkopníkem byl právě Turing. (Li 2016)

V druhé polovině 20. století se stal rozvoj počítačů, výpočetní techniky a zavedení elektronického sběru dat hnacím motorem pro zpracování a využívání informací z dat ke zvýšení výdělku, optimalizaci produktů, a hlavně získáním výhody oproti konkurenci.

Na takové objemy dat už standardní statistické metody nebyly vhodné a bylo potřeba nalézt metody, které dokážou vydolovat i složitější nelineární vztahy. Pojem Data Mining se začal objevovat v devadesátých letech, kdy vznikla i první definice od W. J. Frawleyho, G. Piatetsky-Shapira a Ch. J. Matheuse: "*The non trivial extraction of implicit, previously unknown, and potentially useful information from data.*", (StatSoft Team 2014) respektive autorsky přeloženo jako: „*Netriviální extrakce implicitních, dříve neznámých, a potencionálně užitečných informací z dat.*“

V 21. století je Data Mining rozšířen například v podnikání, vědě, inženýrství a medicíně. Těží se informace z transakcí kreditních karet, pohybů na burze, národní bezpečnosti nebo z nositelných zařízení například chytrých hodinek.

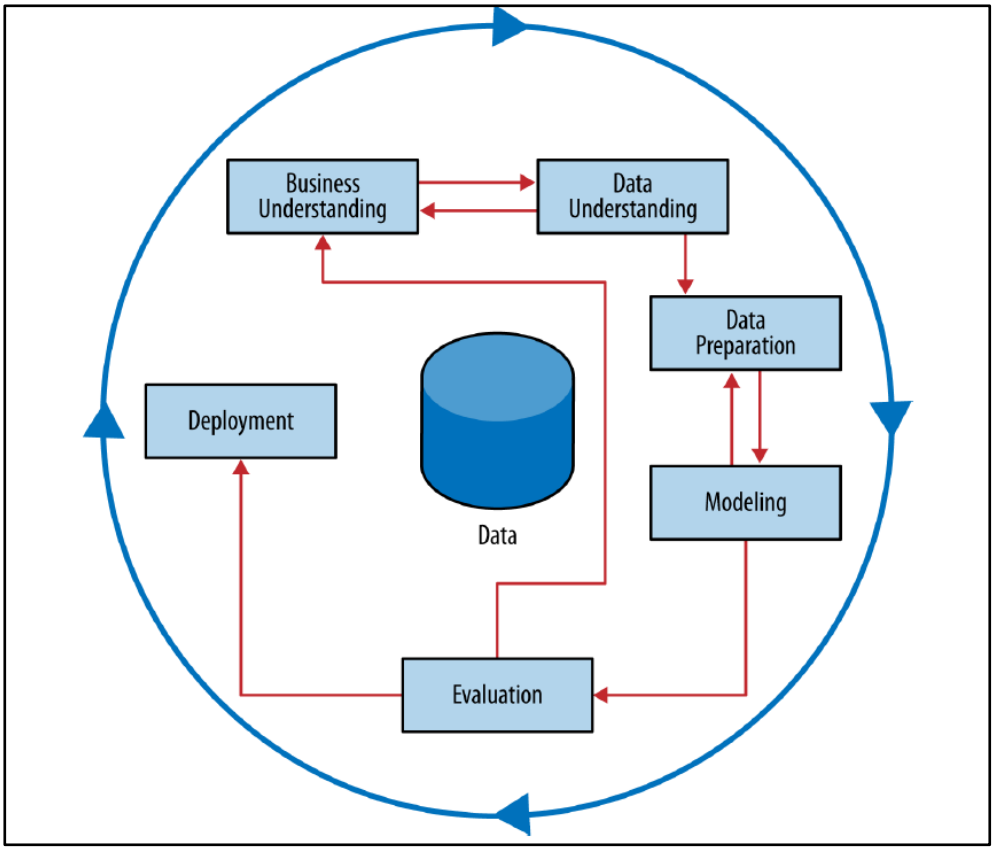
Jednou z nejaktuálnějších technik, které se dnes zkoumají je Deep Learning (hloubkové učení). Tato technika dokáže zachytit závislosti a složité vzorce daleko lépe než jiné techniky a znovu otevírá některé z největších výzev ve světě Data Miningu, datové vědy a umělé inteligence. (Li 2016)

2.3 Standardizovaný proces CRISP-DM

Zkratka CRISP-DM (Cross Industry Standard Process for Data Mining), kterou je možné přeložit jako mezioborový standardní proces pro dolování dat. Je celosvětově uznávanou standardizovanou metodologií pro Data Mining. Jednotlivé fáze procesu jsou zobrazeny na Obrázku 6. Tyto fáze jsou:

1. **Pochopení obchodních souvislostí (Business Understanding)** – Zpočátku je nezbytné porozumět problému, který je třeba vyřešit. Definice problému a návrh jeho řešení je často opakující se proces, jak je zobrazeno i na Obrázku 6. Počáteční formulace problému nemusí být úplná nebo exaktní. Proto je potřeba opakovat tento proces až do fáze, kdy je formulováno optimální řešení.
2. **Pochopení dat (Data Understanding)** – Je důležité pochopit silné a slabé stránky daného datového souboru, protože jen zřídkakdy se stává, že existují data, která se přesně shodují s daným problémem. Historické údaje jsou často sbírány pro účely, které nesouvisejí se současným problémem. Důležité je vzít do úvahy náklady spojené se sběrem dat. Některá data jsou k dispozici zdarma, některá se dají koupit a jiná zase neexistují a budou vyžadovat podpůrné projekty k zajištění jejich sběru. Velmi důležitou částí této fáze je tedy odhad nákladů a přínosu každého datového zdroje a rozhodnutí, zda se vyplatí další investice.
3. **Příprava dat (Data Preparation)** – Současné technologie jsou výkonné, ale vyžadují určité požadavky na data k analýze. Z toho důvodu fáze přípravy dat často probíhá společně s fází pochopení, pro přeformátování dat do podoby, která přináší lepší výsledky. Typickým příkladem přípravy dat je převod dat do tabulkového formátu a odstranění chybějících hodnot. Některé techniky Data Miningu jsou určeny pouze pro kategorická data, jiné zase pouze pro číselné hodnoty. Číselné hodnoty je často nutné normalizovat nebo škálovat tak, aby byly srovnatelné.

4. **Modelování (Modeling)** – Ve fázi modelování jsou techniky Data Miningu aplikovány na data. Výstupem modelování je model nebo vzor zachycující pravidelnost v datech.
5. **Vyhodnocení modelu (Evaluation)** – Účelem této fáze je důsledně posoudit výsledky a získat jistotu, že jsou platné a spolehlivé, a teprve poté pokračovat dál. Po důkladné analýze kteréhokoliv souboru dat je možné nalézt určité vzory. Nicméně je důležité mít jistotu, že tyto vzory nejsou pouhé anomálie v datovém souboru. Stejně důležité je také posoudit, zda vytvořený model splňoval požadavky na původní obchodní cíle. Data Mining vznikl jako podpora při rozhodovacím procesu podniků, a tak je na něj nutné nahlížet pouze jako na část tohoto procesu.
6. **Nasazení modelu (Deployment)** – Nasazení modelu do obchodního procesu je poslední fází, u té se očekává určitá návratnost investice. Nasazení modelu do produkčního systému obvykle vyžaduje, aby byl model překódován pro vyšší rychlost a kompatibilitu s podnikovým systémem. To může být spojeno se signifikantními dodatečnými náklady. V mnoha případech je tým Data Science zodpovědný pouze za vytvoření funkčního prototypu společně s jeho vyhodnocením. Prototyp je následně předán vývojovému týmu, který model přizpůsobí produkčnímu systému. (Provost a Fawcett 2013)



Obrázek 6: Diagram standardizovaného procesu CRISP-DM
zdroj: (Provost a Fawcett 2013)

2.4 Rozdělení technik Data Miningu

Techniky Data Miningu se dělí na: **kontrolované (Supervised)**, známé také jako prediktivní nebo cílené, a **nekontrolované (Unsupervised)**, známé také jako popisné nebo necílené. (Cawley 2014)

2.4.1 Kontrolované techniky

Kontrolované techniky Data Miningu jsou vhodné, pokud je definovaná konkrétní cílová hodnota, kterou je potřeba u dat předpovědět. Cílové hodnoty mohou mít dvě nebo více možných výsledků, ale může se jednat i o spojitou číselnou hodnotu.

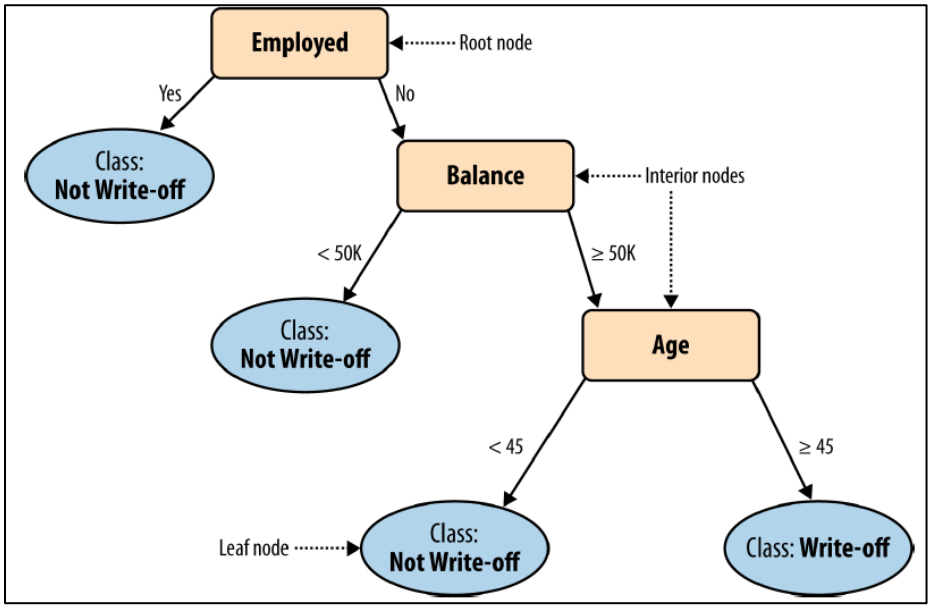
Kontrolované techniky se dále dělí na dvě kategorie algoritmů, a to na **klasifikaci (Classification)** a **regresi (Regression)**. V případě klasifikace je výstupní proměnnou určitá kategorie. U regrese je výstupní proměnnou hodnota. (GeeksforGeeks team 2021)

Algoritmy klasifikace

Klasifikace je proces hledání modelu (nebo funkce), který popisuje a rozlišuje třídy dat. Model se odvozuje na základě analýzy souboru trénovacích dat (tj. datových objektů, pro které jsou známé výsledné třídy). Model se používá k předpovídání tříd u objektů, u kterých je výsledná třída neznáma. (Han a Kamber 2012)

Odvozený model může být prezentován v různých formách, například jako klasifikační pravidla (tj. IF-THEN), rozhodovací strom (Decision Tree), matematický vzorec nebo neuronová síť (Neural Network). Níže bude blíže představena technika rozhodovacího stromu a neuronové sítě. (Han a Kamber 2012)

Rozhodovací strom je stromová struktura podobná diagramu (viz Obrázek 7), kde každý uzel (Node) obsahuje test na hodnotu atributu, každá větev (Branch) představuje výsledek testu, a listy (Leaf) stromu představují třídy nebo rozdělení tříd. Rozhodovací stromy jsou poměrně oblíbeným nástrojem, protože jsou lehké na porozumění a jednoduše převeditelné na klasifikační pravidla. (Han a Kamber 2012)



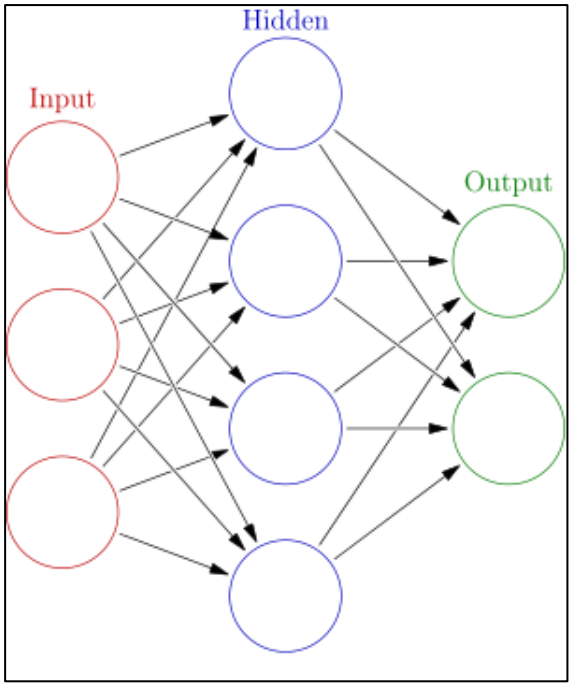
Obrázek 7: Jednoduchý rozhodovací strom

zdroj: Provost a Fawcett 2013

Neuronová síť, pokud se používá ke klasifikaci, je obvykle soubor výpočetních jednotek podobný neuronům s váženými spojeními mezi jednotkami. Neuronové sítě jsou používané pro klasifikace, kde jsou omezené znalosti o vztahu mezi vstupními a výstupními proměnnými. (Han a Kamber 2012)

Fungování neuronové sítě lze definovat ve čtyřech bodech (viz Obrázek 8):

1. Vstupní vrstva (Input) je tvořená uzly, které jednoduše přijímají vstupní hodnoty.
2. Signály mezi neurony jsou předávány prostřednictvím spojovacích článků. Na Obrázku 8 jsou znázorněné jako šipky mezi uzly. Každému spojovacímu článku je přiřazena váha, která násobí přenášený signál. (Han a Kamber 2012)
3. Každý uzel ve skryté (Hidden) vrstvě aplikuje aktivační funkci, která rozhoduje o tom, zda má být neuron aktivován, nebo ne. Mezi základní aktivační funkci patří například binární kroková funkce (Binary Step Function), ta závisí na prahové hodnotě, která rozhoduje o tom, zda má být neuron aktivován. Vstup přiváděný do aktivační funkce se porovnává s danou prahovou hodnotou; pokud je vstup větší než prahová hodnota, neuron je aktivován, jinak je deaktivován, což znamená, že jeho výstup není předán další skryté vrstvě.
4. Výstupní (Output) vrstva je poslední vrstva sítě, která přináší informace získané ze skryté vrstvy a poskytuje konečnou hodnotu. (Pragati 2022)



Obrázek 8: Ukázka třívrstvé neuronové sítě

zdroj: Sahil Chaudhary 2018

Algoritmy regrese

Nekontrolované techniky Data Miningu nepředpovídají cílovou hodnotu ani kategorii proměnné. Zaměřují se na vnitřní strukturu, vztahy a vzájemné propojení dat. Tyto modely se označují také jako deskriptivní. Cílem je seskupit nesetříděné informace podle podobností, vzorů a rozdílů bez předchozího tréninku dat. Mezi základní techniky patří shlukování (Clustering), asociační pravidla (Association Rules) nebo profilování (Profiling). (Provost a Fawcett 2013)

2.4.2 Nekontrolované techniky

Nekontrolované techniky Data Miningu nepředpovídají cílovou hodnotu ani kategorii proměnné. Zaměřují se na vnitřní strukturu, vztahy a vzájemné propojení dat. Tyto modely se označují také jako deskriptivní. Úkolem je seskupit nesetříděné informace podle podobností, vzorů a rozdílů bez předchozího tréninku dat. Mezi základní techniky patří shlukování (Clustering), asociační pravidla (Association Rules) nebo profilování (Profiling). (Provost a Fawcett 2013)

Shlukování

Shluková analýza identifikuje shluky obsažené v datech. Shluk (Cluster) je soubor datových objektů, které jsou si v určitém smyslu podobné. Kvalitní shluková analýza vytváří kvalitní shluky tím, že je podobnost mezi shluky nízká a podobnost uvnitř shluku vysoká. Jinak řečeno, datové objekty uvnitř shluku jsou si navzájem podobnější než objekty jiného shluku.

Shlukování může také sloužit jako užitečný krok před zpracováním dat k identifikaci homogenních skupin, na nichž lze postavit kontrolované modely. (Oracle Team 2022)

Často využívaný shlukovací algoritmus se nazývá **K-means Clustering**. Využívá metody, kdy je každý shluk reprezentován jeho středem neboli centroidem. V K-means jsou centroidy reprezentovány aritmetickými průměry hodnot na ose x a y. Písmeno „K“ značí počet shluků, které chce uživatel najít v souboru dat. (Provost a Fawcett 2013)

Asociační pravidla

Asociační modely zachycují společný výskyt položek nebo události ve velkých objemech dat. Často se používají pro analýzu tržního koše (market basket analysis). Analýza tržního koše zkoumá všechny položky dostupné v určitém médiu, například výrobky na pultech obchodů nebo v katalogu. Díky pokroku v technologii je nyní možné, aby organizace shromažďovaly a uchovávaly obrovské množství údajů o svých zákaznících, a na základě těchto dat upravovaly své služby a obchody. (Oracle Team 2022)

Profilování

Profilování se pokouší charakterizovat typické chování jednotlivce nebo skupiny. Častým využitím je například odhalování podvodů, kdy pomocí profilování model charakterizuje běžné chování a následně hledá případy, které se výrazně odchylují od normálního chování – zejména způsobem, který dříve svědčil o podvodu.

Profilování může ve své podstatě zahrnovat i metodu shlukování, pokud existují podskupiny populace s odlišným chováním. (Provost a Fawcett 2013)

2.5 Testování Data Miningových modelů

Jedny z nejdůležitějších základních pojmů při tvorbě Data Miningových modelů jsou přeučení modelu (Overfitting) a zobecnění (Generalization). Pokud je uživatel při hledání vzorů v datech dostatečně flexibilní, vždy nějaký najde. To ale neznamená, že daný vzor je smysluplný. (Provost a Fawcett 2013)

2.5.1 Zobecnění

Zobecnění je vlastnost modelu nebo procesu modelování, při kterém se model aplikuje na data, která nebyla použita k vytvoření modelu. Kvalita modelu se pozná až poté, co je vyzkoušen na celém datovém souboru, ze kterého pochází i testovací data[2].

[2] Testovací (validační) soubor dat (v anglické literatuře se často vyskytuje pojem Holdout Data) se používá k měření výkonu Data Miningového modelu.

2.5.2 Přeučení modelu

Přeučení modelu je tendence přizpůsobovat modely tréninkovým datům[3] na úkor zobecnění. Všechny techniky Data Miningu způsobují do určité míry přeučení modelu.

Obecné pravidlo tvrdí, že modely mají tendenci lépe predikovat výsledky na tréninkových datech s rostoucí složitostí modelu, ale nefungují tak dobře na nových (testovacích) datech. Tomuto pravidlu se říká kompromis mezi zkreslením a odchylkou (Bias-Variance Trade-Off).

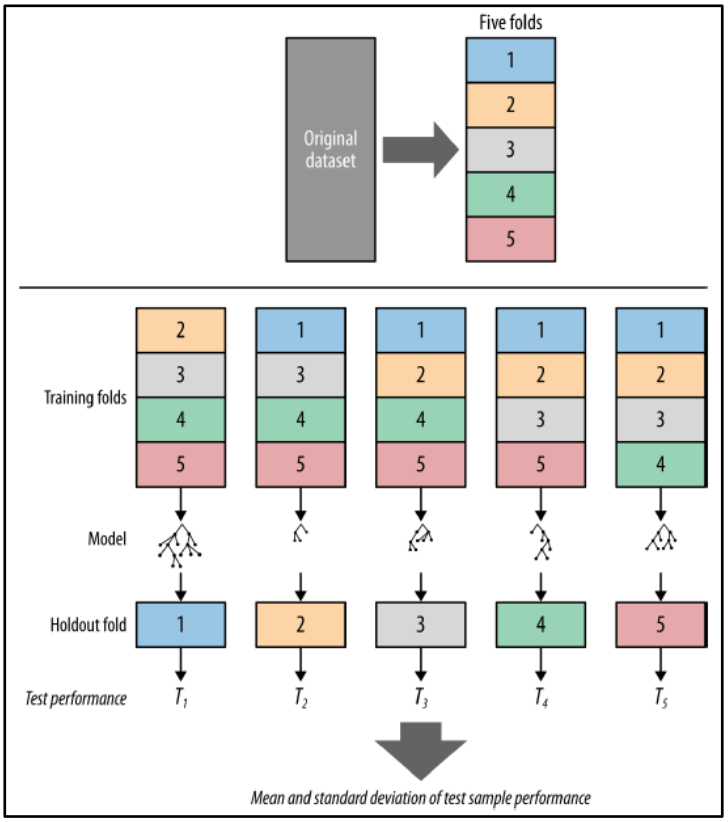
Nicméně existují techniky, jak přeučení modelu rozpoznat, a následně i minimalizovat. Jedná se například o techniku křížové validace (Cross Validation) nebo učících křivek (Learning Curves). (Handel 2021)

Křižová validace

Křížová validace je založena na principu rozdělení množiny dat na podmnožiny. Vždy je jedna z podmnožin testovací a zbylé množiny fungují jako tréninkové pro model. Model se vždy „natrénuje“ na tréninkových množinách a pomocí testovací otestuje přesnost a výkonnost modelu.

Proces se opakuje na všech množinách. Pokaždé je testovací množina jiná než v předchozím opakování. Výstupem na Obrázku 9 je pět různých výsledků přesnosti, které se dále použijí k výpočtu průměrné přesnosti a jejího rozptylu.

[3] Tréninkový soubor dat se používá k vytvoření Data Miningového modelu.



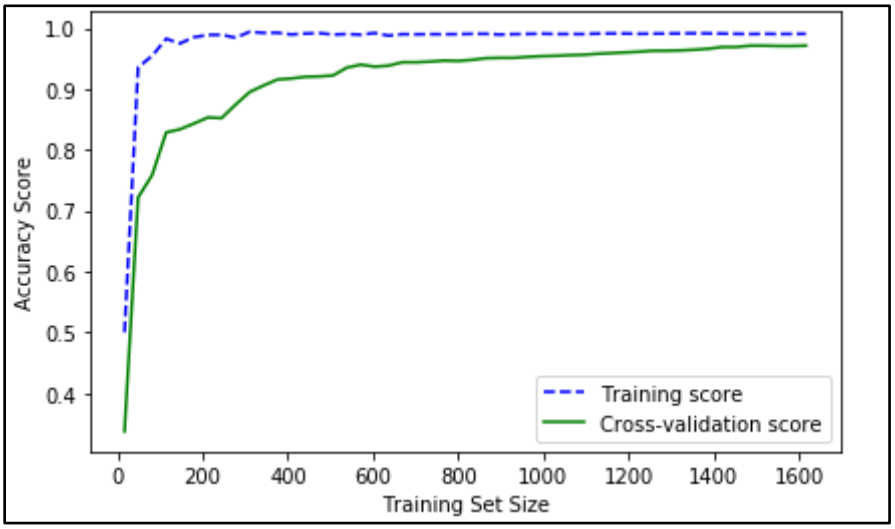
Obrázek 9: Ilustrace křížové validace

zdroj: Provost a Facett 2013

Křivky učení

Výkonnost modelu se obecně zlepšuje do určitého bodu s tím, jak je k dispozici více tréninkových dat. Graf výkonu modelu oproti procentu dat, které byly použity k tréninku modelu, se nazývá křivka učení.

Křivky učení mají většinou charakteristický tvar jak při učení, tak i testování modelu. Zpočátku jsou strmé, protože model najde nejdříve nejzřetelnější datové vzory. Postupně se křivka zplošťuje, jak je možné vidět na Obrázku 10. (Provost a Fawcett 2013)



Obrázek 10: Ukázka křivky učení

zdroj: Geeks for Geeks 2020

Během trénování modelu lze vyhodnocovat stav modelu v každém kroku tréninkového algoritmu. Lze jej vyhodnotit na trénovací množině dat, pro získání představy o tom, jak dobře se model „učí“. Následně je možné vyhodnotit model na testovací množině dat a zjistit, jak kvalitně model zobecňuje. (Brownlee 2019)

2.6 Data Mining v bankovním sektoru

Bankovní sektor si celosvětově prošel v posledních třech desetiletích obrovskými změnami ve způsobu podnikání. S využitím elektronického bankovnictví se zachycování transakčních údajů stalo jednodušším a současně objem dat vzrostl natolik, že není v lidských silách je pro získání užitečných informací analyzovat ručně. (Moin a Ahmed 2012)

Tento sektor nepochybně patří mezi nejbohatší, co se týče informací o klientech. Jedná se například o demografické a transakční údaje nebo vzorce používání platebních karet. Vzhledem k tomu, že bankovnictví patří do odvětví služeb, je velmi důležité udržovat efektivní řízení vztahů se zákazníky (Customer Relationship Management, CRM). (Moin a Ahmed 2012)

Mezi základní oblasti, ve kterých lze techniky Data Miningu použít, patří segmentace a ziskovost zákazníků, úvěrové hodnocení (Credit Scoring), odhalování podvodných transakcí, řízení hotovosti a optimalizace akciových portfolií.

Řízení rizik v bankovnictví

Data Mining je široce využíván pro řízení rizik. Banky potřebují informace, zda zákazník, se kterým jednají, je spolehlivý, nebo ne. Nabízením kreditních karet novým zákazníkům nebo schvalováním úvěrů se banky vystavují určitému riziku, že zákazník nedostojí svým závazkům.

Banky poskytují úvěry svým zákazníkům na základě ověření jejich schopnosti splácet, v souladu s tím banka zákazníkovi nabídne objem, úrokovou sazbu, dobu splácení a jiné parametry úvěru. Zákazníci, kteří jsou u banky po delší časový úsek a patří do vyšší příjmové skupiny, pravděpodobně získají úvěr snadněji. I přesto jsou banky při poskytování úvěrů obezřetné a vytvářejí si rezervy na situace, kdy zákazník úvěr nesplatí. (Moin a Ahmed 2012)

3 Očekávané úvěrové ztráty (ECL)

Tato kapitola je zaměřena na problematiku očekávaných úvěrových ztrát (ECL). Model ECL je koncepčně ukotven v mezinárodním účetním standardu pro finanční nástroje (IFRS 9).

V rámci kapitoly je prvně popsán úvod do problematiky ECL, z jakého důvodu je model vytvořen a jakým způsobem se očekávané úvěrové ztráty počítaly v minulosti. V kapitole je dále popsáno jaké parametry tvoří model ECL. Poslední část se týká očekávaných budoucích ekonomických podmínek (FLI), které musí být do výpočtu ECL zahrnuty.

3.1 Úvod do problematiky ECL

Světová finanční krize z roku 2008 zvýšila zájem bankovních regulátorů o přiměřenosti tvorby opravných položek v průběhu hospodářského cyklu. Tvorba opravných položek má zásadní význam pro odolnost bankovního sektoru v krizových dobách. Z toho důvodu byl vyvinut standard IFRS 9 s účinností od 1. 1. 2018. (Malonová a Tesařová 2020)

IFRS 9 nahradil mezinárodní účetní standard IAS 39 (Internation Accounting Standards). Ten byl po finanční krizi kritizován, a to hned z několika důvodů. Byl nekonzistentní se způsobem kterým instituce řídila své podnikání, rizika a pozdější identifikaci ztrát z úvěrů a pohledávek. Standard IAS 39 vycházel z předpokladu, že všechna finanční aktiva instituce budou splacena, a to až do okamžiku, kdy je zjištěn opak a vznikne ztráta.

Mezi hlavní témata, která standard IFRS 9 řeší, jsou klasifikace a oceňování finančních instrumentů, zajišťovací účetnictví (Hedge Accounting) a přehodnocování finančních aktiv (Impairment). (Machalec 2017)

IFRS 9 přinesl nový model očekávaných úvěrových ztrát ECL. ECL pro všechna finanční aktiva již od začátku zaúčtuje opravné položky ve výši dvanácti měsíčních očekávaných úvěrových ztrát. Pokud však dojde k významnému zvýšení úvěrového rizika (Significant Increase in Credit Risk, zkráceně SICR) od počátečního zaúčtování, zaúčtuje se opravná položka ve výši celoživotních očekávaných úvěrových ztrát. Podstatou ECL modelu je tedy předpovídat budoucí ztráty a předem je zachytit v účetnictví. Tímto dochází k pokrytí jednoho z nejvýraznějších nedostatků standardu IAS 39. (Holan 2017)

Opravné položky (ECL) jsou vypočteny dle vzorce (1):

$$ECL = PD * EAD * LGD * DF \quad (1)$$

Veličiny vstupující do daného vzorce jsou:

- PD (Propability of Default) představuje pravděpodobnost selhání,
- EAD (Exposure At Default) představuje expozici v čase selhání,
- LGD (Loss Given Default) představuje ztrátovost ze selhání,
- DF (Discount Factor) představuje diskontní faktor, kde diskontní sazba odpovídá hodnotě efektivní úrokové sazby k datu počátečního zaúčtování.

IFRS 9 všechny expozice rozděluje do tří rizikových kategorií, které určují odhad ECL (dvanáctiměsíční nebo celoživotní úvěrové ztráty). IFRS 9 je koncepční standard, tudíž k odhadu jednotlivých parametrů instituce přistupují různě. Mezi běžné přístupy patří využití predikčních modelů. Jedná se například o logistickou nebo Coxovu regresi. Do modelů vstupují interní a externí indikátory kreditního rizika, predikce makroekonomických ukazatelů a ekonomická situace dlužníka. (Machalec 2017)

3.2 Přiřazení rizikových úrovní

Jednou z klíčových součástí výpočtu opravných položek dle IFRS 9 je správné přiřazení úrovně (Stage) jednotlivým expozicím. Úrovně pro zařazení expozic v rámci modelu jsou následující:

1. **Úroveň 1 (Stage 1)** – Expozice, u nichž nedošlo k významnému zvýšení úvěrového rizika od prvotního zaúčtování.
2. **Úroveň 2 (Stage 2)** – Expozice, u nichž došlo k významnému zvýšení úvěrového rizika od prvotního zaúčtování.
3. **Úroveň 3 (Stage 3)** – Expozice úvěrově znehodnocené.

Pro expozice **v úrovni 1** instituce vykazuje pouze **dvanáctiměsíční ECL**, zatímco pro expozice **v úrovních 2 a 3** se vykazují **celoživotní ECL**. (Evropská komise 2016)

3.2.1 Zařazení do druhé rizikové úrovně

Pro zařazení do úrovně 2 instituce musí posoudit, zda došlo k významnému zvýšení úvěrového rizika od data prvotního zaúčtování. Vyhodnocení významného zvýšení úvěrového rizika je aplikováno na úrovni jednotlivých expozic.

Významný nárůst úvěrového rizika si každá instituce musí definovat sama. Běžnou praxí je například aktuální prodlení dlužníka nebo kvantitativní a kvalitativní informace, která instituce o dlužníkovi má. (Evropská komise 2016)

3.2.2 Zařazení do třetí rizikové úrovně

Jestliže je expozice zařazena do úrovně 3, je považována za expozici v selhání. Nicméně stále se pro expozici počítá celoživotní ECL. Podobně jako u zařazení do úrovně 2 je na instituci, jak nastaví pravidla pro zařazení do úrovně 3. Mezi běžnou praxi patří aktuální prodlení dlužníka nebo kvalitativní kritéria (například pokud s dlužníkem bylo zahájeno insolvenční řízení). (Evropská komise 2016)

3.2.3 Pohyb mezi jednotlivými úrovněmi

Pokud se prokáže, že u expozice již není významně zvýšené úvěrové riziko, může být expozice zařazena zpět do úrovně 1. Stejně tak pokud již netrvají důvody pro zařazení do úrovně 3, instituce může zařadit expozici do méně rizikové úrovně po uplynutí stanovené lhůty.

3.3 Pravděpodobnost selhání (PD)

Termín PD je definován jako pravděpodobnost, že u dlužníka dojde během určitého časového období (během 12 měsíců nebo do konce životnosti expozice) k selhání, dlužník tedy nebude schopen dostát svým smluvním závazkům. (Kašparovská 2006)

PD se definuje pro každou rizikovou úroveň zvlášť následujícím způsobem:

1. **Úroveň 1** – PD je pravděpodobnost, že k selhání dojde v následujících dvanácti měsících.

2. **Úroveň 2** – PD je pravděpodobnost, že k selhání dojde kdykoliv v průběhu plnění závazku dlužníka vůči instituci.

3. **Úroveň 3** – PD je automaticky rovna 100%. (Evropská komise, 2016)

3.4 Ztrátovost ze selhání (LGD)

Termín ztrátovost ze selhání (Loss Given Default, dále jen LGD) v rámci standardu IFRS 9 nemá exaktní definici. Nicméně bod 55 článku 4 Capital Requirements Regulation (dále jen CRR) definuje ztrátovost ze selhání jako: „*poměr mezi ztrátou z expozice z důvodu selhání protistrany a částkou dlužnou v okamžiku selhání*“. (Evropská komise 2016)

Běžnou praxí je stanovení LGD na základě jednoho ze dvou způsobů (na základě historických dat či regulatorně):

1. Stanovení LGD na základě historických dat – LGD stanovené na základě historických dat zohledňuje všechny minulé ztráty, ke kterým došlo v případě, že nastalo selhání. Na základě historických zkušeností se předpokládá hodnota LGD v budoucnu.

2. Stanovení LGD regulatorně – Regulatorní způsob stanovení LGD vychází z normativních předpisů přijatých nebo vydaných na úrovni Evropské Unie (CRR, Doporučení Evropského orgánu pro bankovnictví) a normativních předpisů, stanovisek, úředních sdělení a doporučení vydaných na národní úrovni. (Evropská komise 2016)

3.5 Hodnota expozice v čase selhání (EAD)

Termín hodnota expozice v čase selhání (Exposure at Default, dále jen EAD) představuje očekávanou hodnotu expozice v čase selhání.

V rámci standardu IFRS 9 neexistuje definice EAD, protože standard vysloveně nepožaduje modelování hodnoty expozice v čase nastání defaultu. Nicméně je velice důležité při výpočtu očekávaných úvěrových ztrát vědět, jak se daná expozice bude vyvíjet, především u expozic zařazených do druhé rizikové úrovně, kde selhání může nastat i několik let v budoucnu.

Pokud by se budoucí vývoj hodnoty expozice ignoroval, došlo by ke zkreslení hodnoty opravné položky, například z důvodu, že by se do úvahy nebralo snížení dané expozice v důsledku splátek.

Definice selhání

Stejně jako EAD, ani definice selhání není součástí standardu IFRS 9, ale zároveň se jedná o důležitý údaj, který má dopad na modelování hodnoty EAD.

Běžnou praxí je využití definice selhání z článku 178 CRR:

„Má se za to, že u konkrétního dlužníka došlo k selhání, pokud nastane jedna nebo obě situace:

a) instituce má za to, že dlužník pravděpodobně v plném rozsahu nesplatí své úvěrové závazky vůči instituci, jejímu mateřskému podniku či některému z jejích dceřiných podniků, aniž by instituce přistoupila ke krokům, jako je realizace zajištění;

b) některý podstatný úvěrový závazek dlužníka vůči instituci, jejímu mateřskému podniku či některému z jejích dceřiných podniků je více než 90 dní po splatnosti.“ (Evropská komise, 2016)

3.6 Očekávané budoucí informace (FLI)

Jedna z povinností IFRS 9 je do výpočtu PD zahrnout i očekávané budoucí ekonomické podmínky jak pro PD v následujících 12 měsících, tak i pro celoživotní PD. FLI představují rozšířený soubor informací, který zahrnuje úvěrové informace týkající se budoucího vývoje (makroekonomického vývoje). Zahrnutí FLI společně s historickými daty se považuje za vytvoření komplexních informací o úvěrovém riziku. (Open Risk Manual Team nedatováno)

Běžnou praxí predikce budoucí PD v České Republice je využití makroekonomických veličin, jejichž predikce jsou zveřejňovány například Českou národní bankou (ČNB) nebo Českým statistickým úřadem (ČSÚ).

Modelové scénáře FLI

IFRS 9 vyžaduje, aby účetní jednotka oceňovala ECL způsobem, který odráží nezkreslenou a pravděpodobností váženou částku, která je stanovena na základě vyhodnocení několika možných scénářů. Je to z důvodu, že v praxi existují nelineární vztahy mezi různými scénáři a s nimi spojenými úvěrovými ztrátami.

Nelineární vztah mezi scénáři a úvěrovými ztrátami lze zobrazit například na portfoliu hypotečních úvěrů na bydlení:

- Pokud se hodnota rezidenčních nemovitostí sníží o 10 %, pak se ECL zvýší pouze o 1 %, a to z důvodu významného přezajištění nemovitostí.
- Ale pokud se hodnota rezidenčních nemovitostí sníží o 20 %, pak se ECL zvýší o 10 %, protože podstatně více úvěrů se stane nedostatečně zajištěnými a utrpí ztráty. (PricewaterhouseCoopers GmbH 2017)

4 Tvorba modelu FLI

Tato kapitola je zaměřena na tvorbu modelu pro výpočet očekávaných budoucích ekonomických podmínek (FLI) na základě makroekonomických veličin inflace, nezaměstnanosti a hrubého domácího produktu.

V rámci kapitoly jsou prvně popsány makroekonomické veličiny, které vstupují do regresního modelu. Dále je popsán sběr vstupních dat a popis přípravy vysvětlujících proměnných a vysvětlované proměnné. V rámci kapitoly je popsána tvorba modelu a jeho testování. Poslední část se týká interpretace výsledků modelu.

V rámci kapitoly jsou prvně popsány makroekonomické veličiny, které vstupují do regresního modelu. Dále je popsán sběr vstupních dat a popis přípravy vysvětlujících proměnných a vysvětlované proměnné. V rámci kapitoly je popsána tvorba modelu a jeho testování. Poslední část se týká interpretace výsledků modelu.

Tento regresní model umožní z predikovaných vysvětlujících proměnných určit koeficient, kterým se vynásobí PD. Pokud je koeficient FLI větší než 1, tak se počítá, že se budoucí ekonomické podmínky zhorší, a tudíž je pravděpodobnost selhání klienta větší. Naopak pokud bude FLI menší než 1, tak se počítá, že budoucí ekonomické podmínky se zlepší a pravděpodobnost selhání klienta bude nižší.

4.1 Makroekonomické veličiny vstupující do modelu FLI

Do modelu pro výpočet FLI vstupují tyto makroekonomické veličiny:

1. Hrubý domácí produkt (dále jen HDP) – HDP je suma peněžní hodnoty statků a služeb nově vytvořených v daném období (obvykle jeden rok) na určitém území. (Český statistický úřad 2015)

2. Měnově politická inflace (dále jen inflace) – Inflace znamená obecně růst cenové hladiny, respektive jde o oslabení kupní síly měny oproti službám a statkům. Měnově politická inflace je základní inflace bez započtení dopadů změn nepřímých daní. (Česká národní banka nedatováno)

3. Nezaměstnanost – Nezaměstnanost reprezentuje skupinu lidí ve věku od 15-64 let, kteří nemají práci, ale aktivně ji hledají (například jsou zapsaní na úřadu práce a odpovídají na nabídky zaměstnání). (Česká národní banka nedatováno)

4. Klientské úvěry (výkonné a nevýkonné) podle odvětví (CZ-NACE) – Výkonné klientské úvěry jsou všechny úvěry, které banky považují za aktivní a klienti je splácejí. Nevýkonné klientské úvěry jsou ty, které banka vyhodnotila jako úvěry v selhání. (Česká národní banka nedatováno)

4.2 Vstupní data modelu FLI

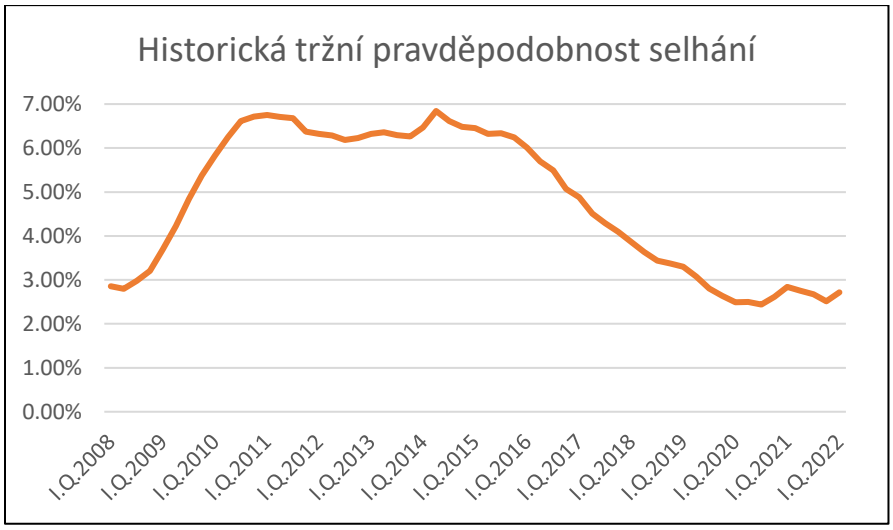
Historická data z období Q1 2008 až Q1 2022 jednotlivých veličin jsou čerpána ze systému časových řad ARAD, který je spravován ČNB. Vysvětlovanou proměnnou je tržní pravděpodobnost selhání a vysvětlující proměnné jsou HDP, nezaměstnanost a měnově politická inflace.

Systém časových řad ARAD

ARAD je veřejně dostupná databáze, kterou spravuje ČNB v rámci informačního servisu veřejnosti. Většina zveřejněných údajů je vytvořena statistickým zpracováním ČNB, nicméně část dat je přebrána z Českého statistického úřadu s jeho svolením. (Česká národní banka nedatováno)

4.2.1 Tržní pravděpodobnost selhání

Pro tržní pravděpodobnost selhání byla využita databáze výkonných a nevýkonných klientských úvěrů podle odvětví (CZ-NACE) a sekcí (Kč + cizí měna). Tato sestava ukazatelů je vykazována na měsíční bázi, pro účely vytvoření modelu a této bakalářské práce jsou však využívány čtvrtletní hodnoty. Pro získání čtvrtletních hodnot byl použit klasický aritmetický průměr ze tří po sobě jdoucích měsíců. Výsledná hodnota tržní pravděpodobnosti selhání je poměr mezi výkonnými a nevýkonnými expozicemi. V době sběru dat nebyla k dispozici data za Q1 2022. Tato hodnota byla vyvozena jako klasický aritmetický průměr předchozích tří let. Na Obrázku 11 je zobrazen historický vývoj tržní pravděpodobnosti selhání.

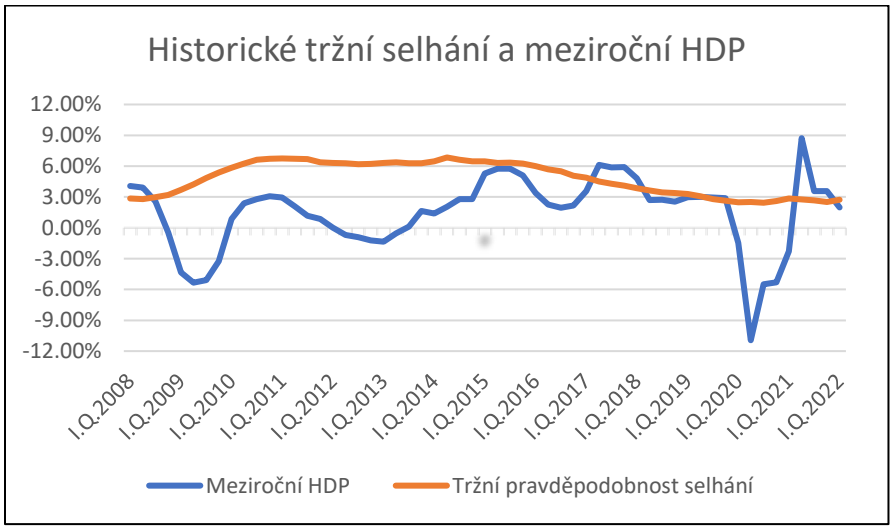


Obrázek 11: Historická tržní pravděpodobnost selhání v procentech (%)

zdroj: autor dle dat ČNB

4.2.2 Hrubý domácí produkt (HDP)

Pro HDP byla použita čtvrtletní data sezonně očištěna ve stálých cenách roku 2015. Tato sestava ukazatelů je vykazována na čtvrtletní bázi. Podobně jako v případě tržní pravděpodobnosti selhání, ani pro HDP nebyla v době sběru dat k dispozici data za Q1 2022. Tato hodnota byla opět vyvozena jako klasický aritmetický průměr předchozích tří let. Na Obrázku 12 je zobrazeno srovnání historického tržního selhání s meziročním vývojem HDP v procentech (%). Z Obrázku 12 je zřejmé, že se historické tržní selhání a meziroční HDP k sobě přibližovali až do vypuknutí celosvětové pandemie covid-19, kdy meziroční HDP skokově kleslo o více než 10 %.

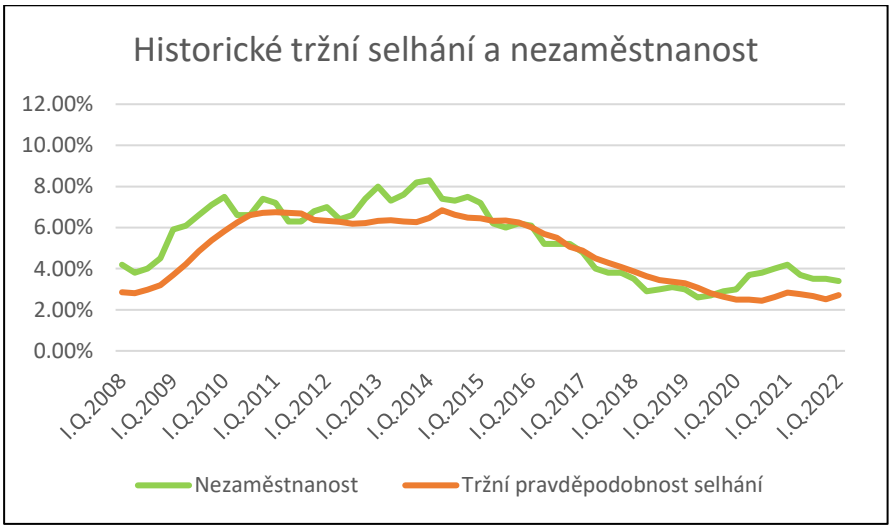


Obrázek 12: Vývoj historického tržního selhání a meziročního HDP v procentech (%)

zdroj: autor dle dat ČNB

4.2.3 Nezaměstnanost

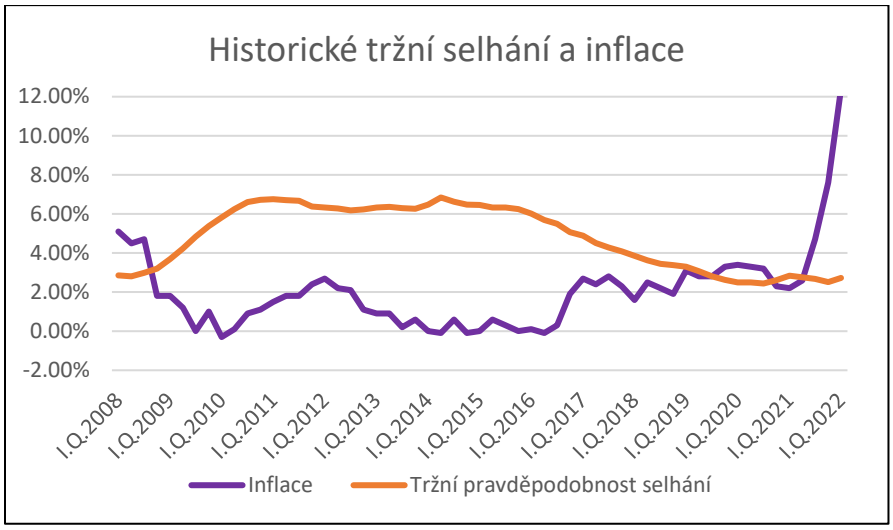
Pro nezaměstnanost byla použita data v procentech na měsíční bázi. Konkrétně se jedná o podíl nezaměstnaných osob na obyvatelstvo ve věku od 15 do 64 let. Pro získání čtvrtletních dat byl použit klasický aritmetický průměr ze tří po sobě jdoucích měsíců. Na Obrázku 13 je zobrazeno srovnání historického tržního selhání s nezaměstnaností v procentech (%). Z Obrázku 13 je zřejmé, že historické tržní selhání a nezaměstnanost jsou z historického hlediska korelující.



Obrázek 13: Vývoj historického tržního selhání a nezaměstnanosti v procentech (%)
zdroj: autor dle dat ČNB

4.2.4 Inflace

Pro inflaci byla použita data o měnově politické inflaci v procentech meziroční změny na měsíční bázi. Pro získání čtvrtletních dat inflace byl použit klasický aritmetický průměr ze tří po sobě jdoucích měsíců. Na Obrázku 14 je zobrazeno srovnání historického tržního selhání s inflací v procentech (%). Z Obrázku 14 je zřejmé, že historické tržní selhání a inflace se k sobě v průběhu historie přibližovali až do vypuknutí celosvětové pandemie covid-19, kdy inflace skokově vzrostla o přibližně 10 %.



Obrázek 14: Vývoj historického tržního selhání a inflace v procentech (%)

zdroj: autor dle dat ČNB

4.2.5 Zahrnutí očekávaných budoucích informací

Pro výpočet koeficientu FLI je nutné vytvořit predikci budoucích hodnot zvolených vysvětlujících proměnných. Tento model predikuje makroekonomické veličiny na čtyři roky dopředu. Predikce jsou získány dvěma způsoby. Prvních šest kvartálů je získáno z prognózy ČNB, která byla zveřejněna 3. 2. 2022, a je založena na datech dostupných k 21. 1. 2022. Zbylých deset kvartálů je odhadnuto lineární interpolací hodnot. Hodnoty jsou směřovány k průměru jednotlivých makroekonomických veličin. Tento Přístup byl zvolen z důvodu zahrnutí efektu hospodářského cyklu. Kvartální predikce jsou zobrazeny v Tabulce 1.

Tabulka 1: Predikce budoucích makroekonomických hodnot

Kvartál	Meziroční HDP	Nezaměstnanost	Inflace
Q2 2022	3.34 %	2.22 %	9.51 %
Q3 2022	2.30 %	2.42 %	8.17 %
Q4 2022	1.95 %	2.33 %	6.63 %
Q1 2023	2.60 %	2.42 %	3.03 %
Q2 2023	3.22 %	2.11 %	2.26 %
Q3 2023	3.72 %	2.32 %	1.99 %
Q4 2023	4.01 %	2.23 %	2.03 %
Q1 2024	3.73 %	2.32 %	2.03 %
Q2 2024	3.45 %	2.01 %	2.04 %
Q3 2024	3.17 %	2.21 %	2.04 %
Q4 2024	2.89 %	2.12 %	2.04 %
Q1 2025	2.60 %	2.76 %	2.04 %
Q2 2025	2.32 %	3.41 %	2.05 %
Q3 2025	2.04 %	4.05 %	2.05 %
Q4 2025	1.76 %	4.69 %	2.05 %
Q1 2026	1.48 %	5.33 %	2.05 %

zdroj: autor dle ČNB a pomocí lineární interpolace

4.3 Vícenásobná lineární regrese

Vícenásobná lineární regrese (Multiple Linear Regression, dále jen MLR) je statistická technika, která používá několik vysvětlujících proměnných k předpovědi výsledku vysvětlované proměnné. Cílem MLR je modelovat lineární vztah mezi vysvětlujícími (nezávislými) proměnnými a vysvětlovanými proměnnými (závislými). MLR je v podstatě rozšířením obyčejné regrese nejmenších čtverců (Ordinary Least Squares), protože zahrnuje více než jednu vysvětlující proměnnou. (Hayes 2022)

Z výše popsaných řad je vytvořena MLR, kde vysvětlovanou proměnnou je pravděpodobnost tržního selhání. Vysvětlující proměnné jsou vývoj meziročního HDP, nezaměstnanost a inflace. Finální model je zobrazen vzorcem (2):

$$historické\ PD = \alpha * meziroční\ HDP + \beta * nezaměstnanost + \gamma * inflace + \varepsilon, \quad (2)$$

kde:

- α značí koeficient u lineárního členu meziročního HDP,
- β značí koeficient u lineárního členu nezaměstnanosti,
- γ značí koeficient u lineárního členu inflace,
- ε značí chybovou proměnnou, tedy náhodnou proměnnou, která zachycuje ostatní faktory vstupující do lineárního vztahu. V modelu se s proměnnou ε dále nepočítá z důvodu její nízké důležitosti.

Po dosazení koeficientů do lineárního vztahu vychází model dle vzorce (3):

$$historické\ PD = 0.115373 * meziroční\ HDP + 0.76688 * nezaměstnanost - 0.120465 * inflace \quad (3)$$

4.3.1 Testování vhodnosti modelu

Odhadnutý model je nezbytné podrobit dalším testům, aby se ověřilo, že jejich odhad neporušuje žádný z předpokladů lineárního modelu.

Augmented Dickey-Fullerův test na stacionaritu

Stacionarita je vlastnost časové řady, kdy se průměr a rozptyl časové řady v čase nemění. Pomocí Augmented Dickey-Fullerova testu jsou postupně otestovány všechny vysvětlující proměnné:

- meziroční HDP – p-hodnota je 0.004393, tedy na 5% hladině významnosti je **vyvráceno**, že data mají jednotkový kořen a jsou nestacionární.
- nezaměstnanost – p-hodnota je 0.02245, tedy na 5% hladině významnosti je **vyvráceno**, že data mají jednotkový kořen a jsou nestacionární.
- inflace – p-hodnota je 0.007045, tedy na 5% hladině významnosti je **vyvráceno**, že data mají jednotkový kořen a jsou nestacionární.

Bresch-Paganův test o existenci homoskedasticity

Pro vytvoření robustního modelu je nutné vyvrátit heteroskedasticitu. Heteroskedasticita znamená zvyšující se rozptyl s rozšiřováním vzorku. Pomocí Breusch-Paganového testu o existenci homoskedasticity je předpoklad homoskedasticity (stálého rozptylu) přijat s p-hodnotou 0.5678. Předpokladem je tedy **homoskedasticita modelu.**

Kolmogorov-Smirnův test normálního rozdělení

Pomocí Kolmogorova-Smirnova testu normálního rozdělení je vyvrácena hypotéza o normálním rozdělení u všech vysvětlujících proměnných s p-hodnotami:

- meziroční HDP – p-hodnota je 7.987E-12
- nezaměstnanost – p-hodnota je 2.54E-13
- inflace – p-hodnota je 9.612E-13

Testem je na 5% hladině významnosti **vyvrácena** hypotéza, že data pochází z normálního rozdělení.

4.3.2 Interpretace modelu

Celkový F-test modelu

Na základě F-statistiky modelu (3;53) 127.5 s p-hodnotou menší než 2.20E-16 je zamítnuta hypotéza o nevýznamnosti modelu jako celku.

t-testy jednotlivých parametrů

t-test pro proměnnou meziroční HDP má p-hodnotu 3.09E-06, tedy je pro model na hladině významnosti 5% statisticky významný.
t-test pro proměnnou inflace má p-hodnotu 1.23E-02, tedy je pro model na hladině významnosti 5% statisticky významný.
t-test pro proměnnou nezaměstnanost má p-hodnotu 2.00E-16, tedy je pro model na hladině významnosti 5% statisticky významný.

V Tabulce 2 je zobrazen výstup modelu z programu RStudio.

HDP

Hodnota koeficientu proměnné meziroční HDP vysvětluje, že pokud se meziroční HDP zvýší o 1 % vyvolá to za jinak stejných podmínek **zvýšení** pravděpodobnosti tržního selhání v průměru o 0.115 %. Tato proměnná se ukázala jako významná na 1% hladině významnosti.

Inflace

Hodnota koeficientu proměnné inflace vysvětluje, že pokud se inflace zvýší o 1 %, vyvolá to za jinak stejných podmínek **snížení** pravděpodobnosti tržního selhání v průměru o 0.12 % na 5% hladině významnosti.

Nezaměstnanost

Hodnota koeficientu proměnné nezaměstnanost vysvětluje, že pokud se nezaměstnanost zvýší o 1 %, vyvolá to za jinak stejných podmínek **zvýšení** pravděpodobnosti tržního selhání v průměru o 0.767 % na 1% hladině významnosti.

Tabulka 2: Regresní model

Regresní model	Koeficient	Směrodatná odchylka	t-podíl	p-hodnota	Významnost
Meziroční HDP	0.115373	0.22119	5.216	3.09E-06	***
Inflace	-0.120465	0.046456	-2.593	1.23E-02	*
Nezaměstnanost	0.76688	0.056054	13.681	2.00E-16	***

zdroj: autor dle výstupu z RStudio

Koeficient determinace modelu

Koeficient determinace je statistická míra, která vyjadřuje podíl rozptylu závislé proměnné, který je vysvětlen nezávislou proměnnou nebo proměnnými v regresním modelu. Koeficient determinace tohoto modelu je 87.14 %. Tedy 87.14 % variability vysvětlované proměnné, lze vyjádřit pomocí nezávislých proměnných.

4.4 Interpretace koeficientu FLI

V rámci výpočtu koeficientu FLI je do výpočtu zahrnut normální, optimistický a pesimistický scénář vývoje PD. Pro odhad optimistického vývoje PD se využil 20. percentil veličin HDP a inflace a 80. percentil nezaměstnanosti. Obdobně pro odhad pesimistického vývoje PD se využil 80. percentil veličin HDP a inflace a 20. percentil nezaměstnanosti.

Výsledné tržní PD bylo odhadnuto pomocí popsaných scénářů a vah pro tyto scénáře určené. Pro normální scénář je váha 0,6, pro pesimistický a optimistický scénář je váha 0,2. V Tabulce 3 jsou zobrazeny výsledky normálního scénáře. Jednotlivé parametry znamenají:

1. Kvartál (predikované období) – Časová vzdálenost od současnosti v kvartálech.
2. PD – Inverzní funkcí určená hodnota PD za pomoci regresního modelu.
3. PD inv – Vyjadřuje určenou hodnotu PD + hodnotu PIT PD inverzní funkce k součtovému standardnímu normálnímu rozdělení.
4. Odhad – Hodnota PD inv ve standardním normálním rozdělení.
5. PIT PD – Aktuální PD (průměr posledních čtyř kvartálů)
6. Poměr – Poměr mezi PD prediction a PIT PD.
7. Váha – Váha dané predikce. Predikce v bližší budoucnosti mají výrazně větší váhu.
8. Vážený efekt – Vážený efekt dané predikce.

Tabulka 3: Výsledky koeficientu FLI v normálním scénáři

Kvartál	PD	PD_inv	Odhad	PIT PD	Poměr	Váha	Vážený efekt
Q2 2022	0.0094	-1.9233	2.72%	2.66%	102.14%	10	10.2145
Q3 2022	0.0114	-1.9120	2.79%	2.66%	104.85%	10	10.4853
Q4 2022	0.0121	-1.8998	2.87%	2.66%	107.81%	6	6.4685
Q1 2023	0.0179	-1.8819	2.99%	2.66%	112.30%	6	6.7383
Q2 2023	0.0172	-1.8647	3.11%	2.66%	116.76%	2	2.3352
Q3 2023	0.0197	-1.8450	3.25%	2.66%	122.03%	2	2.4405
Q4 2023	0.0192	-1.8258	3.39%	2.66%	127.37%	2	2.5475
Q1 2024	0.0196	-1.8061	3.54%	2.66%	133.03%	2	2.6606
Q2 2024	0.0169	-1.7892	3.68%	2.66%	138.06%	1	1.3806
Q3 2024	0.0182	-1.7711	3.83%	2.66%	143.64%	1	1.4364
Q4 2024	0.0171	-1.7539	3.97%	2.66%	149.07%	1	1.4907
Q1 2025	0.0217	-1.7322	4.16%	2.66%	156.19%	1	1.5619
Q2 2025	0.0263	-1.7058	4.40%	2.66%	165.19%	1	1.6519
Q3 2025	0.0309	-1.6749	4.70%	2.66%	176.29%	1	1.7629
Q4 2025	0.0355	-1.6394	5.06%	2.66%	189.77%	1	1.8977
Q1 2026	0.0401	-1.5992	5.49%	2.66%	205.96%	1	2.0596

zdroj: autor

Stejným způsobem se vypočítali hodnoty pro pesimistický i optimistický scénář. V souhrnné Tabulce 4 jsou zobrazeny jednotlivé predikce všech scénářů.

Tabulka 4: Výsledky normálního, pesimistického a optimistického scénáře

normální scénář				pesimistický scénář				optimistický scénář			
	PD	PIT PD	Ratio		PD	PIT PD	Ratio		PD	PIT PD	Ratio
Q1	2.72%	2.66%	102.14%	Q1	2.82%	2.66%	105.84%	Q1	2.73%	2.66%	102.58%
Q2	2.79%	2.66%	104.85%	Q2	2.96%	2.66%	111.15%	Q2	2.76%	2.66%	103.39%
Q3	2.87%	2.66%	107.81%	Q3	3.12%	2.66%	117.04%	Q3	2.78%	2.66%	104.37%
Q4	2.99%	2.66%	112.30%	Q4	3.32%	2.66%	124.58%	Q4	2.84%	2.66%	106.66%
Q5	3.11%	2.66%	116.76%	Q5	3.51%	2.66%	131.58%	Q5	2.90%	2.66%	108.83%
Q6	3.25%	2.66%	122.03%	Q6	3.73%	2.66%	139.80%	Q6	2.97%	2.66%	111.39%
Q7	3.39%	2.66%	127.37%	Q7	3.94%	2.66%	148.01%	Q7	3.07%	2.66%	115.32%
Q8	3.54%	2.66%	133.03%	Q8	4.18%	2.66%	156.97%	Q8	3.18%	2.66%	119.41%
Q9	3.68%	2.66%	138.06%	Q9	4.39%	2.66%	164.87%	Q9	3.30%	2.66%	123.66%
Q10	3.83%	2.66%	143.64%	Q10	4.64%	2.66%	174.01%	Q10	3.41%	2.66%	128.08%
Q11	3.97%	2.66%	149.07%	Q11	4.88%	2.66%	183.04%	Q11	3.54%	2.66%	132.66%
Q12	4.16%	2.66%	156.19%	Q12	5.15%	2.66%	193.42%	Q12	3.66%	2.66%	137.43%
Q13	4.40%	2.66%	165.19%	Q13	5.47%	2.66%	205.30%	Q13	3.79%	2.66%	142.37%
Q14	4.70%	2.66%	176.29%	Q14	5.83%	2.66%	218.83%	Q14	3.93%	2.66%	147.51%
Q15	5.06%	2.66%	189.77%	Q15	6.24%	2.66%	234.20%	Q15	4.07%	2.66%	152.85%
Q16	5.49%	2.66%	205.96%	Q16	6.70%	2.66%	251.61%	Q16	4.22%	2.66%	158.39%
PD 12M		105.96%		PD 12M		113.12%		PD 12M		103.94%	
PD Lifetime		119.03%		PD Lifetime		133.29%		PD Lifetime		111.64%	

zdroj: autor

Po vynásobení jednotlivými vahami vychází koeficient FLI pro **dvanáctiměsíční PD 1.06** a pro **celoživotní PD vychází 1.204**.

Dle FLI modelu se tedy očekává zhoršení ekonomických podmínek v následujícím roku i v následujících čtyřech letech. Expozicím v první rizikové úrovni, tedy těm expozicím kterým se počítá pouze dvanáctiměsíční ECL, bude upraven parametr PD FLI koeficientem 1.06. Expozicím v druhé rizikové úrovni, tedy těm expozicím kterým se počítá celoživotní ECL, bude upraven parametr PD FLI koeficientem 1.204. Expozice ve třetí rizikové úrovni mají parametr PD rovný 100 % a FLI koeficientem se neupravují.

4.5 Zhodnocení modelu

MLR model pro výpočet koeficientu FLI je sestaven na základě kvartálních dat z let 2008 až 2022. Na tři vysvětlující proměnné byla velikost vzorku 57 pozorování. V odborné literatuře se uvádí, že minimální velikost vzorku, který umožňuje dobrý odhad, pro jednu vysvětlující proměnnou v lineárně regresním modelu je 10 až 15 pozorování.(Babyak 2004)

Kvartální data byla vybrána z důvodu proměnné meziroční HDP, ke které neexistují spolehlivá data na měsíční bázi.

Data využitá v modelu byla otestována na stacionaritu (Augmented Dickey-Fullerův test), homoskedasticitu (Bresch-Paganův test), normální rozdělení (Kolmogorov-Smirnův test). Všechny testy vyšli dle požadavků kvality modelu, tedy data jsou stacionární, homoscedastická a nepochází z normálního rozdělení na 5% hladině významnosti.

Model byl otestován celkovým F-testem a jednotlivými t-testy na statistickou významnost proměnných. Model i všechny jeho proměnné jsou statisticky významné na 5% hladině významnosti.

Predikce ekonomických podmínek je získána na šest kvartálu dopředu z oficiální prognózy ČNB, zbylých deset kvartálů je dopočítáno lineární interpolací hodnot tak, aby vysvětlující proměnné dosáhli hodnot svého dlouhodobého průměru z důvodu zohlednění efektu hospodářského cyklu.

Závěr

Tato závěrečná práce se věnuje aplikaci technik dolování dat pro podporu rozhodování v moderních podnicích. Primárním cílem práce je vytvořit vícenásobný regresní model pro stanovení koeficientu FLI.

V teoretické části jsou představeny výstupy literární rešerše pro ekonometrickou analýzu v empirické části. Rešerše se prvně věnuje problematice Big Dat, kde se postupně definují Big Data, jejich vlastnosti, historický vývoj a pohled na očekávaný budoucí vývoj v této oblasti. Druhá část rešerše je zaměřena na problematiku technik Data Miningu, ve které je prvně definován pojem Data Mining a krátce představena historie. Dále jsou rozděleny základní typy úloh, které se pomocí Data Miningu řeší, a jaké mohou nastat problémy při tvorbě modelů, včetně efektivních řešení těchto problémů.

Podstatná část rešerše je věnována očekávaným úvěrovým ztrátám (ECL). Prvně je popsán úvod do problematiky ECL, z jakého důvodu je model vytvořen a jakým způsobem se očekávané úvěrové ztráty počítaly v minulosti. Dále je popsáno, jaké parametry tvoří model a v poslední části je pozornost věnovaná očekávaným budoucím ekonomickým podmínkám (FLI), které výpočtu ECL musí být zahrnuty.

Primární cíl bakalářské práce je splněn, respektive je vytvořen vícenásobný regresní model, který vysvětluje vliv HDP, nezaměstnanosti a inflace na vysvětlovanou proměnnou, kterou je tržní pravděpodobnost selhání.

Ekonometrický model provedený vícenásobnou regresní analýzou na historických datech z let 2008 až 2022 zahrnuje meziroční růst HDP v procentech (%), nezaměstnanost a měnově politickou inflaci. Všechna data jsou čerpána ze systému časových řad ARAD, který spravuje ČNB.

V závěrečné části empirické části je vytvořený model aplikovaný na výpočet koeficientu FLI. Koeficient FLI je vypočten na základě čtyřleté predikce budoucích hodnot zvolených vysvětlujících proměnných.

Expozicím v první rizikové úrovni je upraven parametr PD FLI koeficientem 1.06 a expozicím v druhé rizikové úrovni je upraven parametr PD FLI koeficientem 1.204.

Na základě vytvořeného modelu nebo jemu podobnému se banky v reálném prostředí rozhodují o stanovení koeficientu FLI. Model tedy lze využít jako základ pro výpočet koeficientu FLI nebo jako komparativní model. Dle vytvořeného modelu se tedy očekává zhoršení ekonomických podmínek v následujícím roku i v následujících čtyřech letech.

Jako hlavní možnost návaznosti na tuto práci lze označit detailní sledování vysvětlujících proměnných a aktualizování regresního modelu na jejich základě. Další možností je vytvořit podobné modely s jinými vysvětlujícími proměnnými a sledovat a poměřovat jejich výkonnost.

Například místo proměnné HDP, která je k dispozici pouze na čtvrtletní bázi, by bylo možné využít jinou makroekonomickou veličinu (například index spotřebitelských cen (CPI)), která je dostupná na bázi měsíční, a tím by se velikost vzorku pozorování výrazně navýšila.

V kontextu dnešní doby, ve které je svět stále silně ovlivněn celosvětovou pandemií covid-19 a konfliktem na Ukrajině, který vyeskaloval do nepředvídaného rozsahu, není jisté, zda jsou predikované makroekonomické hodnoty směrodatné. Situace ve světě je nestabilní a predikce budoucího vývoje makroekonomiky je obtížná, až nemožná.

Seznam použité literatury

ADILIN, Beatrice, 2021. ALL ABOUT THE BASICS OF BIG DATA: HISTORY, TYPES AND APPLICATIONS. *Analytics Insight* [online] [vid. 2022-03-30]. Dostupné z: https://www.analyticsinsight.net/all-about-the-basics-of-big-data-history-types-and-applications

APACHE HADOOP TEAM, 2022. HDFS Users Guide. *Apache Hadoop* [online] [vid. 2022-04-14]. Dostupné z: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html

BABYAK, Michael A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine* [online]. **66**(3), 411–421. ISSN 1534-7796. Dostupné z: doi:10.1097/01.psy.0000127692.23278.a9

BIG DATA FRAMEWORK TEAM, 2019. Where does 'Big Data' come from? *Big Data Framework* [online] [vid. 2022-03-30]. Dostupné z: https://www.bigdataframework.org/short-history-of-big-data

BROWNLEE, Jason, 2019. *How to use Learning Curves to Diagnose Machine Learning Model Performance* [online] [vid. 2022-04-17]. Dostupné z: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance

CAWLEY, Keith, 2014. WHEN TO USE SUPERVISED AND UNSUPERVISED DATA MINING. *CloudTweaks* [online] [vid. 2022-04-16]. Dostupné z: https://cloudtweaks.com/2014/09/supervised-unsupervised-data-mining

ČESKÁ NÁRODNÍ BANKA, nedatováno. ARAD. *ARAD* [online] [vid. 2022a-04-24]. Dostupné z: https://www.cnb.cz/docs/ARADY/HTML/index.htm

ČESKÁ NÁRODNÍ BANKA, nedatováno. Co to je měnověpolitická inflace? *Česká národní banka* [online] [vid. 2022b-04-24]. Dostupné z: https://www.cnb.cz/cs/casto-kladene-dotazy/Co-to-je-menovepoliticka-inflace

ČESKÁ NÁRODNÍ BANKA, nedatováno. *Metodický list - Nezaměstnanost* [online]. nedatováno. [vid. 2022c-04-24]. Dostupné z: https://www.cnb.cz/docs/ARADY/MET_LIST/nez_cs.pdf

ČESKÁ NÁRODNÍ BANKA, nedatováno. Metodický list - Úvěry klientské. *Česká národní banka* [online] [vid. 2022d-02-24]. Dostupné z: https://www.cnb.cz/docs/ARADY/MET_LIST/tuvob_cs.pdf

ČESKÝ STATISTICKÝ ÚŘAD, 2015. Hrubý domácí produkt (HDP) - Metodika. *Český Statistický Úřad* [online] [vid. 2022-04-24]. Dostupné z: https://www.czso.cz/csu/czso/hruby_domaci_produkt_-hdp-

DIÁZ, Alba, 2020. THE FOUR V'S OF BIG DATA. *Open Sistemas* [online] [vid. 2022-03-28]. Dostupné z: https://opensistemas.com/en/the-four-vs-of-big-data

EVROPSKÁ KOMISE, 2016. *Mezinárodní standard účetního výkaznictví (IFRS) 9 Finanční nástroje*. 22. listopad 2016.

GEEKSFORGEEKS TEAM, 2021. Supervised and Unsupervised learning. *Geeks for Geeks* [online] [vid. 2022-04-16]. Dostupné z: https://www.geeksforgeeks.org/supervised-unsupervised-learning

HAN, Jiawei a Micheline KAMBER, 2012. *Data mining: concepts and techniques*. 3rd ed. Burlington, MA: Elsevier. ISBN 978-0-12-381479-1.

HANDEL, Andreas, 2021. Model Performance and Overfitting. *Andreas Handel* [online] [vid. 2022-04-22]. Dostupné z: https://andreashandel.github.io/MADAcourse/Model_Evaluation_Overfitting.html

HAYES, Adam, 2022. Multiple Linear Regression (MLR). *Investopedia* [online] [vid. 2022-05-01]. Dostupné z: https://www.investopedia.com/terms/m/mlr.asp

HOLAN, P, 2017. Připravujete se na nový standard IFRS 9, i když nejste finanční institucí? *Auditor*. **24**(8), 8–9.

KAŠPAROVSKÁ, Vlasta, 2006. *Řízení obchodních bank: vybrané kapitoly*. V Praze: C.H. Beck. ISBN 978-80-7179-381-6.

LI, Ray, 2016. History of Data Mining. *KD nuggets* [online] [vid. 2022-04-14]. Dostupné z: https://www.kdnuggets.com/2016/06/rayli-history-data-mining.html

MACHALEC, Milan, 2017. IFRS 9 jako klasická dataminingová úloha. *System Online* [online] [vid. 2022-04-17]. Dostupné z: https://www.systemonline.cz/business-intelligence/ifrs-9-jako-klasicka-dataminingova-uloha.htm?mobilelayout=false

MALONOVÁ, Simona a Žaneta TESAŘOVÁ, 2020. Úvěrové ztráty a opravné položky bank v průběhu hospodářského cyklu: implikace pro IFRS 9. *ČNB* [online] [vid. 2022-04-17]. Dostupné z: https://www.cnb.cz/cs/o_cnb/cnblog/Uverove-ztraty-a-opravne-polozky-bank-v-prubehu-hospodarskeho-cyklu-implikace-pro-IFRS-9

MCARDLE, Gavin a Rob KITCHIN, 2016. What makes Big Data, Big Data? *SAGE journals* [online] [vid. 2022-03-27]. Dostupné z: https://journals.sagepub.com/doi/10.1177/2053951716631130

MOIN, Kazi Imran a Qazi Baseer AHMED, 2012. Use of Data Mining in Banking. / *International Journal of Engineering Research and Applications (IJERA)*. **2012**(2), 738–742. ISSN 2248-9622.

NEO, Benedict, 2020. Big Data: Its Benefits, Challenges, and Future. *Towards Data Science* [online] [vid. 2022-03-30]. Dostupné z: https://towardsdatascience.com/big-data-its-benefits-challenges-and-future-6fddd69ab927

OPEN RISK MANUAL TEAM, nedatováno. Forward-Looking Information. *Open Risk Manual* [online] [vid. 2022-04-23]. Dostupné z: https://www.openriskmanual.org/wiki/Forward-Looking_Information

ORACLE TEAM, 2022a. Co jsou big data? *Oracle* [online] [vid. 2022-03-27]. Dostupné z: https://www.oracle.com/cz/big-data/what-is-big-data

ORACLE TEAM, 2022b. Unsupervised Data Mining. *Oracle* [online] [vid. 2022-04-17]. Dostupné z: https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/4descriptive.htm

PHILLIPS, Andres, 2021. A history and timeline of big data. *WhatIs* [online] [vid. 2022-03-30]. Dostupné z: https://whatis.techtarget.com/feature/A-history-and-timeline-of-big-data

PICKELL, Devin, 2018. What is Big Data? A Complete Guide. *G2* [online] [vid. 2022-03-30]. Dostupné z: https://www.g2.com/articles/big-data

PRAGATI, Baheti, 2022. 12 Types of Neural Network Activation Functions: How to Choose? *V7 labs* [online] [vid. 2022-04-22]. Dostupné z: https://www.v7labs.com/blog/neural-networks-activation-functions

PRICEWATERHOUSECOOPERS GMBH, 2017. *In depth IFRS 9 impairment: how to include multiple forward-looking scenarios*. 1. srpen 2017. B.m.: PricewaterhouseCoopers GmbH.

PROQUEST. 2021. *Databáze článků ProQuest* [online]. Ann Arbor, MI, USA: ProQuest. [vid. 2021-09-18]. Dostupné z: http://knihovna.tul.cz/

PROVOST, Foster a Tom FAWCETT, 2013. *Data science for business: what you need to know about data mining and data-analytic thinking*. 1. ed., 2. release. Beijing Köln: O'Reilly. ISBN 978-1-4493-6132-7.

SAS TEAM, 2015. Big Data - What it is and why it matters. *SAS* [online] [vid. 2022-03-28]. Dostupné z: https://www.sas.com/cs_cz/insights/big-data/what-is-big-data.html

SKLENÁK, Vilém, 2001. *Data, informace, znalosti a Internet*. Praha: C.H. Beck. ISBN 978-80-7179-409-7.

SLÁNSKÝ, David, 2018. *Data and analytics for the 21st century*. ISBN 978-80-88260-25-7.

SMARI, Waleed W a International Conference on Collaboration Technologies and Systems (CTS) ANNUAL IEEE COMPUTER CONFERENCE International Symposium on Big Data and Data Analytics in Collaboration (BDDAC), International Workshop on Collaborative Mobile Systems and Sensors Networks (CMSSN), International Workshop on E-Transactions Systems (ETS), International Symposium on Collaboration, Social Computing, New Media, and Networks (SoMNet), International Symposium on Security in Collaboration Technologies and Systems (SECOTS), International Workshop on Collaborative Robots and Human Robot Interaction (CR-HRI), International Workshop on Collaborations in Emergency Response and Disaster Management (ERDM), International Workshop on Collaboration and Gaming (CoGames), International Workshop on Collaboration Technologies and Systems in Healthcare and Biomedical Fields (CoHeB), 2013. *International Conference on Collaboration Technologies and Systems (CTS), 2013 20-24 May 2013, Sheraton San Diego Hotel & Marina, San Diego, California, USA ; [including symposia and workshops* [online] [vid. 2022-03-27]. ISBN 978-1-4673-6403-4. Dostupné z: http://ieeexplore.ieee.org/servlet/opac?punumber=6558543

STATISTA RESEARCH DEPARTMENT, 2022. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. *Statista* [online] [vid. 2022-04-14]. Dostupné z: https://www.statista.com/statistics/871513/worldwide-data-created

VAILSHERY, Lionel Sujay, 2021. Internet of Things (IoT) and non-IoT active device connections worldwide from 2010 to 2025. *Statista* [online]. Dostupné z: https://www.statista.com/statistics/1101442/iot-number-of-connected-devices-worldwide