Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informačních technologií

Dolování dat z vybrané datové sady pro predikci podvodných pracovních nabídek

Diplomová práce

Autor: Sharon Moscato
Studijní obor: Informační management

Vedoucí práce: Ing. Tereza Otčenášková, BA, Ph.D.

Konzultant práce: RNDr. Josef Dolejš, Ph.D.

Hradec Králové Srpen 2023

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracovala samostatně a s použitím uvedené literatury.

V Hradci Králové dne 1.9.2023 Sharon Moscato

Poděkování:

Děkuji vedoucí diplomové práce Ing. Tereze Otčenáškové, BA, Ph.D. za metodické vedení práce a přínosné připomínky. Také bych ráda vyjádřila poděkování mému konzultantovi RNDr. Josefovi Dolejši, Ph.D. za jeho odborné připomínky.

Anotace

Diplomová práce se zabývá dobýváním znalostí z databází s důrazem na vývoj predikčního modelu pro identifikaci podvodných pracovních nabídek. V práci je analyzována datová sada obsahující pracovní nabídky kategorizované jako podvodné či pravdivé. Primárním cílem je navrhnout model schopný efektivně rozlišovat mezi těmito kategoriemi nabídek. Dílčí otázky práce se věnují analýze rozložení podvodných nabídek v různých průmyslových odvětvích a zkoumání jazykových vzorů v inzerátech. K analýze byla použita metodika CRISP-DM, zatímco pro samotné predikční modely byly využity metody pro dolování dat, konkrétně logistická regrese a náhodný les. Výstupem práce jsou odpovědi na dílčí analytické otázky a dva predikční modely, u kterých byla následně porovnána přesnost jejich predikce. Celkově práce poskytuje detailní pohled na praktické využití metod dobývání znalostí z databází na reálné datové sadě.

Klíčová slova

CRISP-DM, Data, Dobývání znalostí z databází, Dolování dat, Logistická regrese, Náhodný les, Predikce, Python

Annotation

Title: Data mining from a selected dataset for predicting fraudulent job offers

The thesis deals with knowledge discovery in databases with emphasis on the development of a prediction model for identifying fraudulent job offers. The thesis analyzes a dataset containing job offers categorized as fraudulent or not. The primary objective is to design a model capable of effectively distinguishing between these jobs offer categories. Sub-questions of the thesis address the distribution analysis of fraudulent job offers across various industrial sectors and the examination of linguistic patterns in advertisements. The CRISP-DM methodology was used for the analysis, while data mining methods, namely logistic regression, and random forest, were employed for the predictive models. The thesis outputs include answers to the analytical sub-questions and two predictive models, with a subsequent comparison of their predictive accuracy. Overall, the thesis provides a detailed perspective of the practical application of knowledge discovery methods on a real dataset.

Keywords

CRISP-DM, Data, Data Mining, Knowledge Discovery in Database, Logistic Regression, Prediction, Python, Random Forest

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

V posledních dekádách jsme byli svědky nebývalého rozmachu technologií, což mělo za následek expanzi mnoha podniků do online sféry. Díky tomu dnes můžeme získávat téměř vše, co potřebujeme, z pohodlí našich domovů. Toto pohodlí je však spojeno s rozsáhlým zpracováním dat o spotřebitelích. Dnes nejde jen o jednoduché informace, ale o obrovské množství dat, která mohou poskytnout hluboký vhled do chování, preferencí a potřeb zákazníků.

Historicky byly finanční instituce prvními, kdo rozpoznal potenciál těchto dat a začal je využívat k získání konkurenční výhody. Postupem času, s rostoucí dostupností nástrojů pro analýzu dat, se tento trend rozšířil do širokého spektra odvětví. Od sektoru maloobchodu po komplexní oblast zdravotní péče se datová analytika transformovala v nezbytný nástroj, umožňující podnikům hlubší pochopení preferencí, potřeb a chování svých zákazníků na trhu.

Přestože se tato revoluce v oblasti dat zdá být pozitivním krokem vpřed, neobejde se bez výzev a úskalí. Ačkoliv objem akumulovaných dat prudce narůstá, analytické studie odhalují, že z celkového množství těchto informací je podrobně prozkoumáno pouze 12 %. [1] To naznačuje poměrně velký nevyužitý potenciál, ale také upozorňuje na komplexní problémy spojené s efektivním zpracováním a analýzou těchto dat.

V reakci na tyto výzvy vstupuje do popředí proces dobývání znalostí z databází. Jednou z hlavních částí tohoto procesu je fáze dolování dat, ve které se identifikují vzory a trendy v rámci velkých datových sad. Tato oblast se neustále rozvíjí, a to díky rozšiřujícím se nástrojům a technologiím, které umožňují hlubší analýzu a lepší porozumění uloženým informacím. Důležité je, umět tyto technologie a procesy využít, a aplikovat je na konkrétní a relevantní problémy ve skutečném světě. Díky těmto nástrojům nejenže můžeme efektivněji optimalizovat interní procesy a lépe reagovat na potřeby zákazníků, ale také je lze využít k ochraně před falešnými a zavádějícími informacemi. V digitálním světě, kde falešné informace mohou mít vážné důsledky, vstupuje do popředí prediktivní analytika jako klíčový nástroj pro odhalení potenciálních nesrovnalostí a zajištění bezpečnosti a důvěryhodnosti pro koncové uživatele.

Tato diplomová práce se zabývá tématem dobývání dat na reálném datovém souboru s cílem predikce, zda jsou pracovní nabídky podvodné či pravdivé. Pro účely této práce byl vybrán datový soubor obsahující informace o pracovních nabídkách inzerovaných na americkém trhu, který byl získán z veřejně dostupného webového portálu Keggle.com. Práce je rozdělena do dvou hlavních částí – teoretické a praktické. Nejprve jsou definovány pojmy v oblasti dobývání znalostí z databází, které jsou nezbytné pro vypracování praktické části. Přesněji je popsán proces dobývání dat z databází a jsou vysvětleny související termíny, jako je například data mining, který je součástí tohoto procesu, a má za cíl odhalit netriviální souvislosti v analyzovaném datovém souboru.

Dále je představena metodika CRISP-DM, kterou Rogalewicz a Sika označují jako jednu z nejpoužívanější v oblasti dobývání dat, podle které byla vypracována praktická část práce. [2] Na závěr je stručně představen programovací jazyk Python a jeho knihovny, které byly využity při vytváření predikčního modelu.

Praktická část práce postupuje podle jednotlivých fází CRISP-DM, které jsou aplikovány na vybraný datový soubor. Cílem je získat odpovědi na předem definované analytické otázky a vytvořit predikční model.

Práce poskytuje nejen náhled do problematiky dobývání dat z databází, ale také ukazuje jejich praktické využití na reálném datovém souboru. Predikční model pro odhalování podvodných pracovních nabídek představuje užitečný nástroj, který je možné implementovat do webových portálů nabízejících pracovní příležitosti. Tím by se zvýšila kvalita těchto portálů a jejich věrohodnost.

2 Cíl práce

Cílem této diplomové práce je prozkoumat problematiku dobývání znalostí z databází a demonstrovat aplikaci predikčního modelu na reálných datech. Práce se zaměřuje na analýzu datového souboru obsahujícího pracovní nabídky, které jsou rozděleny do dvou kategorií: podvodné a pravdivé. Hlavním cílem je vytvořit predikční model, který bude schopen identifikovat podvodné nabídky práce. Mezi dílčí cíle této práce patří zodpovězení následujících analytických otázek:

- Jaký průmysl vykazuje největší počet podvodných pracovních nabídek?
- Která pracovní pozice je nejčastěji spojována s podvodnými nabídkami?
- Která slova se vyskytují nejčastěji v textech podvodných a nepodvodných pracovních nabídek?

Pro dosažení těchto cílů bude provedena podrobná analýza datového souboru pracovních nabídek s použitím metodiky CRISP-DM a metody pro dolování dat. Výsledkem práce bude model, který bude schopen klasifikovat pracovní nabídky jako pravdivé nebo podvodné s vysokou přesností. Diplomová práce poskytne náhled do problematiky dobývání znalostí z databází včetně jeho praktického využití na reálném datovém souboru.

3 Metodika zpracování

V rámci této diplomové práce byla použita kombinace teoretického a praktického přístupu pro dosažení stanovených cílů. Nejprve byla v teoretické části práce vymezena východiska na základě relevantní odborné literatury, která je nezbytná k pochopení problematiky dolování dat z databází. Odborná literatura byla čerpána především z knihovních zdrojů s důrazem na její relevanci pro danou práci. Vědecké články a studie byly získány prostřednictvím renomovaných databází, jako jsou Science Direct, Web of Science a ResearchGate. Kromě těchto databází byly pro komplexnost výzkumu využity také různé online zdroje, které poskytovaly aktuální informace a doplňkové materiály pro zpracování prediktivního modelu.

Praktická část práce se zaměřuje na vytvoření prediktivního modelu pomocí programovacího jazyka Python a jeho knihoven. Volba tohoto jazyka byla podmíněna skutečností, že Python je v současnosti považován za jeden z nejpoužívanějších programovacích jazyků nejen v oblasti datové analýzy a strojového učení. [3] Díky jeho významnému postavení v datové vědě je Python doprovázen bohatou dokumentací, což představuje značnou výhodu pro zpracování praktické části práce. Aktivní komunita kolem Pythonu zároveň umožňuje na různých platformách dohledávat řešení k specifickým výzvám, které se mohou v takových pracích objevit. Kromě toho nabízí širokou škálu knihoven, které usnadňují práci s daty, například Pandas, NumPy, Scikit-learn a další.

Pro tvorbu prediktivního modelu byl využit datový soubor dostupný na webové stránce Kaggle.com. Kaggle je online komunitní platforma pro datové vědce a nadšence do strojového učení. Platforma umožňuje uživatelům spolupracovat, vyhledávat a publikovat datové sady a soutěžit v řešení různých výzev v oblasti datové vědy. Zvolený datový soubor měl velikost přibližně 50 MB, což představovalo přijatelný vzorek dat pro vytvoření prediktivního modelu. Výběr této konkrétní datové sady byl dále motivován jeho kvalitou, dostupností a relevancí vzhledem k výzkumným cílům diplomové práce. V procesu vytváření prediktivního modelu byly uplatněny teoretické znalosti z oblasti datového dolování, zejména metodika CRISP-DM. Důvody pro volbu této metodiky zahrnují

jeho systematické fáze, které zajišťují komplexní pokrytí projektů datového dolování. Díky jeho iterativní povaze umožňuje metodika CRISP-DM průběžné hodnocení a optimalizaci modelu v průběhu jeho vývoje. Pro tvorbu modelu jsou použity vhodné Python knihovny, které poskytují potřebnou funkčnost a nástroje pro analýzu a zpracování dat. Predikční model spolu s analýzou dat, které byly zpracovány v Jupyter notebooku, jsou uloženy na přiloženém CD k této diplomové práci a také k dispozici online na GitHubu pod následujícím odkazem: https://github.com/SharonMoscato/fake-job-posting-project (ke dni 29.08.2023).

4 Teoretická část

Teoretická část diplomové práce se zabývá vymezením pojmů, které jsou nezbytné pro zpracování praktické části. Nejprve poskytuje čtenáři informace o pojmu dobývání znalostí z databází a jeho často zaměňované terminologii dolování dat. Dále představuje populární metodiku dolování dat CRISP-DM včetně popisu jednotlivých fází, která bude později využita v praktické části práce. Následně se zaměřuje na hlavní rozdělení prediktivních a popisných metod pro dolování dat a popisuje nejčastěji používané druhy metod. V neposlední řadě přibližuje programovací jazyk Python a jeho knihovny, které jsou nezbytné pro zpracování prediktivního modelu v praktické části.

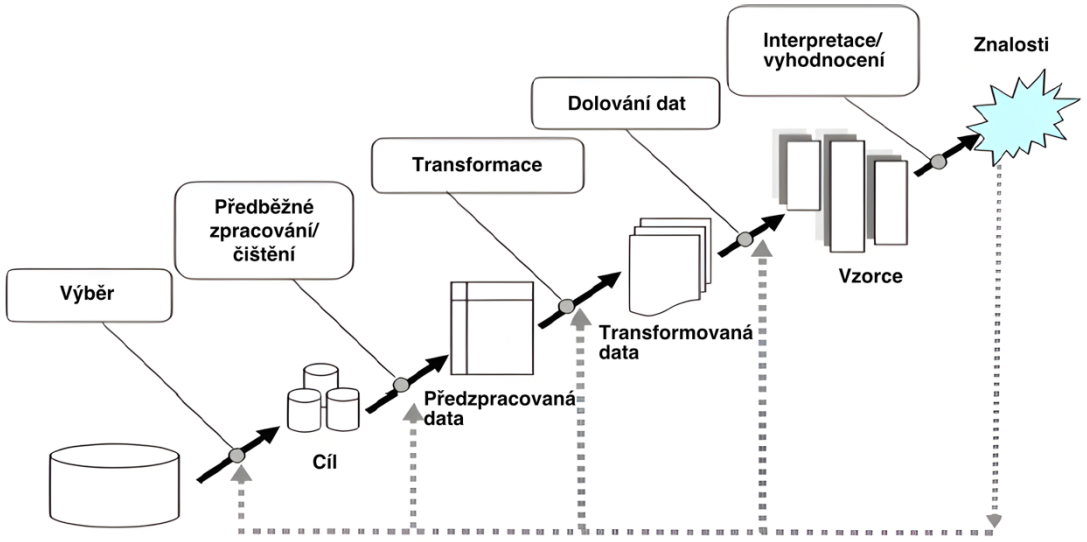
4.1 Dobývání znalostí z databází

S rozšířením digitálních zařízení a internetu vzrostlo množství ukládaných dat. Postupem času se ukázalo, že tato data mohou poskytnout cenné poznatky a podklady pro rozhodování v širokém spektru oborů, například ve zdravotnictví, marketingu, politice a mnoha dalších. Je zřejmé, že abychom vytěžili maximální potenciál takových dat, je zapotřebí je správně zpracovat a najít v nich souvislosti. Za tímto účelem vznikl na počátku 90. let koncept z oblasti umělé inteligence nazvaný dobývání znalostí z databází (KDD, Knowledge Discovery in Database). Impulsem konceptu je reálný problém, který si klade za cíl získat poznatky a znalosti z dat, která lze použít k informování při rozhodování nebo k podpoře dalšího výzkumu. [4] Na jeho vznik měla zásadní vliv rostoucí popularita standardizovaného strukturovaného dotazovacího jazyka (SQL, Structured Query Language), která v 90. letech stále rostla a rozšiřovala se v každodenním využití podniku ke správě dat. Byla navržena s využitím primárních a cizích klíčů a přinesla několik technik, které posílily množství ukládaných dat, a to především indexaci a normalizaci dat. [5]

Dobývání znalostí z databází popsal Usama Flayyad jako netriviální extrakci implicitních, dříve neznámých a potenciálně užitečných informací z dat. [6] Jinými slovy se jedná o holistický přístup k analýze dat, který vyžaduje netriviální proces zahrnující výběr vhodných dat, jejich přípravu, použití algoritmu, následnou interpretaci a vyhodnocení. Jeho hlavní účel je poskytovat podporu pro efektivní analýzu velkých datových souborů, a umožnit tak identifikaci vzorců a trendů, které nemusí být na první pohled zřejmé. Tímto zpřístupnil v několika odvětvích kvalitnější rozhodování na základě přístupu ke znalostem skrytým v datech.

4.2 Dolování dat a jeho metodiky

Obrázek č. 1 znázorňuje, že dolování dat je pouze specifickou částí komplexního procesu KDD, tyto dva pojmy bývají často nesprávně zaměňovány. Dolování dat se zaměřuje na aplikaci algoritmů a statistických modelů na datech, za účelem odhalení vzorců a vztahů, zatímco KDD je širší proces zjišťování znalostí z dat, který zahrnuje také další činnosti, jako je selekce, příprava dat, transformace, vizualizace a vyhodnocení výsledků. [7]



Obrázek 1 Dolování dat jako součást dobývání znalostí z databází [6]

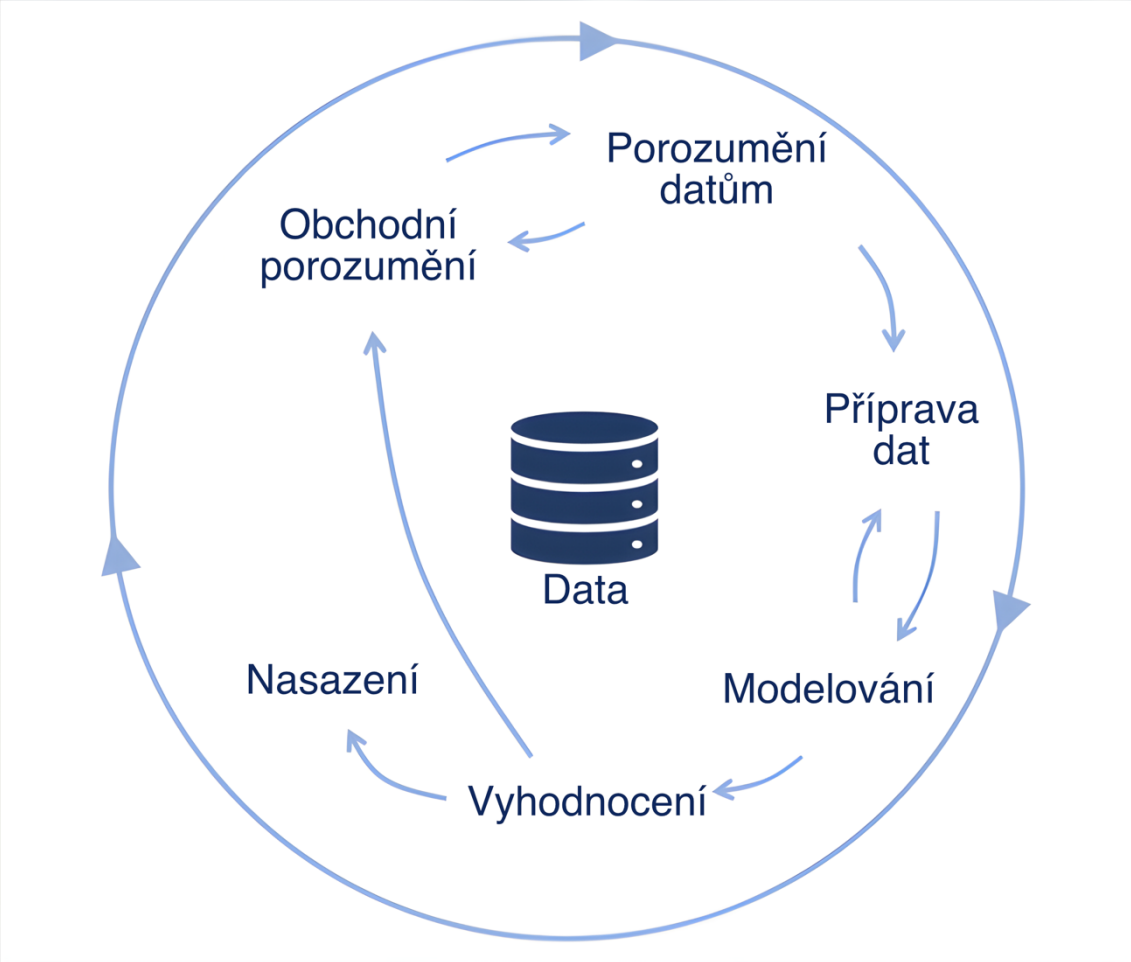
Dolování dat může být klasifikováno různými způsoby v závislosti na konkrétních použitých metodách, typu analyzovaných dat nebo cílech projektu dolování dat. Ve zjednodušeném pohledu můžeme rozdělit modelovací metody v dolování dat na dvě kategorie, a to na prediktivní a popisný model. Prediktivní model zahrnuje vytváření statistických modelů, které mohou předvídat budoucí výsledky nebo události na základě minulých dat. Jedním z mnoha oborů, který využívá takový model, je řízení vztahů se zákazníky (CRM, Customer Relationship Management). Popisný model zahrnuje analýzu dat za účelem identifikace vzorců a trendů, které popisují základní strukturu dat. Na rozdíl od prediktivního modelu nepředvídá budoucí chování a více se zaměřuje na hledání souvislostí v datech, které nejsou zřejmé, a popisuje jejich aspekty. [7] Oblasti použití popisných modelů jsou například analýza trhu, společenské vědy nebo ekonomie, kde je důležité porozumět současnému stavu a charakteristikám sledovaného jevu nebo skupiny.

Vývoj dolování dat vyžadoval vznik metodik, které by standardizovaly proces dolování dat a poskytly efektivní jednotný rámec pro různé úlohy. Metodika představuje soustavu pravidel, postupů, testovacích aktivit, výstupů a procesů, které slouží k řešení konkrétního problému. Metodika stanovuje nejen úkoly, postupy a výstupy, ale také určuje způsob, jakým by měly být úkoly provedeny. [8]

Je obecně známo, že metodiky snižují chybovost a poskytují konzistentnost i strukturovaný postup, který napomáhá sledování progresu v projektu. Níže je popsána metodika CRISP-DM, podle které je zpracována praktická část práce.

4.3 Metodika CRISP-DM

Metodika CRISP-DM (CRoss-Industry Standard Process for Data Mining) je metodika pro projekty týkající se dolování dat, která byla navržena společností NCR, SPSS a Daimler-Benz v roce 1996 a první verze byla vydána v roce 2000. [9] Je to jedna z nejpopulárnějších metodik pro projekty týkající se dolování dat a používá se v mnoha odvětvích.



Obrázek 2 Schéma metodiky CRISP-DM [10]

Obrázek č. 2 graficky zobrazuje CRIPSM-DM, který rozděluje proces dolování dat do šesti hlavních fází: porozumění doméně, porozumění datům, příprava dat, modelování, vyhodnocení a nasazení. Je to cyklický proces, kde není pevně

stanovená posloupnost fází, čímž umožnuje vracet se k předchozím fázím a dle potřeby je upravovat. [10] Flexibilita a modifikovatelnost metodiky pro různé potřeby projektu se projevuje v tom, že některé vnitřní šipky mezi jednotlivými fázemi jsou zpětné, což umožňuje návrat k předchozím fázím, například když zpracováním jedné fáze získáme nové znalosti, které napomohou lépe uchopit předchozí fáze. Vnější kruh na obrázku č. 2 znázorňuje právě cyklickou povahu celého procesu.

4.3.1 Fáze 1: Porozumění doméně

Úvodní fáze se zabývá obchodní perspektivou, zaměřuje se na porozumění cílů a požadavků projektu. Cílem této fáze je převést tyto znalosti do definice problému dolování dat a vytvořit předběžný plán pro dosažení cílů projektu. Je to důležitý proces, který zajišťuje, že se projekt bude ubírat správným směrem a přinese podniku požadovanou hodnotu. [9]

Znalosti získané analýzou dat samy o sobě nemusí stačit, proto je třeba je spojit s odbornými znalostmi a zkušenostmi z odlišných oblastí. Zapojení jednotlivců z různých sektorů, kterých se problém týká, a také těch, kteří budou řešení používat, pomáhá zajistit, aby řešení bylo pro podnik co nejrelevantnější a nejužitečnější.

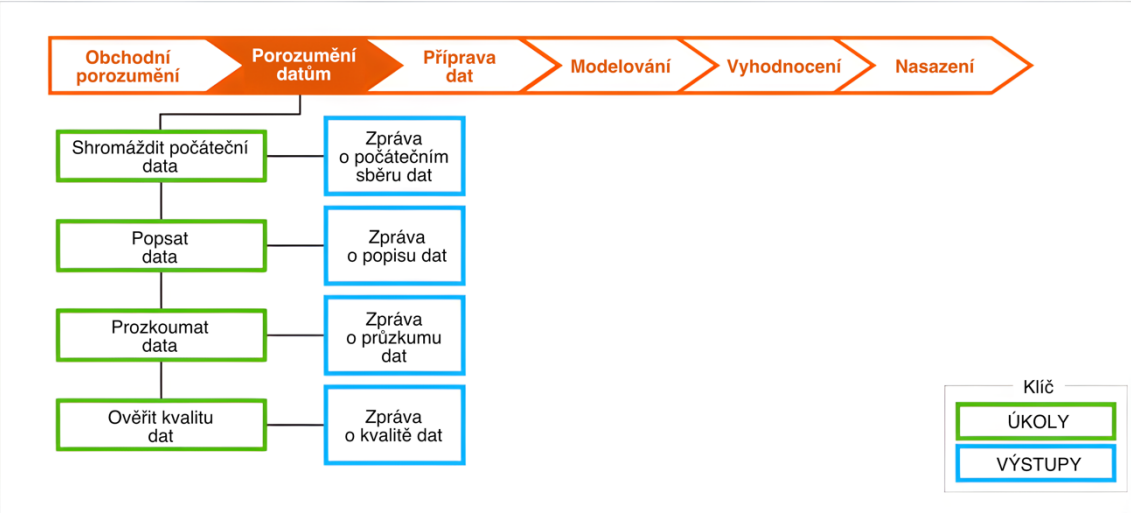
V rámci fáze porozumění doméně je uvedeno několik typů úloh podle dokumentace CRIPS-DM, které musí být vykonány pro pochopení a definování problému, který je žádoucí vyřešit. Mezi tyto úkoly patří:

- Identifikace a definice cílů projektu a požadavků na řešení.
- Identifikace klíčových stakeholderů, kterých se projekt týká.
- Identifikace a analýza kontextu projektu, včetně specifikací, podmínek a omezení.
- Identifikace a analýza datových zdrojů, které budou použity pro projekt.
- Identifikace potenciálních rizik a překážek pro úspěšné zpracování projektu.
- Identifikace příležitostí pro projekt a nalezení způsobů, jak využít výsledky projektu pro podnikání. [10]

Je tedy zřejmé, že porozumění doméně je obsáhlou a náročnou fází, vyžadující interakci s různými subjekty, jako jsou klíčoví stakeholdeři, odborníci na danou oblast a další zainteresované strany, aby bylo možné získat potřebné informace o kontextu a specifikách projektu. Porozumění doméně navíc není jednorázová fáze, ale proces, který se může v průběhu projektu opakovat a upravovat. Mnohdy může být tato fáze náročnější než samotné modelování, protože vyžaduje kontinuální práci a úpravy.

4.3.2 Fáze 2: Porozumění datům

Druhá fáze procesu CRIPSM-DM se zaměřuje na pochopení dat a zahrnuje proces získávání informací o datech z počátečního sběru. Tato fáze bývá často spjata s přípravou dat, kde dochází k nahrazení či odstranění chybějících hodnot, zamezení šumu, vypořádání se s odlehlými hodnotami a celkovou přípravou datové sady pro další analýzu. Cílem této fáze je objevovat první poznatky a potenciálně formulovat hypotézy. Na obrázku č. 3 lze vidět, že jejím výstupem jsou čtyři zprávy, které se týkají získání dat, popisu dat, explorace dat a vyhodnocení kvality dat, které slouží jako podklad pro další fáze procesu. Tyto zprávy poskytují detailní pohled na data a pomáhají pochopit jejich vlastnosti a kvalitu. [10]



Obrázek 3 Porozumění datům [11]

Často se ukazuje, že získání požadovaných dat může být náročným úkolem. Je běžné, že různá oddělení nebo podnikové jednotky v rámci firmy používají různé

systémy pro ukládání a správu dat. Tyto neintegrované databáze, které nejsou propojeny s ostatními částmi systému, se často označují jako „datová sila". V důsledku toho mohou vzniknout izolované skupiny dat, které nelze snadno kombinovat s informacemi z jiných částí organizace. Toto omezení vede k neefektivitě, protože organizace nemůže plně využívat všechna svá data k získání hlubších poznatků. [9] Integrace dat z různých zdrojů do jednoho celku je proces, který obvykle začíná porozuměním datům. V rámci tohoto procesu je provedena analýza zdrojových dat, aby bylo možné identifikovat, jaké informace jsou k dispozici, jak jsou data strukturována a jak mohou být data integrována do sjednoceného datového systému. [10] Pro správné interpretování shromážděných dat je nezbytné neopomenout zprávu zabývající se popisem dat. Zpráva je užitečným nástrojem, který bude obzvlášť nápomocný při integraci dat ve fázi přípravy dat. Obsahuje totiž důležité informace o zdroji dat a vlastnostech datového souboru, jako je jeho velikost, počet atributů a počet záznamů. Atributy jsou rovněž popsány a u každé entity je uveden název, datový typ a jeho využitelnost pro projekt. Zároveň lze do zprávy zahrnout výsledky korelací dat a doplnit názory odborníků v doméně ohledně jejich relevantnosti.

Dalším aspektem pro zprávu je explorace dat, v rámci, která je zaměřena na dotazování a převážně na vizualizaci dat. Analýza datové distribuce pomocí vizualizační metody poskytuje obrázkové reprezentace, jako jsou grafy a diagramy. Je mnohem jednodušší pro lidský mozek identifikovat vztahy mezi entitami pomocí vizuálního zobrazení než prostřednictvím uspořádaných dat ve formě řádků a sloupců. Proto tato technika činí data srozumitelnější tím, že využívá silnou stránku lidí rychle rozpoznat vzory a vytvářet propojení na základě grafického zobrazení. [12] Získané poznatky z explorace dat jsou zaznamenány do zprávy, která zahrnuje nejen výsledky analýzy, ale také počáteční hypotézy o dopadu na celkový projekt. [10]

V neposlední řadě se v rámci zpracování zprávy pro porozumění datům zabývá její tvůrce kvalitou dat. Zjišťuje se, zda se v datové sadě objevují chybná, neplatná a nekonzistentní data, která mohou zkreslit výsledky analýzy a ztížit jejich interpretaci. Měří se jejich četnost, mapuje se jejich nejčastější výskyt a na základě těchto zjištění se navrhuje řešení pro zlepšení kvality dat. Východiskem za cílem

zlepšení konzistentnosti v datové sadě může být nahrazení, ponechání či odstranění hodnot, které jsou problematické z hlediska kvality. [10]

4.3.3 Fáze 3: Příprava dat

Při přípravě dat je důležité provést transformace a úpravy, aby byla data sjednocena a připravena pro následné analytické modelování. Není náhodou, že tento proces obvykle následuje po fázích porozumění doméně a datům, protože umožní lépe pochopit kontext dat a identifikovat možná zkreslení způsobená špatnou kvalitou dat. Využití těchto znalostí a vložení kontextu při přípravě dat může vést k lepším výsledkům a přesnějším analýzám. [13]

Často se stává, že data získaná z databází neodpovídají požadavkům algoritmů pro dolování dat. Proto je důležité věnovat této fázi dostatečnou pozornost, abychom mohli z dat získat co nejvíce užitečných informací. Kvalita dat má významný dopad na úspěšnost všech následujících modelovacích procesů, a proto je klíčové tuto fázi pečlivě provést.

Nejčastějšími problémy, které je třeba vyřešit při přípravě dat, jsou:

- Čištění dat – odstraňování chybějících hodnot, duplicit a nesprávných záznamů.
- Transformace dat – převod dat do požadovaného formátu, úprava datového typu, normalizace a škálování.
- Odstranění anomálií – identifikace a odstranění odchylek nebo chyb v datech.
- Správa chybějících hodnot – identifikace a nahrazení chybějících hodnot, aby byl zajištěn kompletní soubor dat.
- Vyvažování dat – zajištění, že jsou v datech dostatečně zastoupeny všechny kategorie a třídy.
- Správa časových řad – zpracování dat v časových řadách, např. agregace, průměrování a odstranění sezónních vlivů. [9]

Samotné fáze procesu přípravy dat v metodice CRISP-DM, která se zabývá výše zmíněnými problémy, jsou výběr dat, čištění dat, konstrukce dat, integrace dat a formátování dat.

Výběr dat

Po získání dat můžeme dospět k závěru, že ne veškerá data jsou pro budoucí modelování užitečná. Bývají to data, která nejsou relevantní pro danou analýzu, tedy se neshodují s cílem analýzy nebo neodpovídají požadované kvalitě a mohla by prodlužovat čas výpočtu algoritmů. Technický aspekt je také důležitým kritériem při výběru dat, kdy nástroje pro analýzu mohou mít požadavky na množství dat nebo jejich datové typy. Výběr je prováděn na úrovni atributů a záznamů, tedy řádků i sloupců, a jejich začlenění či vynechání je nutné zdůvodnit ve zprávě. [14]

Čištění dat

Po vypracování zprávy o kvalitě dat v rámci předchozí fáze porozumění doméně je proveden výběr opatření, která povedou ke zlepšení kvality dat. Důležitou součástí čištění dat je vypracované hlášení, které obsahuje odůvodnění a opatření vybraných kroků včetně potenciálního vlivu na výslednou analýzu. [10]
Mezi nejběžnější problémy se řadí chybějící hodnoty, duplicitní a redundantní hodnoty a odlehlé hodnoty, které budou níže popsány.

1. Chybějící hodnoty

Jedním z běžných jevů v datových sadách jsou chybějící hodnoty, se kterými je nutné se vypořádat buď jejich vynecháním či nahrazením. Avšak výběr správné techniky pro práci s chybějícími hodnotami není vždy jednoznačný a závisí na mnoha faktorech, jako je typ proměnné, její význam pro analýzu, rozsah a způsob výskytu chybějících hodnot, množství a proporční zastoupení chybějících hodnot v datové sadě nebo i závislost chybějících hodnot na jiných proměnných. Je důležité neztratit žádné informace a zachovat integritu dat při nahrazování chybějících hodnot, aby byla zajištěna spolehlivost výsledků analýzy. [13]
Podle Rubina (1976), který formuloval základní pojmy mechanismů chybějících dat, rozděluje statistická literatura chybějící data do tří kategorií: MCAR, MAR a NMAR, které budou vysvětleny níže. [15]

Určení povahy chybějících dat a jejich zařazení do příslušné skupiny pomáhá při výběru nejvhodnější metody pro řešení chybějících dat v datové sadě. Každá kategorie vyžaduje specifickou analýzu a přístup, a proto je důležité pochopit, do které kategorie chybějící data spadají, aby byla zvolena nejlepší strategie pro jejich nahrazení nebo odstranění.

MCAR (Missing Completely at Random) neboli data, která chybí zcela náhodně. Do této kategorie spadají data, pro které chybějící hodnoty nejsou závislé proměnné, tedy nesouvisí se vzorem chybějících hodnot v souboru dat ani s pozorovanými hodnotami jiných proměnných. Pravděpodobnost, že nějaká hodnota chybí, je stejná pro všechna pozorování. [16] Vhodným příkladem je, kdy v průzkumu chybí data o respondentech ztracená v e-mailu. Tento předpoklad lze otestovat izolováním chybějících a úplných případů a prozkoumáním skupinových vlastností. Pokud charakteristiky obou skupin nejsou stejné, předpoklad MCAR neplatí. [17] U MCAR lze chybějící hodnoty snadno odstranit nebo nahradit vhodnými technikami, jako je nahrazení chybějících hodnot průměrem nebo mediánem. [18]

MAR (Missing at Random) neboli náhodné rozložení chybějících dat. Do této kategorie chybějící hodnoty spadají, pokud pravděpodobnost chybějících hodnot není závislá na hodnotě chybějícího údaje, ale může být závislá na hodnotách jiných pozorovaných proměnných. [19] Příkladem může být datová sada zachycující hodnoty IQ a věku. V tomto kontextu to znázorňuje, že absence hodnot IQ není spojena s jejich samotnými hodnotami (např. absence pouze nižších hodnot IQ), ale je spojena s hodnotou věku, konkrétně chybí hodnoty IQ u jednotlivců starších než 30 let. Zde lze pozorovat závislost mezi chybějícími hodnotami a jinou proměnnou. [17] Definice MAR je zavádějící kvůli použití slova "náhodný", vzhledem k tomu, že chybějící hodnoty nejsou zcela náhodné a korelují s jinými pozorovanými proměnnými. V případě MAR lze k nahrazení chybějících hodnot použít sofistikovanější metody, jako jsou imputační metody založené na regresi. [19]

NMAR (Not Missing at Random) neboli nenáhodné rozložení chybějících dat znamená, že pravděpodobnost vynechání hodnoty závisí na chybějící hodnotě. Jinými slovy chybějí hodnota může záviset na dalších neznámých faktorech, takže

je obtížné absentující hodnotu předpovědět. Například v datové sadě s IQ a věkem, pokud jsou chybějící hodnoty pouze pro lidi s nízkým IQ, existuje nějaký neznámý faktor, který ovlivňuje, zda se tato hodnota zaznamená, nebo ne, bez ohledu na věk nebo jiné známé faktory. Problém s mechanismem MNAR je v tom, že bez znalosti chybějících hodnot není možné zkontrolovat, zda jsou body hodnotami MNAR. [17] Tato kategorie se objevuje nejčastěji a vzhledem k její povaze bývá i nejproblematičtější pro nahrazení chybějících hodnot použít pokročilé metody, jako jsou imputační metody založené na modelování mechanismu chybějících dat. Následující odstavce poskytují stručný přehled vybraných metod pro imputaci chybějících hodnot. Imputace je obecně používaný termín pro doplnění chybějících záznamů o přijatelné hodnoty.

Deduktivní imputace

Jednou z nejjednodušších metod je deduktivní imputace, kdy na základě informací obsažených v jiných proměnných je možné dopočítat chybějící hodnotu. [20]

Imputace průměrnou hodnotu, mediánem nebo modální kategorií

Jednoduchou a rychlou metodou pro čištění dat je nahrazení chybějících hodnot v datové sadě průměrem, mediánem nebo modální kategorií, avšak nelze ji považovat za optimální. Je zřejmé, že nahrazení všech chybějících hodnot stejnou hodnotou nepřinese žádnou novou informaci, pouze rozšíří velikost vzorků. Tímto vede k podhodnocení rozptylu dat a snižuje celkovou variabilitu proměnné. Tyto metody je vhodné využívat pro kategorie chybějících dat MAR nebo MCAR. Pro numerické proměnné lze využít imputaci průměrem nebo mediánem, kdy jsou chybějící hodnoty nahrazeny průměrem nebo mediánem hodnot, které jsou k dispozici v dané proměnné. Medián je střední hodnota, která se nachází uprostřed rozsahu hodnot, když jsou seřazeny od nejnižší k nejvyšší hodnotě. Nahrazení modální kategorií se používá pro kategoriální proměnné. Modální kategorie je kategorie s nejvyšší frekvencí v dané proměnné. Obecně platí, že tyto metody se aplikují v případech, kdy je počet chybějících hodnot relativně malý a kdy neexistují výrazné vztahy mezi chybějícími hodnotami a ostatními proměnnými v datové sadě. Pokud je v datové sadě výrazná korelace

mezi proměnnými, může nahrazení chybějících hodnot mediánem nebo modální kategorií vést k výrazně zkresleným výsledkům. [21]

Regresní imputace

Regresní imputace je technika pro imputaci chybějících hodnot v datové sadě pomocí regresní analýzy. Základní myšlenka je predikovat chybějící hodnoty cílové proměnné na základě úplných proměnných. Nejčastěji používanou metodou je metoda lineární regrese, která pomocí lineární funkce předpovídá hodnoty chybějících dat na základě hodnot jiných silně korelujících proměnných v datové sadě. Tato technika může poskytnout poměrně přesné výsledky, pokud jsou k dispozici informace o vztazích mezi proměnnými v datové sadě. Regresní imputace je užitečná technika v situacích, kdy chybějící data neumožňují využít jiné metody imputace jako například jednoduchou náhradu průměrem nebo mediánem. [22]

Obecně u regrese platí předpoklad, že imputované hodnoty spadají přímo na regresní přímku s nenulovým sklonem, což znamená, že korelace mezi prediktivní proměnnou a chybějící výslednou proměnnou je rovna jedné. Regresní imputace nadhodnocuje korelace, ale podhodnocuje rozptyly a kovariance. [23]

2. Duplicitní a redundantní hodnoty

Duplicitní hodnoty jsou totožné hodnoty v řádcích, které mohou být způsobené například metodou sběru dat nebo sloučením dat z různých zdrojů. Ať je důvod jakýkoliv, duplicitní data mohou vést k chybným závěrům analýzy, protože některá pozorování mohou nastat častěji a nereprezentovat tak realitu. Je však zapotřebí zapojit znalosti domény a dat, aby bylo možné určit, zda se jedná o nechtěná duplicitní data. Pro zcela duplicitní data lze využít funkce pro odstranění těchto hodnot, které jsou integrované v mnoha nástrojích. [24]

Další možností je využití dostupných knihoven Pandas a NumPy v jazyce Python, které nabízejí funkce pro čištění dat. [25]

3. Odlehlé hodnoty

Pojem „odlehlé hodnoty“, často označovaný jako anglický výraz „outliers“, představuje anomální hodnoty, které se výrazně odlišují od standardů nebo očekávání daných dat. Jak uvádí autor Hawkins, odlehlé hodnoty jsou pozorování, která se od ostatních pozorování liší natolik, že vzbuzují podezření, že byla generována odlišným mechanismem. Je důležité zmínit, že určení toho, co je považováno za odlehlou hodnotu, závisí na konkrétním datovém souboru a subjektivním určení hranic mezi běžnými a odlehlými pozorováními, a proto není jejich definice jednoznačná. [9]

Jednou z metod pro identifikaci odlehlých hodnot v datech je použití histogramu, který vizualizuje distribuci hodnot v různých rozsazích. Potenciální odlehlé hodnoty lze nalézt mimo běžný rozsah hodnot, který se většinou u dat vyskytuje. Další metodou je použití krabicového grafu (boxplotu), který zobrazuje rozsah dat pomocí kvartilů a umožňuje identifikovat pozorování, která leží daleko od hlavního shluku hodnot. Lze využít také statistické metody pro detekci odlehlých hodnot v jedné proměnné, jako například z-score, které vyjadřuje, jak daleko je daná hodnota od průměru v jednotkách směrodatné odchylky. Hodnoty s vysokým z-score jsou považovány za potenciální odlehlé hodnoty. Existuje mnoho dalších metod pro vypořádání se s odlehlými hodnotami a také nástrojů, jako jsou statistické programy, programovací jazyky, nástroje pro vizualizaci dat a další. Tyto nástroje často obsahují implementované metody pro zpracování odlehlých hodnot, jako jsou odstranění neobvyklých hodnot, imputace chybějících hodnot, transformace dat a normalizace. Jejich využití umožňuje minimalizovat vliv odlehlých hodnot na analýzu a modelování, a pro zajištění přesnosti a kvality výsledků. [26]

Konstrukce dat

Při zpracování modelu a získání výstupu zkoumání se často objeví potřeba vytvořit nový atribut. V rámci konstrukce dat jsou k dispozici dvě varianty, a to generované a odvozené atributy. Generované atributy jsou vytvářeny pomocí matematických operací nebo jiných funkcí aplikovaných na existující atributy v rámci jednoho

záznamu. Může se například jednat o výpočet plochy na základě délky a šířky záznamu. Odvozené atributy jsou vytvářeny na základě kombinace různých existujících atributů. Tyto atributy mohou být vytvořeny pomocí logických operací nebo výpočtů z různých atributů, které jsou již k dispozici v datovém zdroji. Příkladem může být vytvoření atributu, který popisuje zákazníky, kteří v minulém roce neuskutečnili žádný nákup. Při tvorbě nových atributů je klíčové zvážit, zda jsou relevantní pro řešený problém a zda jsou správně vypočítány. Dále je důležité tyto atributy dokumentovat, aby bylo možné porozumět tomu, jak byly vypočítány a jak se používají v analýze dat. [10]

Pokud se analýza zaměřuje na textová data, je kromě vytváření nových atributů důležité provést lematizaci textu. Jedná se o proces, kde slova jsou v textu transformována na svou základní formu, známou jako lemma. Například slovesa v různém tvaru jsou převedena do svého infinitivu. Tato metoda je v textové analýze široce využívána k redukci variability slovních forem, což významně usnadňuje analýzu a snižuje složitost zpracovávaných dat. Klíčovým cílem lematizace je reprezentovat různé morfologické formy stejného slova jedním základním tvarem. Tento proces může být efektivně realizován s využitím knihovny NLTK v Pythonu. [27]

Dále je v procesu přípravy dat často potřeba kategorická nebo textová data převést na numerický formát, který je vhodný pro algoritmy datové analýzy. Tento proces se nazývá kódování a existuje celá řada technik. Základní rozlišení těchto technik pro kategorické hodnoty spočívá v tom, zda jsou hodnoty ordinální nebo nominální. Jinými slovy je třeba určit, zda dané kategorické hodnoty lze seřadit podle důležitost či přirozeného pořadí či nikoliv.

Pro nominální hodnoty je možné použít metodu „one hot encoding“, často označovaná jako kódování 1 z n, Tato technika spočívá v tom, že pro každou jednotlivou kategorii atributu se vytvoří samostatný binární sloupec. Příkladem může být atribut „Barva“, který může nabývat hodnot jako „Zelená“, „Modrá“ a „Žlutá“. V případě použití této techniky by pro atribut barvy vznikly tři nové sloupce, jeden pro každou barvu. Každý řádek v datové sadě by měl v jednom z těchto sloupců hodnotu 1, pokud daný záznam odpovídá této barvě, a v ostatních sloupcích hodnotu 0. [28]

V případech, kdy existuje mnoho jedinečných kategorických hodnot, může být vhodné využít techniku, která nahrazuje hodnoty kategorické proměnné frekvencí jejich výskytu v celkové datové sadě. Místo konkrétní kategorické hodnoty tedy sloupec ukazuje frekvenci, s jakou se daná hodnota v sadě vyskytuje. [29]
Pro hodnoty, které lze seřadit je ideální ordinální kódování, kdy jsou kategorické hodnoty mapovány na celá čísla, avšak rozdíl oproti jiným metodám kódování je v tom, že pořadí čísel odráží určitou hierarchii nebo uspořádání v kategoriích.
Pro převedení textových dat na numerické hodnoty vhodné pro algoritmy datové analýzy se využívá vektorizace textu. Je to proces transformace textu na číselnou reprezentaci. Jednou z metod vektorizace je TF-IDF (Term Frequency-Inverse Document Frequency). Tato metoda hodnotí váhu slov v dokumentu na základě frekvence jejich výskytu a současně bere v úvahu to, jak je daný výraz ojedinělý. [12] Pro aplikaci vektorizace metodou TF-IDF v jazyce Python lze využít knihovnu scikit-learn. Tato knihovna nabízí třídu TfidfVectorizer, jež umožňuje efektivní vektorizaci textových dat s možností modifikací parametrů. [30]

Integrace dat

V rámci integrace dat dochází k propojení dat z více zdrojů za účelem vytvoření jednotné datové sady. Pokud je tento krok proveden pečlivě, může to snížit možnou nekonzistentnost a duplicitu v datech. [7]
V této fázi je k dispozici zpráva zpracovaná v rámci porozumění datům s podrobným popisem, kde se data nacházejí, včetně jejich specifik. V závislosti na charakteru dat a jejich relacích může být využito různých metod integrace dat. Jednou z možností je sloučení dat. Tato metoda se obvykle používá, když získaná data z různých podnikových systémů jsou multirelační, tedy existuje relace mezi metrikami v datových sadách. Lze je spojit na základě společného atributu nebo klíče a získat tzv. flat dataset. Jinými slovy jednoduchý databázový systém, ve kterém jsou data uspořádána do jednoduché tabulky, kde každý řádek odpovídá jedné entitě a každý sloupec odpovídá jedné vlastnosti této entity. [12] Například v případě datové sady uskutečněných prodejů a datové sady o zákaznících je nutné využít společnou relaci od zákazníka, aby bylo možné vytvořit databázi, která kombinuje informace o názvu položek, celkové ceně, věku zákazníka, trvalém

bydlišti a počtu jeho objednávek. Tímto způsobem lze snadno vytvářet vztahy mezi různými metrikami a získat užitečné informace o prodeji a zákaznících. Další možností je použití agregace, která se využívá v případě potřeby získat souhrnné informace o datech. To znamená, že data jsou seskupena podle určitého kritéria a jsou vypočítávány statistické ukazatele, jako jsou průměr, součet, minimum, maximum nebo počet. V případě datové sady obsahující informace o prodejích může být pro účely projektu výhodné seskupit data podle identifikátoru zákazníka a vypočítat celkovou hodnotu nákupu pro každého zákazníka. Jednou z hlavních výhod využití agregace dat je snížení počtu záznamů, což může vést k výraznému snížení odezvy při provádění dotazů na velkých datových souborech. [10]

Při integraci dat je také důležité zvážit využití externích dat, která jsou veřejně přístupná a mohou poskytnout aktuální informace relevantní pro dané odvětví. Avšak je nutné být obezřetný při volbě identifikátoru pro agregaci dat, aby byly výsledky správně interpretovány. Například v případě výše zmíněné datové sady o prodejích a zákaznících by bylo vhodné využít externích demografických údajů obyvatelstva, aby bylo možné porozumět zákaznické základně a přizpůsobit prodejní strategie. [7]

Formátování dat

Analytické nástroje pro modelování často vyžadují určitou strukturu dat pro správné zpracování a interpretaci výsledků. Pokud jsou data uložena v odlišné struktuře, mohou nastat potíže při analýze a modelování dat. Formátování dat se tak často používá k přeměně dat do požadované struktury, aniž by se změnil jejich význam. Existuje mnoho různých formátů dat jako například XML nebo JSON pro hierarchické struktury, XLS pro binární formát a CSV pro textový formát. Kromě změny formátu dat může být potřeba změnit i pořadí záznamů, které může být požadováno algoritmem pro správné vykonání úlohy. Například pro neuronové sítě je často nutné mít náhodně seřazené záznamy pro dosažení nejlepších výsledků. [10]

4.3.4 Fáze 4: Modelování

Ve fázi modelování, která je předposledním krokem procesu CRISP-DM, je klíčové správně zvolit vhodnou modelovací techniku, která bude aplikována na danou datovou sadu. Nicméně je důležité zdůraznit, že samotné modelování může být ovlivněno předchozími fázemi procesu. To znamená, že i při správném výběru a aplikaci modelovací techniky nemusí být výsledek přesný. Obvykle je k dispozici více vhodných analytických procedur pro zpracování projektu, a proto je nutné experimentovat s výběrem metod. Je běžné, že je třeba se vrátit k předchozí fázi přípravy dat a provést dodatečné úpravy. Fáze modelování se dále dělí na čtyři klíčové podprocesy: výběr vhodné techniky, stanovení způsobu vyhodnocení modelu, vytvoření modelu a konečně jeho vyhodnocení. [29]

1. Výběr vhodné techniky

Tato kapitola se bude věnovat základnímu rozdělení technik strojového učení. Strojové učení je odvětví umělé inteligence zaměřené na vytváření modelů, které dokážou samostatně zpracovat data a učit se z nich. [27]
Rozdělení technik pomůže lépe porozumět metodám v širším spektru a zjednoduší to jejich výběr. Samotné techniky budou rozebrány až v jedné z následujících kapitol. Základní rozdělení je učení s učitelem neboli řízené učení (supervised learning) a učení bez učitele neboli neřízené učení (unsupervised learning). Toto rozdělení se primárně liší v tom, zda je k dispozici datová sada, kde jsou vstupy předem známé, či nikoliv. Využití řízeného učení předpokládá existenci části datového souboru, kde jsou pro určitou množinu vstupů již známy odpovídající výstupy. Tyto vstupy a odpovídající výstupy se použijí k vytvoření modelu, který dokáže předpovědět výstup pro nové vstupy. Model se tedy učí na tzv. trénovacích datech tak, aby se co nejlépe přizpůsobil známým datům. Poté se model použije na nová data, kde jsou vstupy známy, ale výstupy nikoliv. Algoritmus pak na základě modelu předpoví výstupy pro tato nová data. Tímto způsobem lze dosáhnout například klasifikace dat nebo predikce budoucích hodnot.
Na druhé straně učení bez učitele nevyžaduje použití datové sady s předem známými výstupy a spíše se zaměřuje na hledání skrytých vztahů a vzorů v datech

než na předpovídání budoucích hodnot. [9] Například může být použito na rozdělení datové sady s obrázky dvou druhů zvířat do dvou skupin na základě podobností mezi obrázky. V tomto případě nejsou obrázky předem označené a algoritmus se snaží najít skryté vzorce v datech, aby je mohl rozdělit do dvou skupin, které by mohly představovat dva druhy zvířat. Tento přístup se nazývá shlukování dat a je často označovaný anglickým výrazem „clustering“. Jak již bylo zmíněno výše, hlavním rozdílem mezi těmito dvěma technikami je, zda jsou trénovací data předem označena a zda model může být natrénován na těchto datech. Kromě tohoto rozdílu existuje také rozdíl v cíli modelování. Cílem neřízeného učení je získat vhledy z velkých objemů nových dat. Samotné strojové učení určuje, co je v datové sadě odlišné nebo zajímavé. Na druhé straně cílem řízeného učení je předpovědět výsledky pro nová data na základě již známých dat. Dalším rozdílem mezi těmito metodami je náročnost na výpočetní technologie. Řízené učení je relativně jednoduchá metoda strojového učení, která se obvykle vypočítává pomocí programů, jako jsou R nebo Python. Na druhé straně neřízené učení vyžaduje výkonné nástroje pro práci s velkým množstvím neklasifikovaných dat. [31]

2. Definice způsobu vyhodnocení modelů

Před samotným modelováním je důležité stanovit způsob, jakým bude model vyhodnocen. K tomuto účelu se často používá rozdělení datové sady na trénovací a testovací část. Trénovací data slouží k trénování modelu a jeho kvalita se ověřuje na testovacích datech, která nebyla použita při trénování. Při výběru poměru mezi trénovacími a testovacími daty je důležité zohlednit velikost celé datové sady. Pokud je k dispozici rozsáhlá datová sada, lze použít poměr 50:50, avšak u menších datových souborů se častěji používá náhodné rozdělení v poměru 80:20 ve prospěch trénovacích dat. [32] Toto rozdělení umožňuje odhalit, zda je model schopen správně předpovědět výsledky i na datech, které při trénování neměl k dispozici, což se nazývá schopnost modelu generalizovat. Pokud model dokáže dobře předpovědět výsledky pouze na trénovacích datech, ale selže na nových, může se jednat o přetrénování, což je často označováno anglickým výrazem „overfitting“. Při přetrénování dochází ke statistické chybě, kdy je model užitečný

pouze vůči původním datům, ale nikoli pro jiné datové soubory. Je důležité upozornit na to, že extrémy v datové sadě mohou ovlivnit kvalitu modelu. Příliš velká datová sada může být výpočetně náročná a zpomalit, nebo dokonce znemožnit celý proces kvůli nedostatečné výpočetní technice. Naopak příliš malá datová sada může vést k nedostatečné reprezentativnosti a riziku přetrénování modelu. Proto je důležité zvolit vhodné množství dat pro modelování, a pokud je to možné, také zkontrolovat, zda jsou data reprezentativní pro všechny možné situace, na které bude model aplikován. [27]

3. Průběh modelování

Samotný proces strojového učení, kde model rozpoznává vzory nebo predikuje výsledky z dostupných dat, bývá často automatizovaný a provádí se v modelovacím nástroji nebo pomocí algoritmu. Při modelování se často manipuluje s parametry, aby se získaly co nejvhodnější výsledky. Nastavení těchto parametrů je ovlivněno znalostmi získanými v předchozí fázi porozumění dat a doméně. Mnoho modelovacích nástrojů poskytuje možnost automatizovaného hledání nejlepších parametrů na základě definovaného kritéria úspěšnosti. Nicméně takový postup je často časově náročný a vyžaduje vysokou výpočetní kapacitu. Je důležité mít na paměti, že různá nastavení parametrů mohou směřovat k odlišným výsledkům. Proto je nutné experimentovat s různými kombinacemi parametrů a sledovat výsledky. V běžné praxi se často vytváří více modelů, aby bylo možné porovnat jejich účinnost a zvolit ten nejlepší pro daný účel. [32]
Nicméně, jak správně poznamenal George Box, „všechny modely jsou nesprávné, ale některé jsou užitečné“. Cílem je vytvořit co nejrealističtější model, ale v konečném důsledku je klíčové, aby byl užitečný pro daný účel. [33 str. 424]

4. Posouzení modelu

Jak již bylo zmíněno výše, v rámci modelování může být vytvořeno více modelů. v závěrečné fázi procesu modelování se pak provádí hodnocení a porovnání těchto modelů, aby byly vybrány ty nejvhodnější pro dosažení cílů projektu a splnění stanovených kritérií kvality. Tento krok zahrnuje posouzení úspěšnosti

provedených modelovacích technik, včetně testování na předem stanoveném datovém souboru. Hodnocení modelů by mělo být provedeno s ohledem na jejich schopnost dosáhnout stanovených cílů projektu. Finálně zvolený model by měl být schopen nejen dosáhnout požadovaných cílů, ale také by měl být interpretovatelný, což umožňuje porozumění mechanismům, které ovlivňují výsledky predikce nebo klasifikace. Po posouzení vytvořených modelů následuje fáze evaluace, během které jsou výsledky projektu hodnoceny z více úhlů pohledu. Zatímco posuzování modelů je zaměřeno převážně na technickou stránku, fáze evaluace zahrnuje kombinaci technických a doménových znalostí, aby bylo dosaženo co nejlepšího celkového výsledku. [10]

4.3.5 Fáze 5: Vyhodnocení výsledků

Hodnocení výsledků projektu se zaměřuje na to, jak dobře se podařilo dosáhnout obchodních cílů. I když vytvořený model může být sofistikovaný a obsahovat zajímavé vzorce a informace, nemusí nutně být relevantní pro cíle projektu. Proto je nezbytné zhodnotit, zda model skutečně splňuje svůj účel a zda přináší přínosy, které byly očekávány. Na základě této analýzy se určí další postup práce a případné úpravy předchozích fází. V případě nutnosti se může zvažovat návrat k určitým procesům a jejich změny. V konečném důsledku však hodnocení výsledků slouží k zajištění úspěšného naplnění cílů projektu a využití potenciálních přínosů, které projekt může nabídnout. [10]

4.3.6 Fáze 6: Implementace výsledků

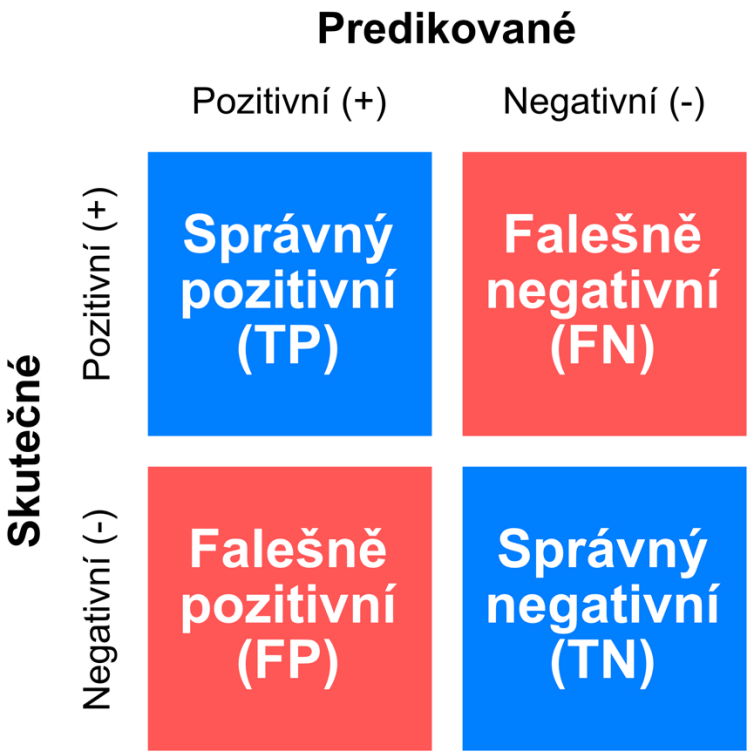
Po vytvoření modelu projekt nekončí, neboť samotný model představuje pouze výsledek aktivit spojených s dolováním dat, který nelze přímo aplikovat bez dalších kroků. Cílem integrace je upravit získané znalosti tak, aby byly přizpůsobeny potřebám zákazníka a aby byl model úspěšně integrován do existujících systémů, čímž se umožní snadný přístup a využívání výsledků pro uživatele. Nasazení modelu se může lišit v závislosti na konkrétních požadavcích zákazníka. Může jít o jednoduchý proces, například generování sestavy nebo naopak složitější metodu, jako je nasazení softwaru, který v sobě implementuje zmiňované modely a je zasazen do procesů dané organizace. [29]

Kromě samotné implementace je zapotřebí sestavit plán monitorování a údržby modelu, který bude zabraňovat nevhodné interpretaci výsledků. Plán monitorování a údržby by měl zahrnovat pravidelnou kontrolu výsledků modelu a aktualizaci modelu. [10]

4.4 Metody dolování dat

Metody dolování dat, jak již bylo naznačeno v předchozí kapitole, lze rozdělit do dvou hlavních kategorií: popisných a prediktivních modelů. Hlavním rozdílem mezi nimi je účel, pro který jsou tyto metody využívány. Popisné dolování dat se zaměřuje na popis dat a identifikaci vzorců a vztahů v datech, zatímco prediktivní dolování dat se zaměřuje na předpovídání budoucích událostí. [7]

V oblasti strojového učení je pro každý model důležité zjistit, jak je účinný a interpretovat toto hodnocení, aby bylo možné jednotlivé modely porovnat. Bez spolehlivé kvantifikace účinnosti modelu je obtížné posoudit jeho praktickou aplikovatelnost. V kontextu této práce je zvláště důležité zaměřit se na hodnocení klasifikačních modelů. Klasifikační modely jsou navrženy tak, aby rozpoznaly a přiřadily jednotlivé atributy do konkrétních kategorií. Aby došlo ke správnému pochopení a vyhodnocení toho, jak dobře model provádí tyto úkoly, je nezbytné porozumět některým klíčovým metrikám. Jedním z nejzákladnějších nástrojů pro hodnocení klasifikačních modelů je matice záměn (Confusion Matrix), která je znázorněna na obrázku č. 4. [34]



Obrázek 4 Matice záměn (Confusion Matrix) [34]

Při modelování klasifikace se hodnotící metrika vypočítává na základě skutečných pozitiv (TP), falešně pozitivních (FP), skutečných negativ (TN) a falešně negativních (FN). Tyto hodnoty představují výsledky klasifikace a lze z nich vypočítat řadu metrik, které poskytují hlubší vhled do výkonnosti modelu. Dvěma základními metrikami, které jsou často používány a jsou zvláště důležité pro tuto práci, jsou přesnost (accuracy) a F1 skóre.

Přesnost (accuracy) je metrika, která měří podíl správně klasifikovaných instancí z celkového počtu instancí. Vypočítá se pomocí vzorce:

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

kde TP je počet skutečných pozitiv, TN je počet skutečných negativ, FP je počet falešných pozitiv a FN je počet falešných negativ. [32]

F1 skóre je průměr mezi přesností a citlivostí. Je to užitečná metrika, zejména v situacích s nevyváženými daty a vypočítá se pomocí vzorce:

$$\text{F1 } score = \frac{2 * Precision * Recall}{Precision * Recall}$$

Kde:

- Precision je metrika, která hodnotí, kolik z pozitivně klasifikovaných instancí je skutečně pozitivních a je definována vzorcem:

$$Precision = \frac{TP}{TP + FP}$$

- Recall metrika někdy také nazývána „sensitivity“, měří schopnost modelu identifikovat všechny pozitivní instance v datech a je definována vzorcem:

$$Recall = \frac{TP}{TP + EN}$$

Ačkoliv každá metrika nabízí svůj jedinečný pohled na výkonnost modelu, je klíčové pečlivě zvážit její relevanci pro konkrétní model a datovou sadu. Ne všechny metriky jsou vhodné pro všechny úkoly. Správná volba metriky může vést k hlubšímu porozumění a lepší interpretaci modelu v rámci daného projektu. [32]
V následující části jsou prezentovány různé metody dolování dat společně s jejich základními principy. Tento teoretický rámec umožňuje hlubší pochopení uvedených metod, z nichž některé najdou uplatnění v praktické části práce.

4.4.1 Logistická regrese

Logistická regrese je metoda používaná v oblasti strojového učení a statistiky, zejména pro binární klasifikační úkoly. Ačkoli je logistická regrese názvem podobná lineární regresi, jejich cíle a výstupy jsou odlišné. Zatímco lineární regrese předpovídá skutečné kontinuální hodnoty, logistická regrese modeluje pravděpodobnost, že daný vstup patří k jedné ze dvou kategorií, tedy 0 a 1.
Pro lepší pochopení principu logistické regrese je zde popsána funkce lineární regrese:

$$y = a + b_1X_1 + b_2X_2 + b_3X_3 \ldots . b_nX_n$$

kde:

- a je průsečík obvykle nazývaná jako „intercept“,
- y je výstupní hodnota,
- $X_1, X_2 \ldots, X_n$ jsou vstupní proměnné,

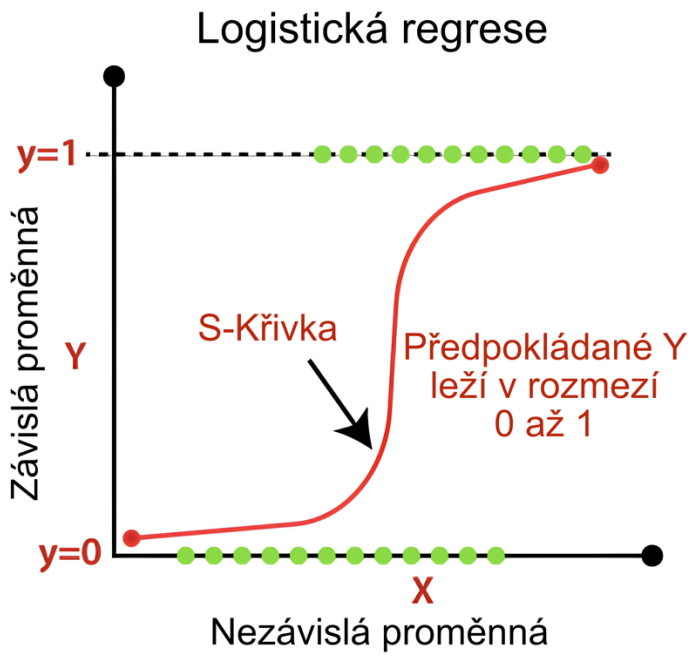
- b_1,b_2…,b_n jsou koeficienty přiřazené k těmto vstupním proměnným, které indikují vliv těchto proměnných na výstupní hodnotu.

Cílem lineární regrese je odhad parametru funkce, který nejlépe vysvětluje vztah mezi nezávislými a závislými proměnnými. Zatímco lineární regrese předpovídá kontinuální hodnoty y, logistická regrese modeluje pravděpodobnost určité kategorie. Jedním z klíčových prvků logistické regrese je její matematický model, který je založen na sigmoidní funkci. Tato funkce je zodpovědná za transformaci lineární kombinace vstupních proměnných na hodnoty mezi 0 a 1:

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

kde:

- $f(y)$ je odhadovaná pravděpodobnost, že pozorování patří do kategorie 1,
- e je základ přirozeného logaritmu (přibližně rovno 2,71828),
- y je výsledek lineární kombinace vstupních proměnných, tedy $y = a + b_1X_1 + b_2X_2 + b_3X_3 \ldots . b_nX_n$ [9]



Obrázek 5 Logistická regrese [35]

Na obrázku č. 5 je znázorněna logistická regrese. Nezávislá proměnná, často označovaná jako vstup, se nachází na ose x, zatímco závislá proměnná, jež je předmětem předpovědi, je na ose y. Křivka ve tvaru „S“, vytvořená sigmoidní funkcí, zajišťuje, že predikované hodnoty zůstávají v rozmezí [0,1], což je zásadní pro interpretaci výsledků jako pravděpodobností. Tímto způsobem logistická regrese upravuje výstupy lineární regrese tak, aby byly kompatibilní s binární klasifikací. [35]

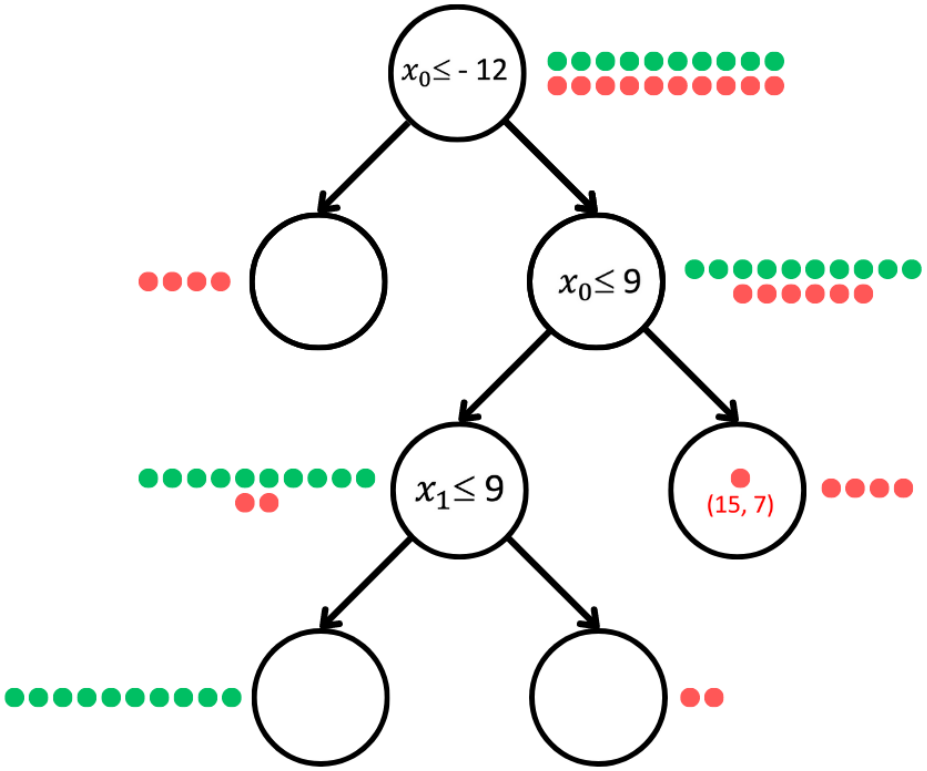
4.4.2 Asociační pravidla

Termín asociační pravidla získal popularitu na počátku 90. let díky Agrawalovi a jeho kolegům. k prosazení této metody došlo zejména v souvislosti s analýzou nákupního košíku, která se zaměřuje na současné nákupy různých druhů zboží v supermarketech jako například listové těsto a jablka. Je široce využívána zejména v oblastech, jako je webové dolování a maloobchodní průmysl, kde kromě analýzy zmíněného nákupního košíku pomáhá analyzovat chování uživatelů na internetu. Cílem asociační analýzy je identifikovat nejčastější kombinace hodnot mezi proměnnými a vytvořit pravidla, která tyto vzorce popisují. [36]

Asociační pravidla jsou zapisována ve tvaru a => B s významem „jestliže předpoklad – pak závěr“ neboli anglicky IF antecedent – THEN consequence Předpoklad představuje kombinaci hodnot proměnných, zatímco závěr vyjadřuje související hodnotu nebo událost. Pro vyhodnocení asociačních pravidel se využívají různé metriky, jako je podpůrnost (support), spolehlivost (confidence) a případně i korelace. Tyto metriky poskytují informace o frekvenci výskytu daného pravidla v datové sadě a o jeho významnosti. [27]

4.4.3 Rozhodovací stromy

Rozhodovací stromy jsou oblíbenou metodou v oblasti datového dolování a rozhodovací analýzy. Jak lze vidět na obrázku č. 6, princip rozhodovacích stromů spočívá v rozdělování datového souboru do předem definovaných tříd na základě rozhodovacích pravidel.



Obrázek 6 Ilustrace rozhodovacího stromu [37]

Tvorba rozhodovacího stromu začíná výběrem atributu, který nejlépe rozděluje data do jednotlivých tříd. Tím se vytvářejí podskupiny, nazývané uzly, ve kterých se nachází jednotlivé instance patřící převážně do jedné třídy. Tento proces je opakován na každém novém uzlu, který vznikne, dokud není možné nebo žádoucí provést další rozdělení datových instancí. Cílem konstrukce rozhodovacího stromu je dosáhnout koncových uzlů, nazývaných listy, které se převážně skládají z jednotlivých instancí jedné třídy. Každá instance je přiřazena do konkrétního listu, a tedy i do příslušné třídy, pokud splňuje všechna pravidla vedoucí k danému listu. Souhrn pravidel pro všechny listy tvoří model klasifikace. [9] [37]

Rozhodovací stromy mají několik výhod, které je činí atraktivními pro analýzu dat. Jsou intuitivní, protože poskytují explicitní pravidla pro klasifikaci. Dobře se vyrovnávají s různými druhy dat, umožňují kombinovat různé typy atributů a účinně zpracovávají heterogenní a chybějící data. Mohou být použity pro kombinaci různých typů atributů, ať už jsou kategoriální či numerické. Navíc umožňují interpretaci výsledků, protože jednotlivá rozhodovací pravidla jsou snadno srozumitelná. Avšak jedním z omezení rozhodovacích stromů je jejich relativní nepřesnost. Mnohdy nedosahují optimální predikční přesnosti, které by bylo možné dosáhnout s dostupnými daty. [27]

4.4.4 Náhodné lesy

V rámci datového dolování a strojového učení si náhodné lesy získaly renomé jako jedna z významných technik. Rozhodovací stromy, ač mají své místo v analýze, mohou čelit výzvám v prediktivní přesnosti. Jak uvádí Hastie: „Stromy mají jeden aspekt, který jim brání stát se ideálním nástrojem pro prediktivní učení, konkrétně nepřesnost.“ [38 str. 352]

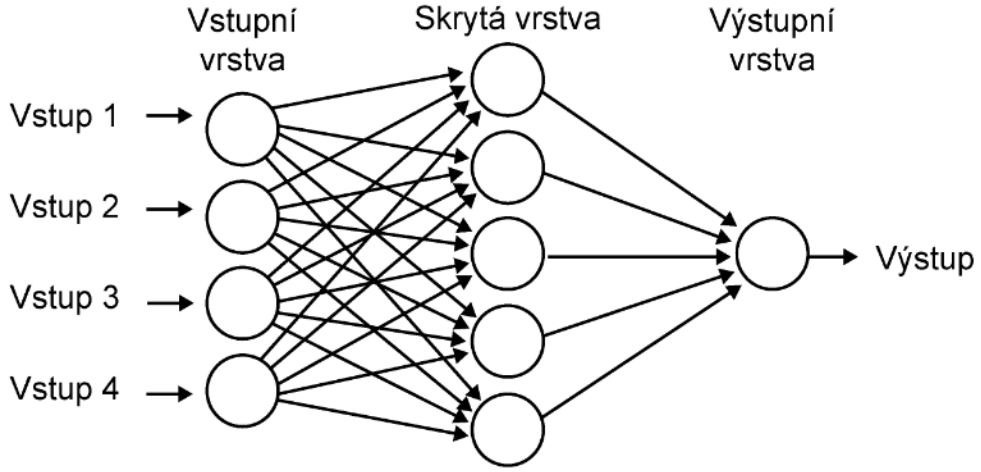
Na rozdíl od jednotlivých rozhodovacích stromů, náhodné lesy kombinují několik takových stromů s cílem maximalizovat přesnost v předpovědích. Kromě toho se náhodné lesy vyznačují rozmanitostí, protože mohou být úspěšně aplikovány jak v klasifikačních, tak v regresních úlohách.

Při modelování náhodných lesů se často využívá metoda „bootstrap“, která zahrnuje náhodný výběr různých podmnožin tréninkových dat s opakováním. Díky tomu se každý rozhodovací strom učí na odlišném vzorku dat, což přináší modelu potřebnou rozmanitost. Během tohoto procesu zůstává část dat nevyužita, která nese název „Out of Bag“ a tyto datové sady se často využívají pro validační účely, což umožňuje odhadovat chybovost modelu. Během rozhodování o vytváření uzlů v modelu se nevyužívají všechny dostupné atributy, ale pouze jejich náhodný výběr. Tento postup nejen podporuje rozmanitost v modelu, ale také snižuje potenciální riziko přeučení. Jakmile je model naučený, pro predikci nebo klasifikaci se výstupy jednotlivých stromů kombinují. Každý strom „hlasuje“ pro určitou klasifikaci nebo výsledek, a konečná predikce je následně určena buď většinovým

hlasováním (v případě klasifikace) nebo průměrováním výsledků (v případě regrese). [39]

4.4.5 Neuronové sítě

Neuronové sítě jsou matematické modely, které se inspirovaly strukturou a fungováním lidského mozku. Jejich využití se začalo zkoumat již v roce 1943 McCullochem a Pittsem, ale jejich další rozvoj proběhl až v 80. letech s nástupem digitálních počítačů. Neuronové sítě umožňují provádět různé matematické výpočty pomocí propojených neuronů. [9]



Obrázek 7 Uspořádání neuronů do vrstev v dopředné neuronové sítě. [40]

Princip fungování neuronových sítí je inspirován biologickými neuronovými sítěmi v lidském mozku. Na obrázku č. 7 je znázorněn princip dopředné neuronové sítě, která získala svůj název díky tomu, že signál se šíří od vstupu jednosměrně směrem k výstupům, přičemž nenastává žádné zpětné šíření signálu. Neuronová síť je složena z mnoha umělých neuronů, které jsou propojeny spoji, přičemž každému spoji je přiřazena váha. Neurony přijímají vstupní signály, zpracovávají je a generují výstupní signály. Disponují schopností učit se a adaptovat se na základě poskytnutých dat. Učení se v neuronových sítích zahrnuje proces nastavování vah spojů mezi neurony tak, aby sítě byly schopny přesně zpracovávat vstupní data a provádět různé úkoly, jako je klasifikace, predikce nebo rozpoznávání vzorů. To se provádí pomocí algoritmů učení, které optimalizují váhy spojů na základě tréninkových dat. Neuronové sítě se skládají z různých vrstev neuronů, vstupní

vrstva přijímá vstupní data, skryté vrstvy zpracovávají a transformují data a výstupní vrstva generuje výstupní hodnoty nebo predikce. Princip fungování neuronových sítí spočívá v šíření signálů od vstupní vrstvy k výstupní vrstvě pomocí vah a aktivačních funkcí. Váhy spojů určují, jakým způsobem se signály šíří mezi neurony a jaký vliv mají na výstupní hodnoty. Aktivační funkce řídí aktivační stav neuronů na základě přijatých signálů. [7] [40]

Použití neuronových sítí má několik výhod, ale také nevýhod. Mezi výhody patří schopnost modelovat nelineární vztahy mezi proměnnými a komplexní interakce. Neuronové sítě jsou také neparametrické, což znamená, že nevyžadují předpoklady o rozložení pravděpodobnosti vstupních proměnných. Díky tomu, že nepředpokládají konkrétní pravděpodobnostní distribuci, jsou neuronové sítě schopné pracovat s různými typy dat, včetně spojitých, kategoriálních a ordinálních proměnných. Tímto způsobem se mohou neuronové sítě přizpůsobit různorodým datovým strukturám, aniž by bylo nutné předem definovat konkrétní modelový předpoklad. Mezi nevýhody neuronových sítí patří jejich tendence k přeučení, často označovaný jako anglický výraz „overfitting“, kdy se příliš přizpůsobí trénovacím datům a ztratí schopnost generalizace na nová data. Navíc neuronové sítě mohou být náročné na výpočetní výkon. [27]

4.4.6 Bayesovské sítě (Bayes Network)

Bayesovská síť je statistický grafický model, který se používá k reprezentaci pravděpodobnostních vztahů mezi různými jevy. Tato síť je založena na Bayesově teorii pravděpodobnosti a pojmenována je po reverendu Thomasu Bayesovi, který vypracoval základní principy této teorie. [9] Grafický model Bayesovské sítě se skládá z uzlů a orientovaných hran, které je spojují. Každý uzel představuje proměnnou nebo jev, který je zkoumán. Hrany mezi uzly reprezentují závislosti a vztahy mezi těmito jevy. Tyto hrany mají směr, který ukazuje, který jev ovlivňuje druhý. V Bayesovské síti závisí každý uzel na svých předchůdcích, což znamená, že hodnota uzlu závisí na hodnotách proměnných, které ho ovlivňují. Každý uzel má také pravděpodobnostní rozdělení, které popisuje, jak se jeho hodnota mění v závislosti na hodnotách jeho předchůdců. Bayesovské sítě umožňují efektivní odvozování, což znamená, že lze odhadovat hodnoty proměnných na základě

dostupných informací. Díky propojení mezi uzly v Bayesovské síti je možné vypočítat pravděpodobnostní rozdělení pro různé kombinace hodnot proměnných v síti. Bayesovské sítě jsou důležitým nástrojem pro rozhodování za nejistých podmínek, kdy je třeba brát v úvahu různé faktory a jejich vzájemné vztahy. [7]

4.4.7 Metoda nejbližšího souseda

Metoda nejbližšího souseda (KNN, K-Nearest Neighbors) je jednoduchý, ale výkonný algoritmus, který se používá v oblasti klasifikace a regrese. Tato metoda je součástí řízeného učení a umožňuje zařazení do více tříd.

Principem algoritmu KNN je porovnání vlastností zkušebních vzorků s již natrénovaným modelem. Příznaky, které slouží k reprezentaci jednotlivých objektů, mohou být různého typu, například histogramy zvoleného barevného prostoru, které obsahují důležité informace o jednotlivých objektech.

Proces porovnání probíhá tak, že pro každý klasifikovaný objekt se vypočítá vzdálenost mezi ním a všemi objekty v trénovací sadě. Vzdálenost slouží k vyjádření míry podobnosti mezi daty, přičemž nižší hodnota vzdálenosti indikuje vyšší podobnost. Klasifikovaný objekt je následně kategorizován do třídy objektů, ke kterému je nejblíže. Euklidovská vzdálenost se většinou používá k výpočtu vzdálenosti, což je metrika, která měří přímou vzdálenost mezi dvěma body v prostoru. Existují i jiné metriky pro výpočet vzdálenosti, ale euklidovská vzdálenost je jednou z nejčastěji používaných a nejznámějších. [9] [12]

Pro příklad, tato metoda může být využita pro klasifikaci objektů, například obrázků květin, kde by bylo možné extrahovat histogramy zvoleného barevného prostoru pro každý obrázek. Tyto histogramy by sloužily jako vstupní data pro trénovací datový soubor. Histogramy obsahují informace o distribuci barev a intenzit v obraze. Trénovací datový soubor by tedy obsahoval extrahované histogramy pro každý obrázek květiny spolu s příslušnými třídami květin. Na základě těchto histogramů by byl natrénován KNN model. Poté lze testovat nový obrázek květiny, extrahovat jeho histogram zvoleného barevného prostoru a porovnat ho s histogramy v trénovacím datovém souboru. Vzdálenost mezi testovacím histogramem a histogramy trénovacích dat by byla vypočítána pomocí euklidovské vzdálenosti.

Metoda nejbližších sousedů je jednoduchá a snadno implementovatelná, avšak má také své omezení. Je citlivá na volbu vhodné hodnoty počtu nejbližších sousedů (k) a vyžaduje správnou volbu a normalizaci příznaků. Navíc může být náročná z hlediska výpočetního výkonu, přičemž nároky rostou s hodnotou konstanty k. Dále se algoritmus převážně zaměřuje na numerické proměnné, ačkoli lze zpracovat i kategorické proměnné, což vyžaduje speciální přístup. [9]

4.5 Nástroje pro dobývání znalostí z databází

Výběr vhodného nástroje pro vypracování modelu je klíčovým krokem pro každého datového analytika. Existuje mnoho možností, které lze zvážit, od programovacích jazyků jako Python, Scala a R až po nástroje, které jsou obecně snadno ovladatelné, ať už se jedná o open-source nebo komerční aplikace. Programovací jazyk Scala zahrnuje rysy objektově orientovaného programování a poskytuje silné nástroje pro vývoj distribuovaných systémů a paralelního zpracování dat. [41] Python a R jsou oba významné programovací jazyky, které jsou open-source pro analýzu dat a získaly si velkou popularitu. Každý z nich má své vlastní přednosti a mohou se navzájem doplňovat. Pokud je nutné vybrat pouze jeden z těchto jazyků, R nabízí větší specializaci na analýzu dat a modelování, zatímco Python je snadněji naučitelný a umožňuje sofistikovanější vizualizaci dat.

Mezi aplikacemi lze zmínit analytický nástroj BigML, který je komerčním produktem provozovaným společností BigML, Inc. Společnost BigML poskytuje tento nástroj jako placenou službu typu Software as a Service (SaaS). To znamená, že uživatelé mohou využívat nástroj prostřednictvím online platformy, a nemusí se starat o správu a provoz infrastruktury nebo softwaru. Další možností je využití platformy RapidMiner, jedná se o populární nástroj pro dolování dat, který je volně dostupný a umožňuje uživatelům analyzovat a zpracovávat data. [42] [43]

V následující kapitole bude podrobněji popsán programovací jazyk Python a jeho příslušné knihovny, které jsou vhodné pro vytvoření predikčního modelu v rámci této diplomové práce.

4.5.1 Python

Python je interpretovatelný, vysokoúrovňový, všeobecně použitelný programovací jazyk vynalezený Guidem van Rossumem. Byl poprvé uveden na trh v roce 1991 a jeho designová filozofie klade důraz na čitelnost kódu pomocí významného použití odsazení. Jeho jazykové konstrukce a objektově orientovaný přístup pomáhají programátorům psát srozumitelný a logický kód pro malé i velké projekty. Jazyk python má dynamické typování, což znamená že typ proměnných je definován za běhu programu a není nutné jejich typ předem deklarovat. Dále automaticky spravuje operační paměť. Toto automatické spravování paměti znamená, že programátoři nemusí ručně alokovat místo pro nové objekty nebo uvolňovat paměť, když objekty již nejsou potřeba. Tím se zjednodušuje proces programování a snižuje se potřeba manuální správy paměti. [44]
Python je preferován oproti jiným nástrojům pro analýzu dat z důvodu jeho využitelnosti v různých oblastech. Disponuje rozsáhlým ekosystémem knihoven, které rozšiřují jeho funkčnost a umožňují programátorům řešit různé problémy efektivněji a rychleji. [45] Následující knihovny jsou pouze některé z mnoha dostupných:

Pandas

Pandas je základní programovací knihovna používaná pro zpracování a analýzu dat v Pythonu. Jedná se o softwarovou knihovnu, která poskytuje funkce a nástroje pro správu datových struktur polí. Rozšiřuje možnosti Pythonu poskytováním objektů, jako jsou Series, což je jednorozměrné pole, které podporuje homogenní data, a DataFrame, dvourozměrné pole, které podporuje heterogenní data. Tyto objekty umožňují snadno přidávat, odebírat a upravovat datové sloupce, spojovat datové sady a vytvářet výběry na základě podmínek. Poskytuje možnost vytvářet objekty DataFrame přímo ze souborů a databází, převádět textové záznamy do časové řady a pracovat s daty vyjadřujícími čas. Pandas také umožňuje vytvářet grafy přímo z DataFrame pomocí knihovny Matplotlib. [46]

NumPy

NumPy, zkráceně Numerical Python, je významnou knihovnou používanou pro vědecké výpočty. Je široce využívána vzhledem k efektivní manipulaci s velkými multidimenzionálními poli. Pole, která tvoří jádro NumPy, představují kolekci prvků nebo hodnot s jednou nebo více dimenzemi. Jednorozměrná pole jsou obvykle nazývána vektory, zatímco dvourozměrná pole jsou známá jako matice. Na rozdíl od pole v Pythonu využívá NumPy menší paměťové nároky, je schopný provádět výpočty mnohonásobně rychleji a poskytuje rozsáhlou kolekci matematických funkcí. Z tohoto důvodu je široce využíván při zpracování velkých datových souborů a při provádění složitých výpočtů. [47]

Matplotlib

Matplotlib je významnou a vysoce ceněnou knihovnou specializující se na vizualizaci dat. Poskytuje uživatelům široké spektrum možností pro vytváření grafů, nejen jednoduchých, jako jsou liniové, sloupcové a bodové grafy, ale také pro pokročilé vizualizace, jako jsou 3 D grafy, animace a interaktivní grafy. Také je dobře integrován s dalšími knihovnami pro analýzu dat v ekosystému Pythonu, jako je NumPy a Pandas. [48]

4.5.2 Anaconda

Anaconda je open-source distribuce programovacích jazyků Python a R, která je zaměřena na vědecké výpočty, datovou vědu, strojové učení a další aplikace. Obsahuje velké množství předinstalovaných balíčků a knihoven, které jsou navrženy pro analýzu a zpracování dat.

Anaconda dále umožňuje vytvoření odděleného virtuálního prostředí pro každý projekt. Tímto zajišťuje, že specifické verze knihoven nevytváří konflikt s ostatními projekty. Poskytuje uživatelům užitečné grafické prostředí nazvané „Anaconda Navigator“, které usnadňuje správu balíčků a umožňuje přístup k dalším aplikacím. Díky tomu uživatel nemusí manuálně stahovat a instalovat jednotlivé komponenty potřebné pro svou práci. [49]

4.5.3 Jupyter Notebook

Jupyter Notebook je interaktivním prostředím určeným pro vývoj a spouštění kódu v různých programovacích jazycích, včetně Pythonu. Poskytuje přístup prostřednictvím webového prohlížeče a slouží nejen jako integrované vývojové prostředí (Integrated development environment, IDE) pro psaní kódu, ale také jako nástroj pro prezentaci, což jej činí ideálním pro sdílení poznatků a výsledků. Díky své flexibilitě může být provozován na různých platformách. Uživatelé mohou pracovat s Jupyter Notebook lokálně na svém počítači bez přístupu k síti nebo využít cloudová řešení jako například projekt Collaboratory od Googlu. [50]

5 Praktická část

Praktická část diplomové práce je zaměřena na analýzu datové sady, která byla získána z renomované platformy pro sdílení dat Kaggle.com. Tato část práce je zpracována v souladu s metodikou CRISP-DM, uznávaným standardem pro proces dobývání znalostí z databází. Hlavním cílem praktické části je vývoj spolehlivého predikčního modelu, schopného detekovat, zda je konkrétní nabídka práce podvodná. Součástí analýzy je také odpověď na řadu předem definovaných otázek, které pomáhají hlouběji porozumět charakteristikám dat. Celá analýza dat je provedena pomocí programovacího jazyka Python a jeho knihoven, které umožnují snadnou manipulaci s daty a tvorbou predikčního modelu. Výsledky analýzy a predikční model vytvořený v Jupyter notebooku jsou přiloženy na CD příloze diplomové práce.

5.1 Použité nástroje

Pro zpracování praktické části diplomové práce bylo využito prostředí Anaconda, jež poskytuje stabilní platformu pro instalaci a správu potřebných knihoven a balíčků pro analýzu dat. V rámci této práce byly použity následující klíčové balíčky:

- Numpy pro numerické výpočty,
- Pandas pro manipulaci s daty,
- Seaborn a matplotlib pro vizualizaci dat,
- Nltk pro zpracování přirozeného jazyka,
- Re pro práci s regulárními výrazy,
- Html pro zpracování HTML kódů,
- Collections pro manipulaci s datovými strukturami,
- Eordcloud pro vizualizaci slov v oblaku slov,
- Sklearn pro strojové učení a extrakci textových vlastností,
- a další specializované balíčky, jako WordNetLemmatizer, word_tokenize, TfidfVectorizer, a CountVectorizer.

Prostředí Anaconda usnadňuje řešení závislostí a zajišťuje kompatibilitu mezi jednotlivými knihovnami. Jako hlavní nástroj pro analýzu a zpracování dat bylo zvoleno vývojové prostředí DataSpell společnosti JetBrains. Toto vývojové prostředí (IDE, Integrated Development Environment) je navrženo speciálně pro potřeby analýzy dat a nabízí rozsáhlé funkce pro práci s notebooky Jupyter a interaktivní analýzu dat. Díky školní licenci získané pro DataSpell bylo možné využívat všechny pokročilé funkce tohoto nástroje, což velkou měrou přispělo k efektivitě a plynulosti zpracování dat.

V průběhu zpracovávání dat a jejich analýzy bylo pro sdílení výsledků a konzultace s vedoucím práce a odborným konzultantem využito prostředí Google Colab. Díky integraci s Google Drive bylo možné rychle a efektivně sdílet Jupyter notebook. Ve výchozím nastavení Google Colab poskytuje přístup k široké škále knihoven, jako jsou Numpy, Pandas, Matplotlib, Seaborn a Sklearn. Pokud byla potřeba další

specializovaná knihovna, bylo možné ji snadno nainstalovat přímo v notebooku pomocí příkazu „pip“.

5.2 Výběr a získání datového souboru

Základním kamenem každého datového projektu je výběr vhodného datového souboru. Pro účely této diplomové práce byla zvolena datová sada dostupná na platformě Kaggle.com, která je široce uznávaným zdrojem datových sad pro různé účely, včetně výzkumu a analýzy dat. Vybraná datová sada se nazývá „Fake Job Postings" a obsahuje téměř 18 000 záznamů o pracovních nabídkách. Tyto nabídky práce jsou označeny jako autentické, nebo podvodné, což činí tuto datovou sadu ideální pro účely této studie – vytvoření predikčního modelu, který dokáže identifikovat podvodné nabídky práce. Po výběru datové sady byl datový soubor stažen přímo z Kaggle.com ve formátu CSV. Tento formát je běžně používán pro ukládání tabulkových dat a je kompatibilní s širokou škálou nástrojů pro analýzu dat, včetně Pythonu, který je použit pro tuto práci.

5.3 Porozumění doméně

Ve světě, kde se digitální technologie stále více stávají nedílnou součástí našeho života, se nevyhnutelně setkáváme s přesunem mnoha oblastí do online sféry. Tento vývoj zasáhl i trh práce, což potvrzuje narůstající množství pracovních nabídek v online sféře. Přestože digitalizace přináší výhody, například v podobě snadného přístupu a rychlé komunikace, nese s sebou i rizika, zejména podvodné pracovní nabídky. Podle informací z Forbesu jsou tyto nabídky příčinou ztrát ve výši dvou miliard dolarů ročně. Jen v prvním čtvrtletí 2022 bylo 14 milionů lidí vystaveno riziku podvodů spojených s prací. Tyto nabídky se mohou objevovat v různých formách, od nabídek fiktivních pozic po požadavky na platby za „školení“ nebo „startovací balíčky“. [51]

V obtížném ekonomickém období, zejména po celosvětové pandemii covidu-19, se mnozí lidé stávají oběťmi těchto podvodů, které často předstírají, že zastupují známé společnosti. Zvýšený zájem o práci z domova zase vede k nárůstu podvodných nabídek v oblasti práce na dálku, což podvodníkům vytváří ideální pole působnosti.

V kontextu této práce bude podrobně analyzovaná datová sada obsahující reálné pracovní nabídky. Tato data poskytují základ pro vývoj predikčního modelu, jehož cílem je detekovat potenciální podvody. Ačkoliv v rámci této práce nebude implementace modelu provedena z důvodu jejího rozsahu, v budoucnosti lze toto téma rozšířit a zkoumat možnosti jeho implementace na různých pracovních portálech. Tím by mohlo dojít k ochraně uchazečů o práci a ke zvýšení pocitu bezpečí při online hledání zaměstnání.

5.4 Rozbor datového souboru

Rozbor datového souboru představuje základní krok před jakoukoliv další manipulací s daty. Kvalitně provedený rozbor vede k hlubšímu pochopení dat a umožňuje identifikovat možné problémy, které by mohly ovlivnit další analýzu. Následující podkapitoly přinášejí podrobnosti o procesu sběru dat, jejich popisu, výsledcích průzkumu a zhodnocení kvality.

5.4.1 Sběr a následný popis dat

Datový soubor této práce obsahuje informace o pracovních nabídkách zveřejněných online. Struktura datového souboru je členěna do 18 sloupců, z nichž každý reprezentuje jedinečný aspekt pracovní nabídky. Zde je výpis 18 atributů, včetně popisu každého z nich:

- job_id: jedinečný identifikátor každé pracovní nabídky,
- title: název pozice,
- location: lokalita pracovního místa,
- department: specifické oddělení společnosti,
- salary_range: nabízený platový rozsah,
- company_profile: stručný popis zaměstnavatele,
- description: podrobnosti o pracovním místě,
- requirements: specifikace požadavků,
- benefits: výhody, které pracovní místo nabízí,
- telecommuting: atribut, zda práce nabízí možnost práce na dálku,
- has_company_logo: přítomnost loga společnosti v nabídce,

- has_questions: přítomnost kontrolních otázek v nabídce,
- employment_type: typ pracovního poměru,
- required_experience: úroveň zkušeností požadovaných pro pracovní pozici,
- required_education: požadovaná úroveň vzdělání pro práci,
- industry: průmyslové odvětví společnosti,
- function: název pracovní pozice,
- fraudulent: ukazatel, zda je pracovní nabídka podvodná.

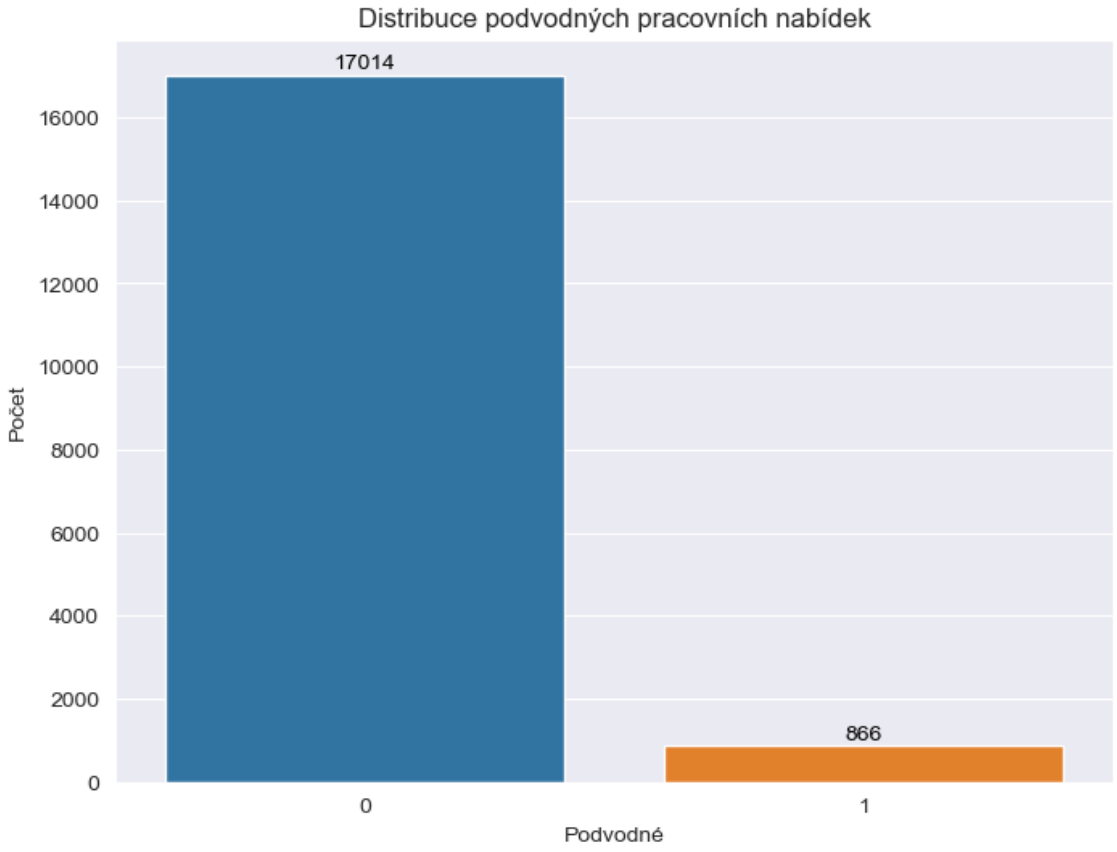
Datová sada má celkem 17 880 záznamů, které reprezentují jednotlivé pracovní nabídky. Datové typy jsou nejednotné, rozdělují se na celá čísla (int64) a textové řetězce (object). Při prvním průzkumu dat bylo identifikováno 281 duplicitních záznamů, které je nutné odstranit, neboť nenesou novou informaci. Klíčové sloupce obsahují hodnoty v očekávaném rozsahu, což potvrzuje validitu dat. Nicméně sloupce, jako je „salary_range“, „department“, „company_profile“ a „benefits“, vykazují značné množství chybějících hodnot, což by mohlo ovlivnit následné analýzy a modelování. Tato datová sada je z hlediska požadavků na tvorbu predikčního modelu optimální. Klíčový sloupec „fraudulent“ definuje jasný směr pro trénink modelů. Pro analýzu či modelování může být nezbytné data dále upravovat a přizpůsobovat, zejména s ohledem na požadavky konkrétního modelovacího algoritmu.

5.4.2 Průzkum dat

V rámci analýzy byla datová sada rozdělena na číselná a kategoriální data podle jejich datových typů. Je důležité zdůraznit, že některé kategoriální proměnné obsahují vysoký počet jedinečných hodnot, což je přibližuje k charakteristikám textových dat.

Průzkum dat se soustředí na číselná data, do nichž spadají sloupce: „job_id“, „telecommuting“, „has_company_logo“, „has_questions“ a „fraudulent“. Při detailnějším prozkoumání cílového sloupce „fraudulent“, který kategorizuje pracovní nabídky jako pravdivé či podvodné, vyplývá, že data nejsou rovnoměrně distribuována. Toto nerovnoměrné rozložení je patrné z grafu na obrázku č. 8. Většina záznamů, konkrétně 17 014, reprezentuje pravdivé pracovní nabídky,

zatímco podvodných nabídek je pouze 866. Toto nepoměrné zastoupení jedné třídy vůči druhé je třeba zohlednit v dalším zpracování dat, zejména při rozdělení datové sady na trénovací a testovací části. Je důležité zajistit, aby obě sady měly adekvátní zastoupení obou kategorií.



Obrázek 8 Distribuce podvodných pracovních nabídek (na základě dat z platformy Kaggle.com sestavil autor)

Sloupec „job_id“ je určen k jednoznačné identifikaci každého záznamu v datové sadě. Ostatní číselné hodnoty v datové sadě představují binární hodnoty (0 nebo 1) a nevykazují odlehlé hodnoty.

Následně byla provedena analýza výskytu podvodných pracovních nabídek v kontextu tří klíčových binárních proměnných: „telecommuting“, „has_company_logo“ a „has_questions“. Tato analýza posloužila k identifikaci potenciálních vzorů a charakteristiky, spojených s podvodnými nabídkami. Bylo zjištěno, že nabídky, které umožňují práci na dálku a zároveň nemají firemní logo, měly vyšší podíl podvodných nabídek.

Dále se zkoumalo procento podvodných pracovních nabídek v závislosti na kombinaci tří klíčových binárních proměnných: „telecommuting“, „has_company_logo“ a „has_questions“. V tabulce č. 1 jsou zobrazeny výsledky, z nichž vyplývá, že inzeráty, které umožňují práci na dálku, ale zároveň nemají firemní logo a nekladou žádné otázky, vykazují nejvyšší míru podvodných nabídek, konkrétně 29.27 %. Naproti tomu inzeráty, které umožňují práci na dálku, jsou opatřeny firemním logem a obsahují otázky, mají nejnižší míru podvodných nabídek, a to pouhých 1.42 %. Tato zjištění ukazují, že uvedené proměnné mohou sloužit jako užitečné predikční proměnné pro model, neboť absence některých informací v inzerátu může značně zvýšit riziko podvodné nabídky.

Tabulka 1 Procenta podvodných nabídek práce v závislosti na kombinaci tří binárních proměnných (na základě dat z platformy Kaggle.com sestavil autor)

	Telecommuting	*Company logo*	*Has questions*	*Fraudulent percentage*
0	0	0	0	17.590081
1	0	0	1	9.630459
2	0	1	0	1.787455
3	0	1	1	2.030728
4	1	0	0	29.268293
5	1	0	1	11.111111
6	1	1	0	6.956522
7	1	1	1	1.424501

5.4.3 Kvalita dat

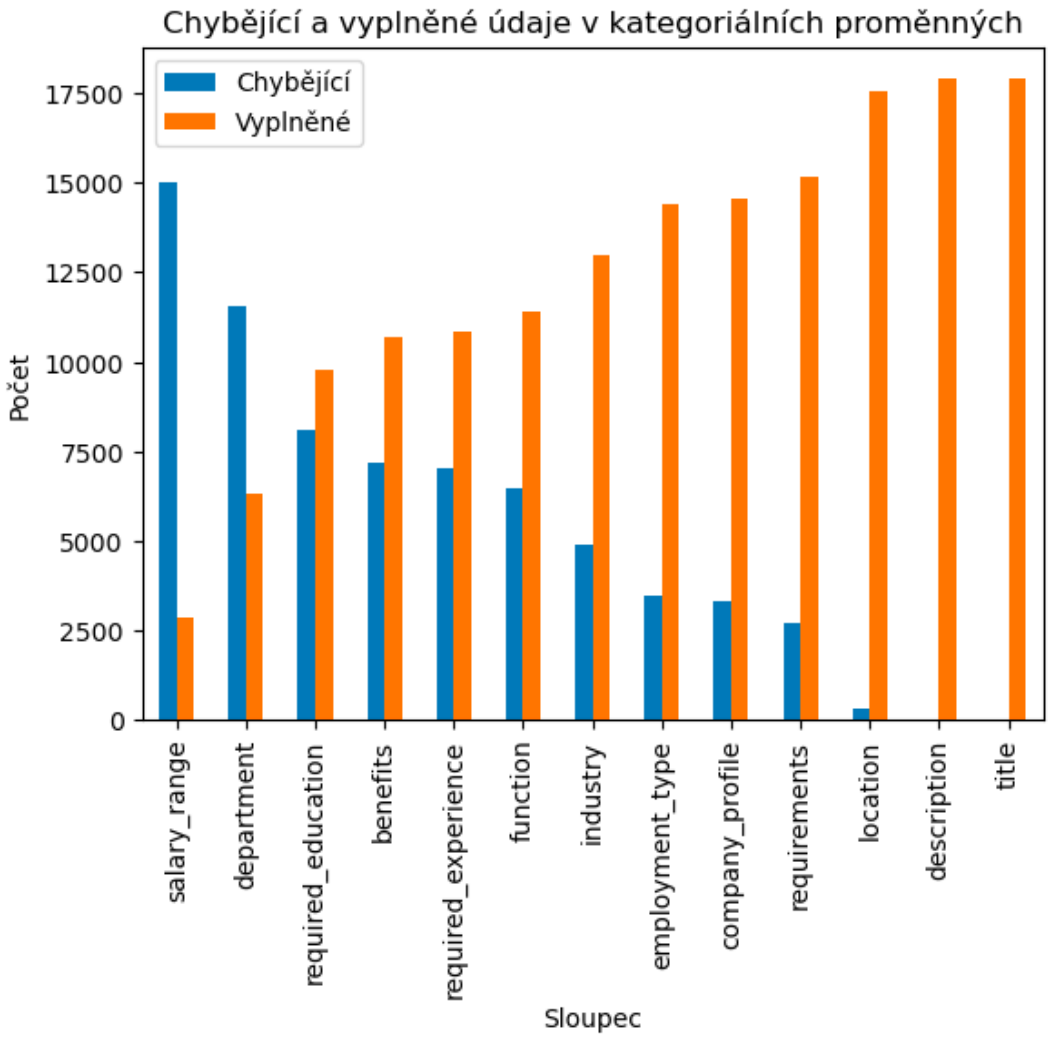
Ve vybrané datové sadě převažují kategorická data, což přináší řadu výzev týkajících se kvality dat. Jedním z hlavních problémů, které mohou u kategorických dat nastat, je nekonzistence v zápisu. Různé varianty zápisu pro stejné hodnoty mohou způsobit fragmentaci a nekonzistenci v datech. Například kategorické hodnoty mohou mít různé zápisy, jako „USA“, „U.S.A.“ nebo „Spojené státy“, což může komplikovat analýzu.

Dalším problémem je dvojznačnost mezi kategoriemi. Některé kategorie mohou být podobné nebo se překrývat, což může způsobit nejasnosti při analýze. To může zahrnovat situace, kdy různé kategorie mají stejný nebo podobný význam, ale jsou zapsány různými způsoby.

Je také možné, že se v datech objeví kategorie, které nenesou žádnou přidanou hodnotu pro analýzu a mohly by být považovány za redundantní. To může zahrnovat kategorie, které se vyskytují velmi zřídka nebo které nejsou relevantní pro konkrétní analýzu. Kromě toho může způsobit komplikace vysoká kardinalita, kdy některé kategorické proměnné mají příliš mnoho jedinečných hodnot.

Ve vybrané datové sadě nalezneme i číselné údaje, avšak jsou jednoduché, protože představují pouze binární hodnoty. Tato binární data by měla být snadno integrovatelná do predikčního modelu. Přesto je nezbytné prověřit, zda všechny hodnoty skutečně odpovídají hodnotám 0 nebo 1, a zda neexistují nějaké nesrovnalosti.

Další výzvou je řešení chybějících hodnot v kategorických datech. Nahrazení či doplnění těchto hodnot může být složitější než v případě číselných dat. Analýza se nejprve zaměřuje na chybějící hodnoty v kategoriálních proměnných, jež jsou zobrazeny v grafu na obrázku č. 9. Tento graf podrobně ilustruje rozložení chybějících hodnot v kategoriálních proměnných. Z vizualizace vyplývá, že většina sloupců disponuje kompletními nebo téměř kompletními daty. Nicméně některé sloupce, především „salary_range“, vykazují významné množství chybějících hodnot.



Obrázek 9 Chybějící a vyplněné údaje v kategoriálních proměnných (na základě dat z platformy Kaggle.com sestavil autor)

Chybějící hodnoty mohou vznikat z různých příčin. Mezi ně mohou patřit chyby v procesu sběru dat, neúplné informace poskytnuté zaměstnavateli, nebo záměrné vynechání informací kvůli jejich citlivosti, či z toho důvodu, že nebyly povinné při vyplňování pracovních nabídek. Pro tyto případy je důležité zvážit, zda je vhodnější tyto chybějící hodnoty nahradit či je ponechat v původním stavu. Pokud absence určitých hodnot nemá negativní dopad na kvalitu prediktivního modelu, může být vhodnější je ponechat beze změny. V další části analýzy se bude zkoumat problematika chybějících hodnot v jednotlivých sloupcích podrobněji.

Sloupec „salary range“

Ve sloupci „salary_range“ je zaznamenáno 15 012 chybějících hodnot. Vyplněné hodnoty jsou nekonzistentní, často prezentují platový rozsah oddělený pomlčkou, což reprezentuje minimální a maximální nabízený plat. Možnost převedení těchto rozpětí do kategorií, například „Nízké“, „Střední“ a „Vysoké“, existuje, avšak tento přístup není bez nedostatků. Definice hodnot jako „Nízké“ nebo „Vysoké“ může být subjektivní a může se lišit v závislosti na regionu či odvětví. Využití střední hodnoty platového rozpětí by mohlo tuto variabilitu částečně eliminovat. V tomto kontextu by se vypočítal průměr minimálního a maximálního platu, a ten by se následně využíval jako kontinuální proměnná. Ovšem tento způsob přináší dílčí komplikace: odlišná měřítka uvedených platů (hodinové, měsíční, roční) by vedla k tomu, že střední hodnoty by napříč různými nabídkami práce nebyly konzistentní, což by mohlo výsledky analýzy zkreslit.

Často se stává, že zaměstnavatelé v nabídkách práce záměrně neuvádějí platy, ať už z důvodu zajištění flexibility při vyjednávání či obav z možného odrazení kandidátů. V takových situacích může být informace o tom, zda je plat specifikován, mnohem významnější pro predikční model než samotná částka. Jako alternativa by se tedy mohl vytvořit binární sloupec, který by zaznamenával, zda je plat uveden či ne, což by mohlo být pro modelování užitečnější, než pokusy o imputaci chybějících hodnot nebo jejich kategorizaci.

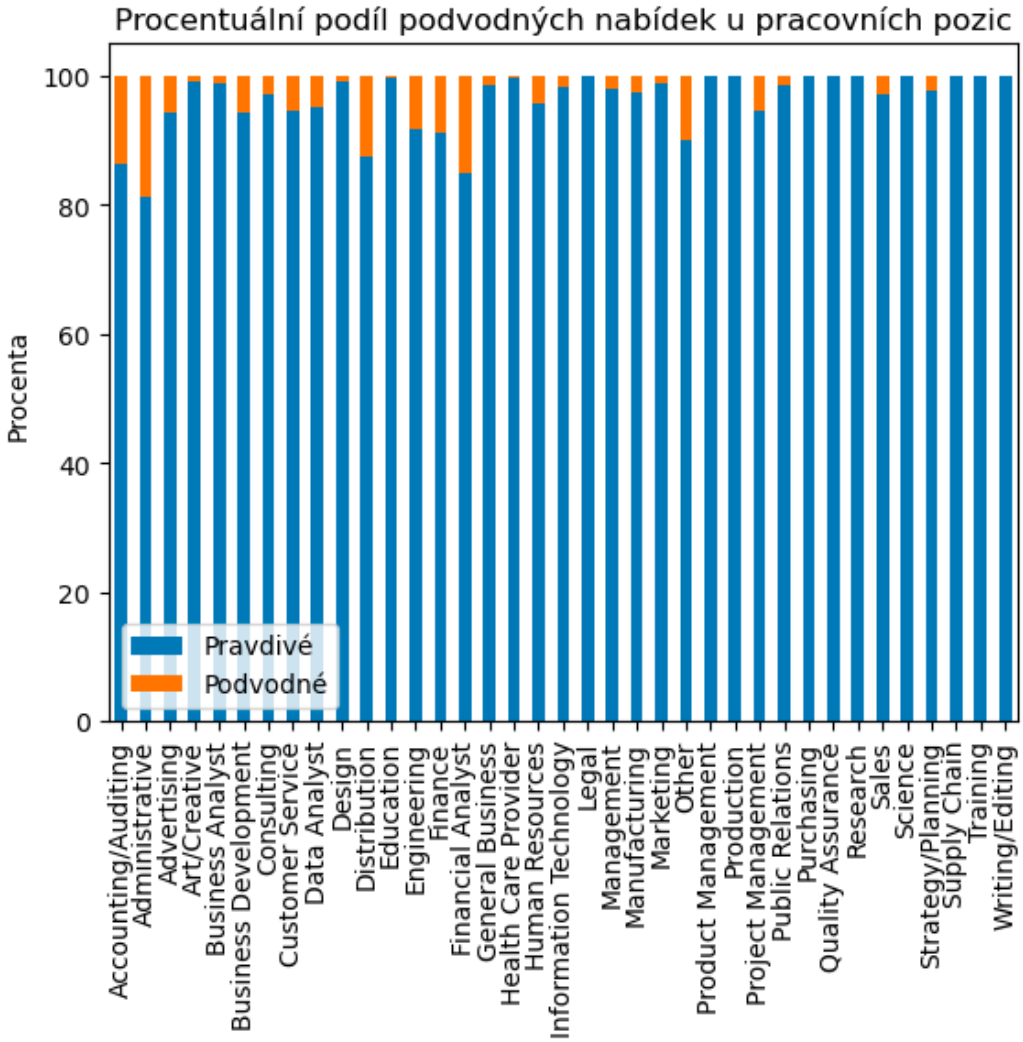
Sloupec „location“

Sloupec „location“ v datovém souboru reprezentuje geografickou polohu nabízeného pracovního místa. Z analýzy několika záznamů je patrné, že formát hodnot v tomto sloupci se liší v závislosti na konkrétní zemi. Nabídky ze Spojených států amerických obvykle následují formát země, stát, město, zatímco u evropských a jiných zemí je formát stát, okres, město. Tato variabilita formátu může ztěžovat analýzu a interpretaci dat. Jedním z možných přístupů k řešení této nekonzistence by bylo zachování pouze prvního údaje před čárkou, což by v mnoha situacích odpovídalo názvu země. Takové řešení by ovšem vedlo ke ztrátě podrobnějších informací o lokalitě, zejména v případě nabídek ze Spojených států amerických.

Po zvážení všech aspektů se zdá být optimálním řešením rozdělení hodnot ze sloupce „location“ do čtyř nových sloupců, konkrétně „location_country“, „location_state“, „location_district“ a „location_city“. Tento způsob nejenže zachová všechny relevantní geografické informace, ale také přinese strukturovanější formát pro následné analýzy.

Sloupec „function“

Při analýze kvality dat byl zkoumán sloupec „function“, který reprezentuje specifickou pracovní pozici nabídky práce. Z výsledků analýzy vyplývá, že největší počet pracovních nabídek spadá do kategorie informačních technologií.



Obrázek 10 Procentuální podíl podvodných nabídek u pracovních pozic v jednotlivých oborech(na základě dat z platformy Kaggle.com sestavil autor)

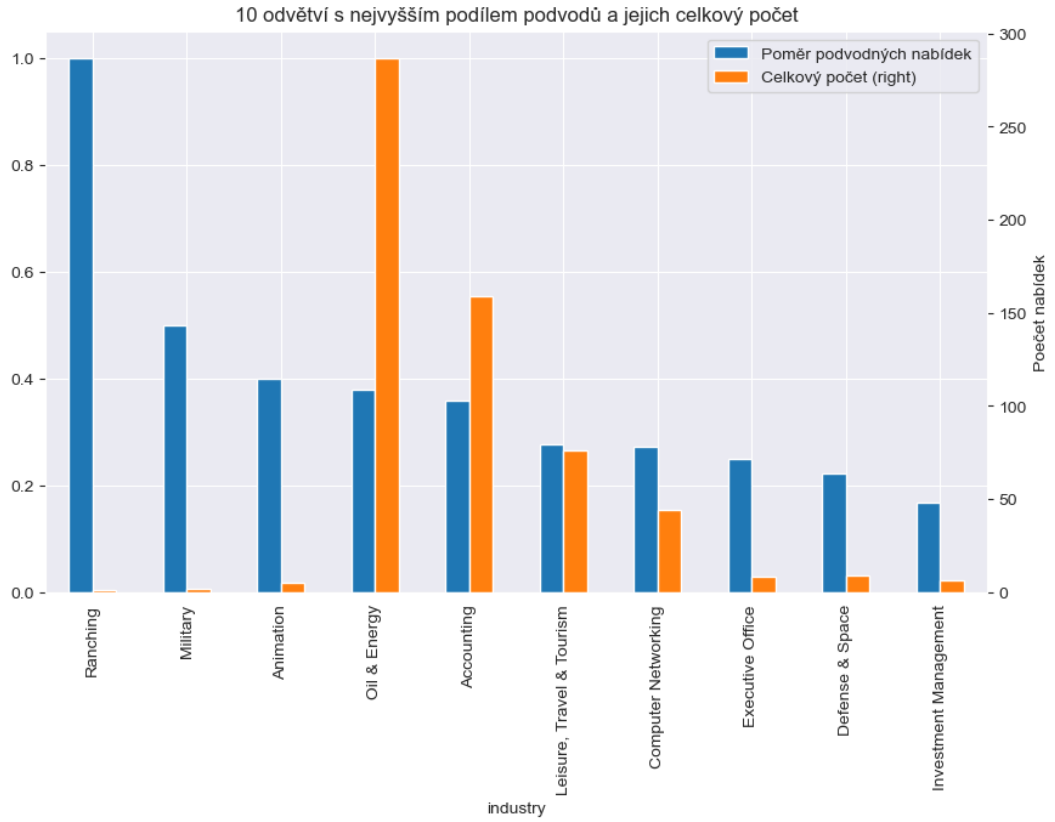
Z obrázku č. 10, na němž je zobrazen procentuální podíl podvodných nabídek vůči celkovému počtu, vyplývá, že nejvíce podvodných nabídek je v oblasti administrativy, následované účetnictvím/auditem. Tyto obory se zdají být častým cílem podvodníků, což může poukazovat na zvýšenou poptávku v těchto oborech nebo na nižší práh pro ověření pravosti nabídek.

Následovala analýza možného vztahu mezi sloupci „function“ a „industry“ s použitím chí-kvadrát testu. Předpokládaná nulová hypotéza představovala nezávislost kategorií ve sloupci „function“ na kategoriích ve sloupci „industry“. Alternativní hypotéza předpokládala existenci určité závislosti mezi těmito sloupci. Na hladině významnosti 0,05 byla nulová hypotéza zamítnuta, což ukazuje na významnou spojitost mezi danými sloupci. Vzhledem k této spojitosti mezi sloupci „function“ a „industry“ se jeví jako vhodné řešení vytvoření kontingenční tabulky, zobrazující četnost kombinací hodnot z obou sloupců pro lepší přehled ve vzájemných vztazích. Na základě těchto informací by bylo možné navrhnout imputaci založenou na nejčastějším výskytu funkcí v daných průmyslových odvětvích.

V následující části práce, zaměřené na čištění dat, bude nezbytné detailněji zkoumat problematiku sloupce „function“ a zvážit možné strategie imputace či transformace dat.

Sloupec „industry“

V další části práce byl podrobně zkoumán sloupec „industry“, který reprezentuje odvětví nabízené pozice. Jedním z dílčích cílů této diplomové práce je identifikace odvětví s nejvyšším počtem podvodných nabídek pracovních míst. Z analýzy vyplývá, že odvětví „Oil & Energy“ vykazuje nejvyšší podíl podvodných nabídek, což je patrné z obrázku č. 11.



Obrázek 11 Odvětví s nejvyšším podílem podvodů a jejich celkový počet (na základě dat z platformy Kaggle.com sestavil autor)

Bylo rovněž zjištěno, že sloupec obsahuje množství unikátních hodnot. Pro hlubší pochopení byly zkoumány hodnoty v odvětvích spojených s technologiemi, zejména těch, které obsahují klíčová slova „Technology“, „Computer“, „Software“ a „Internet". V rámci této analýzy bylo identifikováno několik variant těchto názvů v různých kombinacích. Toto zjištění vedlo k otázce, zda by některé z těchto hodnot měly být sjednoceny do širších kategorií.

Po další analýze bylo zjištěno, že zatímco sloupec „industry“ popisuje odvětví, ve kterém firma působí, sloupec „function“ definuje konkrétní pracovní pozici nabízenou v daném odvětví. Konkrétně pracovní pozice „Engineering“ může být zastoupena v různých odvětvích, jako je letecký průmyslu nebo softwarové inženýrství, přičemž každé odvětví má svá specifika. Slučování hodnot v sloupci „industry“ by mohlo vést ke ztrátě nuancí o konkrétních odvětvích, která mohou mít odlišné charakteristiky i v rámci širší kategorie.

Sloupec „department“

Ve sloupci „department“, obsahujícím celkem 17 880 záznamů, bylo zjištěno 1 337 unikátních hodnot. Tato variabilita v názvech oddělení vedla k otázce, zda neexistují nějaké skryté nekonzistence, například způsobené rozdílnou velikostí písmen v názvech. Pro identifikaci těchto nekonzistencí byly všechny názvy oddělení převedeny na malá písmena, při nichž se snížil počet unikátních hodnot na 1 224, tudíž by tato změna měla být aplikována v rámci čištění dat.

Dále byla využita tokenizace hodnot ve sloupci, díky níž bylo možné identifikovat několik klíčových slov, která by mohla signalizovat nekonzistence, jako jsou například „tech“ a „technology“ nebo „dev“ a „development“. Tato analýza odhalila některé hodnoty, které se lišily pouze velikostí písmen.

Kromě identifikovaných nekonzistencí byla provedena analýza slov, která se ve sloupci „department“ vyskytují nejčastěji. Přitom bylo zjištěno, že termíny „Technology“ a „Technical“ bývají velmi časté. Mohlo by se zdát vhodné je sloučit s pojmy jako „Information Technology“, avšak hlubší analýza odhalila, že různá oddělení s podobnými názvy mohou mít zcela odlišné činnosti. Například, zatímco oddělení „Technology“ v jedné společnosti může být odpovědné v rámci pracovní pozice za vývoj letadel, oddělení „Information Technology“ v jiné společnosti může být zaměřeno na vývoj softwaru. Tato jedinečnost různých oddělení je zásadní, a proto bylo rozhodnuto ponechat většinu těchto unikátních hodnot, aby se minimalizovala možná ztráta informací.

Dále je třeba poznamenat, že ve sloupci „department“ bylo zaznamenáno 11 547 chybějících hodnot. i když nepřítomnost těchto hodnot může působit jako omezení, z hlediska predikčního modelu pro odhalení pravosti nabídek pracovních míst by jejich nepřítomnost neměla významně ovlivnit kvalitu predikcí. Zároveň z hlediska počtu unikátních hodnot se tento sloupec jeví textového rázu a pro další analýzu by neměl být zařazen mezi kategorická data.

Sloupce „title“, „description“, „requirements“, „benefits“ and „company profile“

V rámci analýzy kvality dat datové sady nabídek práce bylo rozhodnuto se zaměřit na sloupce, které byly v popisné analýze klasifikovány jako kategorické na základě

jejich datového typu. Při bližším zkoumání, jak je zobrazeno v tabulce č. 2, bylo zjištěno, že sloupce „company_profile“, „description”, „requirements“, „benefits“ a „title“ obsahují velký počet unikátních hodnot, což je vedlo k jejich považování za textové sloupce. U sloupce „title“ nebyly identifikovány žádné chybějící hodnoty, což naznačuje, že tato informace může být při vytváření inzerátu povinná.

Dále byl proveden průzkum k identifikaci tzv. whitespaces, tedy oddělovacích znaků. Oddělovací znaky v programování a zpracování dat obvykle odkazují na mezery, tabulátory a nové řádky, které mohou být na začátku nebo na konci řetězců a mohou způsobovat nepřesnosti v datech. Po odstranění nadbytečných oddělovacích znaků bylo odhaleno, že některé hodnoty, které se původně zdály být odlišné, jsou po této úpravě stejné.

Tabulka 2 Počet nekonzistencí pro zkoumané sloupce (na základě dat z platformy Kaggle.com sestavil autor)

Sloupce	*Dvojité mezery*	*HTML tagy*	*HTML entity*	*Oddělovací znaky*
company_profile	105	0	2775	1991
description	359	0	3834	3525
requirements	134	0	2152	2233
benefits	104	0	1324	1256
title	165	0	0	2202

Následná analýza všech výše zmíněných textových sloupců ukázala, že data obsahují nejen dvojité mezery a nadbytečné oddělovací znaky, ale také specifické HTML entity. Na druhou stranu nebyly nalezeny žádné HTML značky. Rozdíl mezi HTML entitami a značkami spočívá v tom, že zatímco HTML značky definují strukturu a formát obsahu webové stránky, příkladem mohou být „<h1>“ nebo „<p>“, HTML entity představují specifické kódové sekvence symbolizující charaktery, které nelze přímo vložit do HTML kódu, jako je to v případě „&“ reprezentující znak „&“. Zvláště výrazné byly nadbytečné oddělovací znaky na začátku a konci řetězců, zejména ve sloupcích „description” a „requirements“.

Na základě těchto zjištění je doporučeno provést důkladné čištění textových sloupců s cílem odstranit nadbytečné oddělovací znaky, dvojité mezery a nahradit HTML entity jejich odpovídajícími znaky.

Sloupce „employment type“, „required experience“ a „required education“

Ve sloupcích „employment_type“, „required_experience“ a „required_education“ bylo identifikováno několik unikátních hodnot. Vzhledem k omezenému počtu těchto hodnot se konsolidace hodnot v jednotlivých kategoriích nejeví jako potřebná. Data v těchto sloupcích se zdají být čistá a strukturovaná.

Konkrétně ve sloupci „employment_type“ je pět unikátních hodnot, s „Full-time“ jako dominantní hodnotou. Více než 34 % hodnot v tomto sloupci chybí. Ve sloupci „required_experience“ je sedm unikátních hodnot s nejčastějším zastoupením „Mid-Senior level“, přičemž více než 38 % hodnot chybí. V „required_education“ se objevuje 13 unikátních hodnot, s „Bachelor's Degree“ jako nejčastější hodnotou, avšak více než 45 % hodnot zde chybí. Tyto chybějící hodnoty mohou hrát klíčovou roli při predikci, zda je nabídka práce podvodná, či nikoli. Absence některých informací může ukazovat na nedbalost nebo záměrnou neúplnost při vytváření nabídky, což může být indikátorem podvodných nabídek.

Z analýzy distribuce těchto sloupců je důležité zdůraznit, že některé hodnoty, například „Some High School Coursework“ v „required_education“, vykazují vysokou míru podvodných nabídek. Tento vysoký podíl může být však ovlivněn nízkým výskytem těchto hodnot v datech. Celkově lze konstatovat, že tato data nepotřebují významné čištění či úpravy. Pro modelování je však nezbytné provést techniku 1 z n („one-hot encoding“) těchto kategorií, aby byly správně interpretovány predikčními modely.

Numerické sloupce

V datové sadě nalezneme numerické sloupce: „telecommuting“, „has_company_logo“, „has_questions“, „fraudulent“ a „job_id“. Kromě sloupce „job_id“, který funguje jako unikátní identifikátor pro nabídky práce, obsahují všechny ostatní numerické sloupce pouze dvě unikátní hodnoty. Tato charakteristika naznačuje, že se jedná o binární proměnné. Při pohledu na data

v těchto numerických sloupcích se jeví jako čistá a konzistentní, což znamená, že v rámci přípravy dat pro analýzu není potřeba další čištění. Avšak pro účely modelování bude vhodné odstranit sloupec „job_id“, neboť neposkytuje žádnou relevantní informaci pro prediktivní modelování.

5.4.4 Závěrečné zhodnocení kvality dat

V průběhu zkoumání datové sady byly identifikovány klíčové oblasti týkající se kvality dat. Kategorické sloupce, zejména ty s větším počtem unikátních hodnot, mohou obsahovat nekonzistence v zápisu, což může komplikovat analýzu. Je nezbytné zajistit, aby hodnoty v těchto sloupcích byly konzistentní a jednoznačné. Textové sloupce ukázaly přítomnost nadbytečných oddělovacích znaků a HTML entit, což vyžaduje důkladné čištění pro správnou analýzu.

Co se týká numerických sloupců, s výjimkou unikátních identifikátorů, většina z nich obsahuje binární hodnoty, což zjednodušuje analýzu a modelování. Tyto sloupce vypadají čistě i konzistentně a nevyžadují další úpravy. Chybějící hodnoty v některých sloupcích mohou ovlivnit kvalitu predikčních modelů. Je proto důležité zvážit různé metody imputace nebo se rozhodnout je ponechat v původním stavu, pokud to nebude negativně ovlivňovat výsledky analýzy. Tyto kroky budou provedeny v rámci přípravy dat.

Z celkového hodnocení plyne, že i když datová sada vyžaduje určité úpravy a čištění, její kvalita je dostatečná pro další analýzu a modelování. V následujících fázích analýzy bude nezbytné provést doporučené úpravy dat, aby bylo zajištěno, že výsledky analýzy budou co možná nejpřesnější.

5.5 Příprava dat

Na základě analýz a zjištění z předchozích fází CRISP-DM budou aplikována specifická doporučení pro optimalizaci datové sady. Začátek této fáze je věnován čištění dat, během něhož budou detailně popsány kroky a techniky použité pro úpravu jednotlivých sloupců v datové sadě. Veškeré provedené úpravy budou ukládány do nového dataframe nazvaného „df_processing“, což umožní zachovat původní data v nezměněné formě pro případnou další referenci.

Po čištění dat následuje fáze konstrukce dat, kde budou zpracovány nové proměnné a případné transformace stávajících dat. Ačkoli integrace dat je důležitou součástí mnoha projektů analýzy dat, v tomto konkrétním případě nebude potřeba, jelikož je k dispozici pouze jedna datová sada, a není tedy nutné kombinovat data z různých zdrojů.
Na závěr tohoto procesu bude pozornost věnována formátování dat, aby byla zajištěna jejich správná struktura a kompatibilita pro následující fáze analýzy. Cílem je zajistit, aby data byla připravena tak, aby co nejvíce odpovídala potřebám následného modelování a analýzy.

5.5.1 Čištění dat

Na základě zprávy kvality dat bylo zjištěno, že existují určité oblasti v datové sadě, které vyžadují zásah v oblasti čištění dat. Zvláště kategorické sloupce s velkým množstvím unikátních hodnot ukazují potenciální nekonzistence a textové sloupce mohou obsahovat nadbytečné znaky a HTML entity. Tyto problémy mohou zkreslit výsledky analýzy a představují hlavní cíle tohoto procesu čištění.
V následujících částech budou procházeny jednotlivé sloupce a budou aplikovány úpravy tak, jak bylo navrhnuto v rámci zprávy kvality dat, s cílem zajistit konzistenci a spolehlivost dat pro další analýzu.

Sloupec „function“

V rámci čištění dat ve sloupci „function“ bylo nutné se zaměřit na specifické problémy a výzvy identifikované v analýze kvality dat. Předchozí zkoumání ukázalo potenciální spojitost mezi sloupci „function“ a „industry“ prostřednictvím chí-kvadrát testu. Tento test potvrdil významnou spojitost mezi těmito dvěma sloupci na hladině významnosti 0.05. S ohledem na tuto zjištěnou spojitost byla vytvořena kontingenční tabulka, která zobrazuje četnost kombinací hodnot mezi oběma sloupci. Tento krok poskytl hlubší přehled o vzájemných vztazích mezi „function“ a „industry“.
Na základě těchto informací byla navržena metoda imputace chybějících hodnot ve sloupci „function“ založená na nejčastějším výskytu funkcí v rámci průmyslových odvětví. Pro otestování přesnosti imputace bylo nezbytné porovnat

imputované hodnoty se skutečnými hodnotami v datech. K tomuto účelu byl vytvořen nový testovací sloupec „function_test“. Do tohoto sloupce byly imputovány hodnoty podle stejné logiky, která byla použita pro imputaci chybějících hodnot ve sloupci „function_imputation“. Byla tedy přiřazena nejčastější hodnota „function“ pro dané „industry“. Následně byly porovnány imputované hodnoty ve sloupci „function_test“ s původními hodnotami ve sloupci „function“.

Toto porovnání umožnilo odhalit, v kolika případech se skutečná hodnota „function“ shodovala s nejčastější hodnotou „function“ pro daný průmysl. Výsledky tohoto porovnání ukázaly, že v přibližně 42.66 % případů byla imputovaná hodnota správná. V zbývajících 57.34 % případů se imputovaná hodnota lišila od skutečné hodnoty. Vzhledem k těmto zjištěním bylo rozhodnuto neintegrovat upravený sloupec „function_test“ do finální datové sady pro další analýzu, protože přesnost této metody nebyla dostatečně vysoká.

Sloupec „department"

Jedním z prvních kroků k optimalizaci sloupce „department“ bylo převedení všech hodnot na malá písmena. Odstraněny byly přebytečné oddělovací znaky, což snížilo počet unikátních hodnot ze 1 337 na 1 224. Taková úprava odhalila skryté nekonzistence v důsledku rozdílné velikosti písmen v názvech oddělení.

Dřívější analýza kvality dat odhalila některé varianty klíčových slov ve sloupci „department“, které by mohly být interpretovány jako ekvivalentní, jak bylo uvedeno ve zprávě o kvalitě dat. Na základě těchto poznatků byly vybrané varianty sjednoceny. Například všechny výskyty slova „dev“ byly nahrazeny slovem „development“, výraz „tech“ byly nahrazen slovem „technology“, „it“ bylo změněno na „information technology“, slovo „hr“ bylo substituováno slovem „human resources“ a slovo „administrative“ bylo nahrazeno slovem „administration“. Pro zachování jedinečnosti a specifity různých oddělení bylo však rozhodnuto nesjednocovat všechny názvy oddělení, ale pouze ty, jejichž ekvivalence byla zřejmá ze zprávy o kvalitě dat. Po těchto úpravách zůstalo celkem 1 217 unikátních hodnot ve sloupci „department“.

Ačkoliv se počet unikátních hodnot snížil, stále zde zůstává vysoký počet těchto hodnot. Tato variabilita odráží komplexnost a specifikum různých oddělení v organizacích a může být vnímána jako přirozená vlastnost těchto dat, která by neměla narušit následnou analýzu. V dalších fázích byl tento sloupec zařazen k textovým datům.

Sloupce „title“, „description“, „requirements“, „benefits“ a „company profile“

Sloupce „title“, „description“, „requirements“, „benefits“ a „company profile“ ve zprávě o kvalitě byly tyto sloupce označeny jako textová data. Bylo zjištěno, že obsahují některé nekonzistence, například dvojité mezery, HTML entity či nadbytečné oddělovací znaky. Aby byla zajištěna kvalita dat a jejich připravenost pro další analýzu, rozhodlo se provést detailní čištění těchto sloupců. K čištění dat byla navržena funkce s názvem „clean_text_data“, jejíž úkolem bylo eliminovat dvojité mezery, převádět HTML entity na odpovídající znaky a odstraňovat nadbytečné oddělovací znaky z počátku a konce řetězců.

Během tohoto procesu se ukázalo, že některé HTML entity mohou být vnořené nebo mít komplexní strukturu, což znesnadňovalo jejich kompletní odstranění. V důsledku toho byla funkce „clean_text_data“ upravena tak, aby prováděla iterativní čištění. Tato funkce byla aplikována na všechny zmíněné sloupce.

Po těchto úpravách následovala kontrola v nově vytvořené datové sadě „df_cleansed“, zda byly všechny nežádoucí hodnoty odstraněny. K identifikaci řádků, jež obsahovaly HTML entity, byla použita funkce „find_html_tags“. Tato analýza umožnila identifikovat a extrahovat problematické řádky z každého sloupce, což zaručilo, že upravená data jsou nyní čistší a lépe strukturovaná.

5.5.2 Konstrukce dat

Během konstrukce dat byly nejprve odvozeny nové sloupce s cílem zlepšit interpretaci a srozumitelnost datové sady. Následně se využily různé kódovací techniky k převedení kategoriálních proměnných do formátu vhodného pro modelovací algoritmy.

Sloupec „salary range“

Na základě závěrů z kontroly kvality dat byl vytvořen nový binární sloupec s názvem „has_salary", který stanovuje, zda nabídka práce obsahuje informaci o platu. Tato binární reprezentace může být užitečná pro další analýzu a modelování, neboť může odrážet strategii či záměry zaměstnavatele při tvorbě inzerátu.

Sloupec „location“

Na základě analýzy kvality dat bylo zjištěno nesrovnalosti v hodnotách sloupce „location“, které se lišily v závislosti na zemi. Aby bylo možné získat konzistentní a strukturované údaje o lokalitě, bylo rozhodnuto rozčlenit hodnoty ve sloupci „location“ do tří nových sloupců: „location_country“, „location_state“ a „location_city“. K tomuto účelu byla vytvořena funkce využívající metodu str.split() s parametrem n=2. Tato metoda rozděluje hodnoty ve sloupci „location“ na maximálně tři části – zemi, stát a město. Tyto části byly poté přiřazeny do nově vytvořených sloupců. Pro odstranění oddělovacích znaků byla v hodnotách sloupců „location_state“ a „location_city“ použita metoda str.strip().
Po dokončení zpracování byl původní sloupec „location“ odstraněn, aby se zabránilo duplicitě informací v datovém souboru. Díky tomu byla data ve sloupci „location“ efektivně restrukturalizována a připravena pro další analýzu a modelování.

Vytvoření sloupec „combined_text“

Během fáze konstrukce dat bylo rozhodnuto spojit všechna klíčová textová data do jediného sloupce, což umožňuje jednodušší zpracování textových dat pro další modelování. Sloupce „title“, „company_profile“, „description“, „requirements“, „benefits“, „department“, „industry“ a „function“ byly sjednoceny do nově vytvořeného sloupce s názvem „combined_text“. Pro každý záznam datové sady byly hodnoty těchto sloupců spojeny do jednoho souvislého textového řetězce. Proces spojení byl navržen tak, aby vynechal chybějící hodnoty a zajistil hladké spojení dostupných textů. Tímto postupem lze snadněji uplatnit techniky

zpracování přirozeného jazyka na sjednocený text a extrahovat z něj informace pro následné modelování.

Kódování kategorických proměnných

Během fáze konstrukce dat byly kategorické proměnné upraveny tak, aby lépe vyhovovaly dalšímu modelování. Klíčovým krokem bylo frekvenční kódování sloupců „location_country“, „location_state“ a „location_city“. Frekvenční kódování představuje techniku, při které se hodnoty kategorické proměnné nahrazují frekvencí jejich výskytu v datové sadě. Výsledné frekvenční hodnoty byly uloženy do nově vytvořených sloupců s příponou „_freq“.

Následovalo mapování sloupců „required_experience“ a „required_education“. Mapování pro „required_experience“ bylo navrženo tak, aby odráželo hierarchii pracovních pozic – od stáží až po manažerské pozice. U „required_education“ bylo upřednostněno mapování založené na postupu ve vzdělání, zahrnující různé stupně od střední školy až po doktorský titul. V obou případech byly kategorické hodnoty převedeny na číselné reprezentace odpovídající úrovni vzdělání či pracovních zkušeností.

V posledním kroku byla aplikována techniky 1 z n („one-hot encoding“) na sloupec „employment_type“. Jedná se o standardní metodu pro konverzi kategorických proměnných, při níž se každá unikátní hodnota v kategoriálním sloupci převede do nového binárního sloupce. V důsledku toho byly zavedeny nové sloupce reprezentující různé typy pracovních vztahů, jako jsou „employment_type_Full-time“ a „employment_type_Part-time“.

5.5.3 Formátování dat

Během fáze formátování dat bylo klíčovým úkolem provést syntaktické úpravy, které obsahovaly převod textu na malá písmena, odstranění nadbytečných oddělovacích znaků a další korekce vedoucí ke zvýšení čitelnosti a konzistence dat. Kromě těchto základních syntaktických změn bylo nutné znovu uplatnit funkci pro čištění dat vytvořenou v dřívějších fázích analýzy. Tato funkce, pojmenovaná „clean_text_data“, odstraňovala konkrétní nepravidelnosti, například dvojité mezery nebo potenciální HTML značky.

Následovala syntaktická úprava pomocí funkce „preprocess_text“, která text převedla na malá písmena, vyřadila nealfanumerické znaky a tzv. stop slova, což jsou běžná slova, jako spojky, zájmena nebo slovesa, která jsou bez většího významového obsahu. Výstup z těchto úprav byl uložen do nového sloupce „preprocessed_text“, což přispělo k větší kvalitě textových dat a připravilo je pro následné analýzy a modelování.

5.5.4 Druhá část konstrukce dat

Po dokončení fáze formátování dat se ukázalo, že lemmatizace je nezbytná k zajištění vyšší kvality a srozumitelnosti textových dat pro modelovací algoritmus. Tento postup je považován za klíčový, neboť lemmatizace podporuje konzistenci textových dat tím, že omezuje variabilitu v textu, zejména snižuje výskyt různých gramatických forem téhož slova. Pro lemmatizaci byla použita knihovna NLTK, jež obsahuje nástroje pro tokenizaci a lemmatizaci textu. Implementována byla funkce „lemmatize_text“, která nejprve tokenizuje text na jednotlivá slova a poté každé slovo lemmatizuje k jeho základní formě. Po lemmatizaci se základní formy slov spojí zpět do textového řetězce, což zajistí konzistentní a jednotnou reprezentaci textu.

Pro vektorizaci byla vybrána metoda TF-IDF (Term Frequency-Inverse Document Frequency), jež patří mezi běžné metody vektorizace textu. Byla využita třída „TfidfVectorizer“ z knihovny „scikit-learn“ s omezeným počtem rysů za účelem efektivity a snížení dimenzionality. Parametr „max_features“ byl nastaven na hodnotu 150, což znamená vytváření matice na základě 150 nejčastějších slov v textových datech. Toto nastavení bylo zvoleno v průběhu iterativního procesu výběru modelu a bude podrobněji popsáno v další fázi modelování. Po vektorizaci byla původní textová reprezentace „lemmatized_text“ odstraněna a výsledná TF-IDF matice byla začleněna do hlavního datového rámce, připravujíc tak komplexní datovou sadu pro následující modelovací kroky.

5.5.5 Výběr dat

Ve fázi výběru dat bylo nutné určit, které proměnné budou nejrelevantnější pro použité modelovací algoritmy. Pro algoritmus náhodného lesa byla zvolena

sada proměnných obsahující „telecommuting“, „has_company_logo“, „has_questions“, „fraudulent“ a „has_salary“. Doplněny byly i kategorické proměnné spojené s „location“, „required_education“ a „required_experience“, stejně jako kódované proměnné pro „employment_type“ a frekvenčně kódované proměnné pro „location“. Důležitým rozlišením je zahrnutí vektorizovaných sloupců vycházejících z „lemmatized_text“, které odráží definovaný počet nejčastějších slov. Při výběru proměnných pro logistickou regresi byla sada téměř shodná s náhodným lesem, avšak bez sloupce „lemmatized_text“. Zvolené proměnné byly vybrány na základě předpokládaného dopadu na model a schopnosti předpovědět cílovou proměnnou.

V dalším kroku výběru dat byla data optimalizována odstraněním sloupců, které již nebyly nezbytné pro analýzu a modelování. V předchozích fázích byly vytvořeny některé pomocné sloupce primárně určené k usnadnění analýzy nebo transformace dat. Tyto sloupce však mohou zvyšovat objem datové sady a potenciálně komplikovat následující fáze zpracování. Z tohoto důvodu bylo rozhodnuto vyřadit sloupce jako „combined_text_cleaned“, „combined_text“, „job_id“, „function_imputation“ a „function_test“.

5.5.6 Závěrečné zhodnocení přípravy dat

Při přípravě dat byla provedena řada kroků k zajištění jejich nejlepší kvality pro další analýzy a modelování. Proces začal čištěním dat, během něhož byly prováděny specifické úpravy jednotlivých sloupců na základě doporučení ze zprávy o kvalitě dat. K zachování původních dat pro budoucí potřeby byly všechny změny aplikovány v novém dataframe nazvaném „df_processing“.

Následně probíhala konstrukce dat: byly vytvářeny nové sloupce pro hlubší analýzu informací a byly aplikovány techniky kódování k transformaci kategorických proměnných do formátu vhodného pro modelování. Rozhodnutí nespojovat data bylo učiněno, jelikož všechny nezbytné informace již byly obsaženy v jedné datové sadě.

Během fáze formátování dat byly prováděny syntaktické úpravy k zajištění správné struktury a kompatibility dat s analýzou. V souladu s metodikou CRISP-DM může být často nutné přehodnocovat a upravovat již realizované kroky, vracet

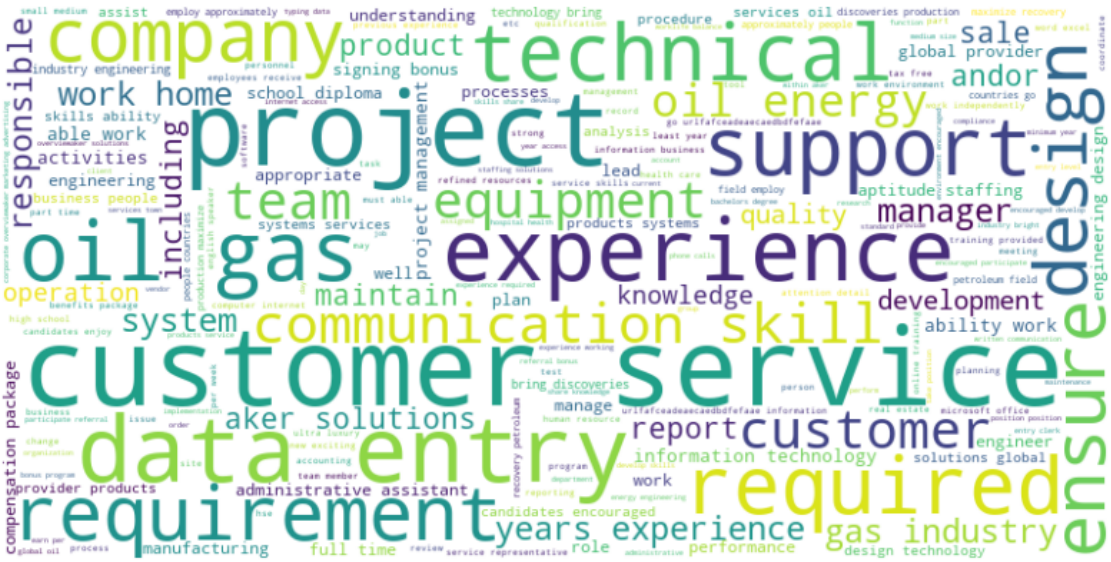
se ke starším fázím a optimalizovat je. V tomto případě, ačkoliv se zdálo, že fáze formátování dat byla ukončena, vyskytla se nutnost vrátit se zpět k fázi konstrukce dat. Důvodem byly potřebné syntaktické úpravy a transformace, závislé na formátování textových dat, jež měly data upravit pro další analýzu.

Závěrem procesu přípravy dat přišla fáze výběru dat. Během ní bylo rozhodnuto, které proměnné budou nejvhodnější pro další modelovací algoritmy. V tomto kontextu proběhl výběr proměnných jak pro náhodný les, tak pro logistickou regresi. Shrnutím celého procesu je, že proces přípravy dat byl pečlivý a důkladný, s cílem zajistit, aby data byla co nejlépe připravena pro následné modelování a analýzu.

5.5.7 Slovní mraky

Po dokončení čištění dat a odstranění běžných, avšak významově méně relevantních slov je vhodné se zaměřit na analýzu, která by odpovídala konkrétnímu dílčímu cíli práce: Která slova se vyskytují nejčastěji v textech podvodných a nepodvodných pracovních nabídek?

Ke grafickému znázornění těchto slov slouží vizualizace ve formě slovních mraků, jež graficky zobrazuje četnost slov z textových sloupců jak pro pravdivé, tak i pro podvodné pracovní nabídky. Díky dřívějším úpravám dat lze předpokládat, že slova zobrazená ve slovních mracích odráží obsah a charakteristiky zkoumaných textů, a jsou proto zbavena zkreslení způsobeného často se opakujícími, ale v kontextu méně důležitými slovy.



Obrázek 12 Vizualizace pomocí slovních mraků nejčastějších slov pro podvodné pracovní nabídky (na základě dat z platformy Kaggle.com sestavil autor)

Z analýzy zobrazené na obrázku č. 12 vyplývá, že v podvodných pracovních nabídkách se převážně objevují termíny jako „project“, „customer“, „services“ a „experience“ a celá řada dalších. Fráze „oil gas“ má rovněž výrazné zastoupení, což může signalizovat spojitost s odvětvím „Oil & Energy“. Uvedené odvětví bylo v minulosti označeno jako segment s větším množstvím podezřelých nabídek v rámci provedených analýz.

Obrázek 13 Vizualizace pomocí slovních mraků nejčastějších slov pro pravdivé pracovní nabídky (na základě dat z platformy Kaggle.com sestavil autor)

Porovnáním obrázků č. 12 a č. 13 zřetelně vyniká, že pravdivé nabídky prezentují širší spektrum slov s méně častým opakováním, což může být důsledkem nevyváženosti datové sady. Ačkoliv se některé termíny jako „customer“ a „services“ objevují v obou typech nabídek, termíny „information“, „technology“, „communication“ a „skill“ charakterizují primárně pravdivé pracovní nabídky.

Na závěr lze říct, že z vizualizace je patrné, že některá slova, např. „project“, „experience“ a „oil gas“, se často objevují v podvodných nabídkách, zatímco slova jako „information“, „technology“, „communication“ a „skill“ mají větší zastoupení v pravdivých nabídkách. Zatímco se termíny typu „customer“ a „services“ vyskytují v obou kategoriích, mohla by charakteristika jejich výskytu a kontext, v němž se objevují, nabízet odlišné interpretace, protože tyto termíny mohou mít různý význam v závislosti na kontextu vět, v nichž se vyskytují.

5.6 Modelování

Po dokončení fáze předzpracování dat, kdy byla data vyčištěna, transformována a připravena pro analýzu, následuje klíčová fáze v procesu analýzy dat – modelování. Na základě typu dat v datové sadě byly vybrány dva modely, které byly využity pro predikci. Vzhledem k tomu, že se jedná o predikci binární hodnoty, logistická regrese byla přirozenou volbou. Její výsledky lze snadno interpretovat, což umožňuje rozpoznání klíčových rysů ovlivňujících predikce. Logistická regrese je navíc rychlá a efektivní jak v tréninku, tak v predikci.

Druhým zvoleným modelem je náhodný les – souborová metoda spojující více rozhodovacích stromů za účelem dosažení větší přesnosti. Tento model vyniká v automatickém zpracování rozsáhlých atributů a v identifikaci nejrelevantnějších z nich, což je zvláště užitečné při práci s daty s mnoha proměnnými, jako je tomu v případě textových sloupců.

Cílem výběru těchto modelů bylo spojit lineární a nelineární přístupy k problému a současně využít přednosti obou metod. Zatímco logistická regrese nabízí srozumitelné výsledky, náhodný les přináší robustnost a schopnost zachytit složité vztahy v datech.

Pro modelování byla data rozčleněna do dvou hlavních proměnných: X a Y. Proměnná Y představuje cílovou hodnotu a indikuje, zda je nabídka podvodná (1),

či pravdivá (0). Proměnná X obsahuje všechny další sloupce vybrané jako vstupní atributy pro modelování.

K rozdělení dat slouží funkce „train_test_split“ z knihovny scikit-learn, která zajišťuje efektivní a náhodné rozčlenění dat. V tomto případě byly aplikovány následující klíčové parametry funkce:

- „arrays" – Sekvence datových množin určených k rozdělení na trénovací a testovací podmnožiny;
- „test_size" – Definice poměru dat, jež mají být vyhrazena pro testovací množinu. V tomto případě byla tato hodnota nastavena na 0.25, což znamená, že 25 % dat bylo alokováno pro testování, zatímco zbývajících 75 % pro trénink;
- „random_state" – Tento parametr slouží k nastavení semena pro generátor náhodných čísel, čímž se zabezpečuje reprodukovatelnost rozdělení dat. Konkrétní hodnota, např. 42, zaručuje, že při opakovaném rozdělení stejné datové sady zůstane rozdělení konzistentní a shodné.

Rozdělením datové sady pomocí metody „train_test_split" byly získány trénovací a testovací části v poměru 75 % k 25 %. Nastavením parametru „random_state" byla zajištěna konzistentnost rozdělení, což znamená, že při každém opakovaném použití funkce na stejné datové sadě byly obdrženy identické trénovací a testovací množiny. Toto stabilní rozdělení dat vytvořilo pevný základ pro vývoj a validaci predikčních modelů. V následujících podkapitolách jsou podrobněji popsány modely včetně jejich výsledků a zhodnocení.

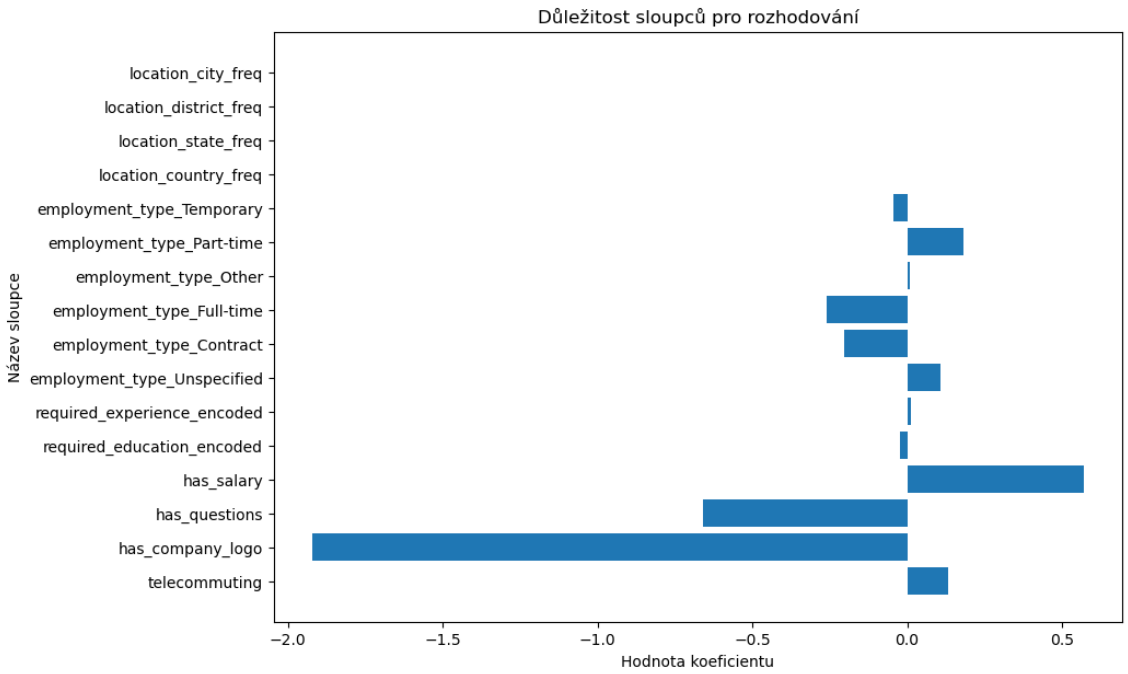
5.6.1 Logistická regrese

Pro model logistické regrese byla použita datová sada obsahující předem definované sloupce z fáze výběru dat. Vzhledem k omezením logistické regrese, zejména co se týká schopnosti zpracovávat velké množství vysoko dimenzionálních dat, bylo rozhodnuto nezahrnovat do modelu vektorizovaná textová data.

Logistická regrese odhadovala pravděpodobnosti zařazení každého záznamu do jedné ze dvou možných kategorií, tedy zda je pracovní nabídka podvodná, či

pravdivá. Koeficienty tohoto modelu popisují, jak nezávislé proměnné ovlivňují logitovou transformaci výsledku. Na základě těchto koeficientů lze pochopit, jak jednotlivé proměnné ovlivňují pravděpodobnost klasifikace do určité kategorie. Jak lze vidět na obrázku č. 14 „has_company_logo“ s hodnotou −1.920629, tato hodnota naznačuje, že pokud nabídka práce neobsahuje logo firmy, zvyšuje se riziko, že by mohlo jít o podvodnou nabídku. Na druhou stranu koeficient pro „has_salary“ s hodnotou 0.568852 naznačuje, že nabídky obsahující platové ohodnocení mají vyšší pravděpodobnost být považovány za pravdivé pracovní nabídky.

Je však důležité zdůraznit, že přestože koeficienty mohou poskytnout cenné informace o vztazích mezi proměnnými a výsledkem, neměly by být považovány za konečný důkaz kauzality.



Obrázek 14 Zobrazení pozitivních a negativních hodnot koeficientů pro jednotlivé sloupce (na základě dat z platformy Kaggle.com sestavil autor)

Při měření pomocí metriky přesnosti byl výsledek velice pozitivní s hodnotou 0.9475, avšak tyto hodnoty jsou zkreslené vzhledem k nerovnováze datové sady. V konečné fázi byla účinnost modelu logistické regrese ověřena pomocí F1 skóre, které dosáhlo hodnoty 0.0176. Tato hodnota je velmi nízká a svědčí o tom, že logistická regrese není vhodná pro analýzu této datové sady. Jednou z příčin může

být skutečnost, že logistická regrese nejlépe pracuje s binárními daty, zatímco mnoho sloupců obsahujících textové hodnoty v těchto datech bylo v analýze opomenuto. Kromě toho je logistická regrese lineárním klasifikátorem, což znamená, že vytváří rozhodovací hranici na základě lineární kombinace vstupních proměnných. Vzhledem k charakteru datové sady se zdá, že data neumožňují jednoduché lineární oddělení mezi pravdivými a podvodnými pracovními nabídkami. Z tohoto důvodu se jeví potřebné využít nelineární klasifikátor, jako je náhodný les, který bude schopen vzít v potaz více proměnných.

5.6.2 Náhodný les

Pro model náhodného lesa byla použita datová sada obsahující vektorizované sloupce, které byly odvozeny z pole „lemmatized_text“. Tato datová sada dále zahrnovala další nezávislé proměnné shodné s těmi, které byly použity pro logistickou regresi. K vektorizaci tohoto textu byla použita třída „TfidfVectorizer“, kde bylo zapotřebí nastavení hodnoty parametru „max_features“, jež určuje maximální počet slov zahrnutých do vektorizace a následně do samotného modelu. Parametr „max_features“ byl nastaven na hodnotu 150, což znamená, že byl zohledněn pouze tento počet nejčastěji se vyskytujících slov z celého textu. Optimalizace tohoto parametru byla provedena na základě experimentu s různými hodnotami: 75, 100, 125, 150, 175 a 200 slov. Z následující tabulky je patrné, jak se mění skóre modelu v závislosti na počtu nejčastějších slov:

Tabulka 3 Skóre modelu v závislosti na počtu nejčastějších slov (na základě dat z platformy Kaggle.com sestavil autor)

Počet slov	*Skóre*
75	0.9736
100	0.9740
125	0.9736
150	0.9745
175	0.9738
200	0.9740

Z uvedených výsledků je patrné, že optimální hodnotou parametru je 150 nejčastějších slov. Toto nastavení bylo tedy použito pro finální modelování s funkcí „Random ForestClassifier“. Pro dosažení nejvyšší přesnosti byly upraveny některé další hyperparametry modelu. Hyperparametry jsou specifické parametry, které se nespecifikují na základě dat, ale je nutné je zadat externě. Patří mezi ně:

- bootstrap – Tento hyperparametr určuje, zda se bude při konstrukci stromů v lese používat náhodný výběr podmnožin tréninkových dat s opakováním. Pokud je nastaveno na „False“, celá datová sada se použije pro výstavbu každého stromu.
- criterion – Určuje metriku kvality rozdělení, pro klasifikační úkoly je hodnota buď „gini“ nebo „entropy". Volba kritéria může ovlivnit, jak stromy v lese rozhodují o rozdělení, a tedy i o výsledné kvalitě modelu.
- max_depth – Určuje maximální hloubku stromů v lese. Nastavením této hodnoty pomáháme zamezit přeučení modelu.
- max_features – Určuje maximální počet vlastností, které model zohlední při rozdělení uzlu, ovlivňujíc tím rozmanitost stromů v lese.
- min_samples_leaf – Určuje minimální počet vzorků, které musí být v listovém uzlu, čímž pomáhá zamezit přílišné fragmentaci dat a přeučení modelu.
- min_samples_split – Určuje minimální počet datových záznamů potřebných pro rozhodnutí o rozdělení uzlu stromu. Podobně jako předchozí parametr pomáhá i tento regulovat komplexnost stromů a zároveň zabraňuje jeho přeučení.
- n_estimators – Určuje počet stromů, které budou vytvořeny v náhodném lese. Zatímco větší množství stromů může zvýšit přesnost modelu, může také prodloužit jeho dobu učení.
- n_jobs – Stanovuje počet procesorových jader využitých při výpočtu. Nastavením hodnoty na −1 se využijí všechna dostupná jádra, což může významně zrychlit výpočet.

- class_weight – Umožňuje modelu vážit třídy během tréninku. Při nastavení na „balanced“ se váhy automaticky přizpůsobí na základě frekvence tříd v tréninkových datech, což je užitečné pro nevyvážené sady.

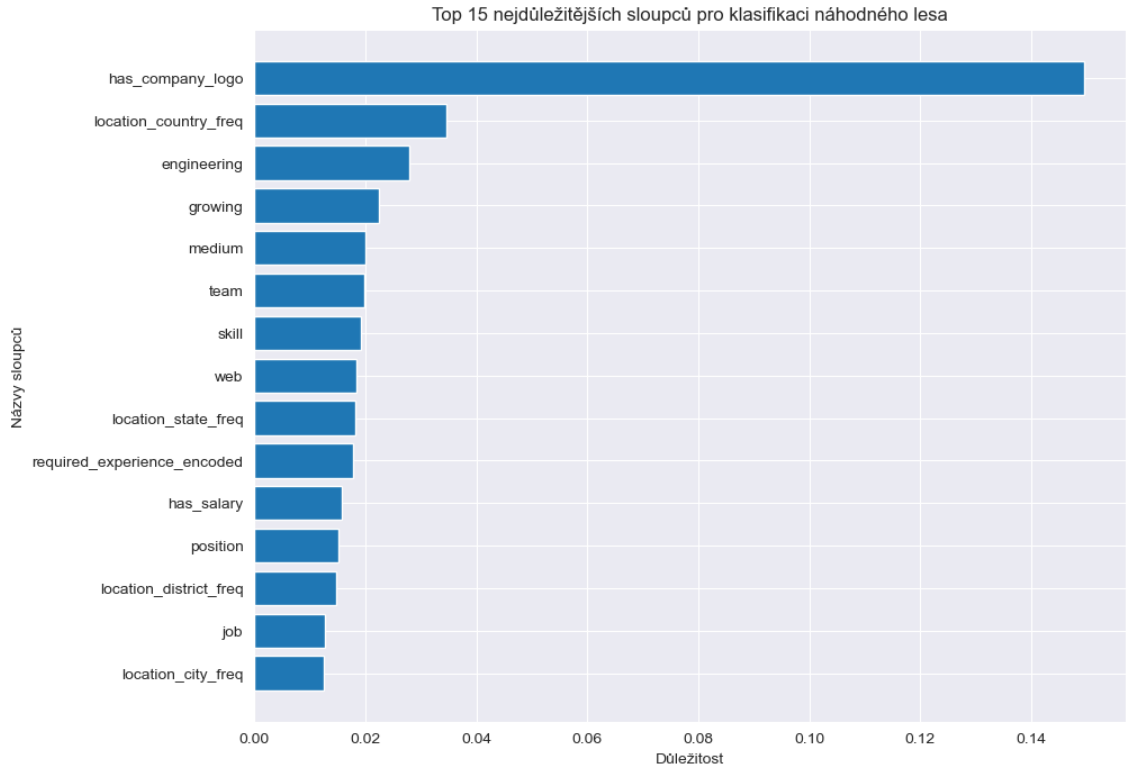
Optimalizace hyperparametrů byla provedena pomocí metody „RandomizedSearchCV“, která vybírá náhodné kombinace hyperparametrů a vyhodnocuje výkonnost modelu pro každou kombinaci s využitím křížové validace. Nejlepší nastavení pro optimalizaci modelu bylo následující:

- bootstrap: False
- criterion: entropy
- max_depth: 32
- max_features: 45
- min_samples_leaf: 6
- min_samples_split: 8
- n_estimators: 499
- class_weight: balanced

Optimalizované hyperparametry byly použity pro finální modelování. S upraveným nastavením dosáhl model hodnoty F1 skóre 0.8127, což je považováno za dobrý výsledek, protože hodnota F1 skóre se pohybuje v rozmezí od 0 do 1, kde 1 značí perfektní přesnost a úplnost. Hodnota 0.8127 tedy indikuje vysokou úroveň správně klasifikovaných instancí a dobrou rovnováhu mezi přesností a úplností modelu.

Po vyhodnocení modelu následovala analýza sloupců, které nejvíce ovlivňují jeho výsledky. Pro lepší vizualizaci těchto významů byl vytvořen graf prezentující 15 nejdůležitějších sloupců řazených podle jejich vlivu. Z těchto informací lze získat cenný vhled do atributů, které model považuje za klíčové při svém rozhodování. Avšak je důležité poznamenat, že na základě tohoto grafu není možné jednoznačně určit, jaké hodnoty těchto atributů vedou k určité klasifikaci. Graf zobrazuje pouze míru jejich důležitosti, ale neinformuje o tom, zda daný atribut zvyšuje, či snižuje pravděpodobnost, že nabídka práce je pravdivá, nebo podvodná, protože se nejedná o hodnoty vyjadřující pravděpodobnost. Na grafu z obrázku č. 15 je

patrné, že některé z těchto sloupců mají zásadní význam pro model. Například sloupec „has_company_logo“ napovídá, že přítomnost loga společnosti v inzerátu může mít významný vliv na klasifikaci, což může odrážet důvěryhodnost a profesionální prezentaci společnosti. Dále klíčová slova jako „engineering“ nebo „data“ poukazují na to, že obsah inzerátu, zvláště ve vztahu k technologickým a datovým oborům, má značný dopad na modelování.



Obrázek 15 Nejdůležitějších 15 sloupců pro klasifikační model náhodného lesa (na základě dat z platformy Kaggle.com sestavil autor)

Z toho vyplývá, že model kladl zvláštní důraz na některá slova, což naznačuje, že textový obsah má zásadní význam pro klasifikaci. Tyto poznatky ukazují, jak kombinace textového obsahu a dalších atributů inzerátu určují výsledné rozhodování modelu.

5.6.3 Závěrečné vyhodnocení výsledků modelování

V rámci modelování byly zkoumány dvě predikční techniky: logistická regrese a náhodný les. Tyto modely byly vybrány na základě specifik dané datové sady a očekávaného cílového výstupu. Přestože logistická regrese je obvykle známá

pro řešení binárních klasifikačních úloh, v tomto případě se ukázala jako méně účinná. Její nedostatečná výkonnost v této datové sadě ukázala na potřebu implementace robustnějšího modelu, schopného odhalit komplexnější vzory v datech. Podle výsledků logistické regrese, zobrazené v tabulce č. 4, byla pro pravdivé nabídky dosažena přesnost 95 % a úplnost 100 % s F1 skóre 0.97. Pro podvodné nabídky však model vykázal nízkou přesnost 33 % a úplnost pouhé 1 %, což vedlo k F1 skóre 0.02.

Tabulka 4 Metriky logistické regrese (na základě dat z platformy Kaggle.com sestavil autor)

Třída	*Přesnost*	*Úplnost*	*F1 skóre*	*Počet záznamů*	*Správně klasifikováno*
0	0.95	1.00	0.97	4171	4167
1	0.33	0.01	0.02	229	2

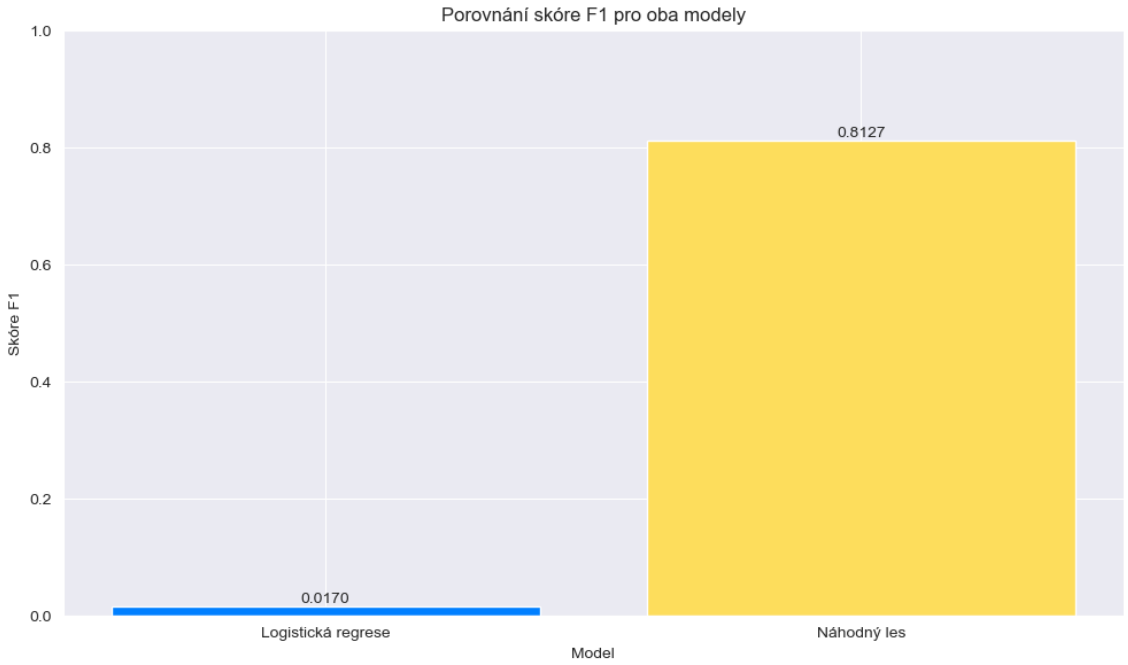
V tomto kontextu se jasně ukázalo, jak je důležité vybrat vhodnou metriku pro hodnocení modelu. I když metrika přesnosti ukazovala hodnotu 0.9475, což by mohlo na první pohled vypadat jako výtečný výsledek, ve skutečnosti nezachytila skutečný stav věci. Důvodem je nevyváženost datové sady. Většina záznamů v datové sadě je z kategorie pravdivých nabídek, což vedlo k tomu, že predikční model byl silně nakloněn k predikci této třídy. V důsledku toho dosáhl model vysoké přesnosti, protože správně klasifikoval většinovou třídu, ale selhal v identifikaci méně zastoupených podvodných nabídek.

Na druhou stranu, jak je patrné z tabulky č. 5, se model náhodného lesa ukázal být mnohem úspěšnějším ve zpracování zkoumané datové sady. Jeho výkonnost byla zvláště patrná u pravdivých nabídek, kde dosáhl vysoké přesnosti 99 % a úplnosti 100 %. Co se týče podvodných nabídek, model dosáhl přesnosti 91 %, ale jeho schopnost správně identifikovat všechny instance této třídy byla nižší, a to s hodnotou úplnosti 73 %. I přes tento nedostatek byl náhodný les schopen poskytnout dobrý kompromis mezi přesností a úplností, jak dokládá F1 skóre 0.81.

Tabulka 5 Metriky klasifikace náhodného lesa (na základě dat z platformy Kaggle.com sestavil autor)

Třída	*Přesnost*	*Úplnost*	*F1 skóre*	*Počet záznamů*	*Správně klasifikováno*
0	0.99	1.00	0.99	4171	4154
1	0.91	0.73	0.81	229	168

Na obrázku č. 16 je vidět sloupcový graf, který porovnává F1 skóre pro oba modely. Z tabulek je patrné, že logistická regrese správně klasifikovala pouze 2 podvodné nabídky, zatímco náhodný les správně identifikoval 168 podvodných nabídek z celkového počtu 229.



Obrázek 16 Sloupcové porovnání F1 skóre pro náhodný les a logistickou regresi (na základě dat z platformy Kaggle.com sestavil autor)

Vzhledem k výsledkům modelování by mohl být užitečný algoritmus „Light Gradient Boosting Machine (LGBM)“, jenž je optimalizován pro vysoký výkon a může zpracovávat velké datové sady. Jeho schopnost přímo pracovat s kategorickými daty a jeho adaptivní povaha by mohly být klíčové pro identifikaci složitějších vzorů ve zkoumané datové sadě, zejména co se týče podvodných nabídek.

Z výsledků modelování vyplývá, že náhodný les nabídl uspokojivé řešení pro predikci podvodných pracovních nabídek, avšak i přes tyto pozitivní výsledky

stále existuje prostor pro optimalizaci. V rámci budoucího výzkumu by bylo vhodné zvážit vzhledem k dobrým výsledkům dosaženým s náhodným lesem implementaci jiných modelů z kategorie stromových algoritmů. Tyto modely by mohly nabídnout robustnější a preciznější výsledky predikce.

6 Shrnutí výsledků

V rámci této diplomové práce bylo hlavním cílem vytvoření prediktivního modelu schopného rozpoznávat pravdivé a podvodné pracovní nabídky. Stanoveny byly také dílčí cíle, na jejichž základě byla vytvořena analytická odpověď. Výsledky pro hlavní i dílčí cíle práce jsou níže shrnuty.

Predikce pravdivých a podvodných pracovních nabídek

Při modelování byly zkoumány dvě predikční techniky: logistická regrese a náhodný les. Ačkoliv je logistická regrese často preferována pro binární klasifikační úkoly, v daném kontextu se ukázala být méně efektivní s F1 skóre 0.0170, což signalizuje její omezenou schopnost identifikace podvodných nabídek. Na druhou stranu náhodný les dosáhl F skóre 0.8127, kdy byl schopen klasifikovat správně 168 podvodných nabídek z celkového počtu 229. Toto vysoké F1 skóre naznačuje, že náhodný les je v tomto případě vhodnější technikou pro detekci podvodných pracovních nabídek.

Jaký průmysl vykazuje největší počet podvodných pracovních nabídek?

Z analýzy vyplývá, že průmysl „Oil & Energy“ je odvětvím s nejvyšším počtem podvodných nabídek. Toto zjištění poukazuje na specifické riziko v tomto odvětví.

Která pracovní pozice je nejčastěji spojována s podvodnými nabídkami?

V oblasti administrativy byl identifikován nejvyšší podíl podvodných nabídek ve srovnání s celkovým počtem, následovaný odvětvím účetnictví/audit. Tyto oblasti se jeví jako časté cíle podvodníků, možná v důsledku vysoké poptávky po zaměstnancích v těchto sektorech.

Která slova se vyskytují nejčastěji v textech podvodných a nepodvodných pracovních nabídek?

Pro vizualizaci četnosti slov v textech nabídek byly vytvořeny slovní mraky. Ty odhalily výrazné rozdíly mezi podvodnými a pravdivými nabídkami. Ve slovních mracích podvodných nabídek se často objevují termíny jako „project“, „experience“

a „oil gas“. Naopak slovní mraky pravdivých nabídek preferují slova jako „information“, „technology“, „communication“ a „skill“. Slova „customer“ a „services“ se objevují v obou typech nabídek, ale jejich kontext může být odlišný, což vyžaduje další zkoumání použití těchto termínů a kontextu jejich výskytu. Taková analýza může poskytnout další vhled do technik a strategií, které podvodníci používají při tvorbě svých inzerátů.

7 Závěr

V této diplomové práci bylo cílem hlouběji proniknout do problematiky dobývání znalostí z databází a následně prezentovat predikční model založený na reálné datové sadě. Vybraný datový soubor dostupný na platformě Kaggle.com sloužil nejen jako základ pro analýzu, ale také pro vytvoření predikčního modelu s cílem identifikovat podvodné nabídky.

V teoretické části jsou prezentovány základní informace o dobývání dat s důrazem na metodiku CRISP-DM. Díky šesti fázím této metodiky lze systematicky strukturovat praktickou část práce a přistupovat k analýze dat. Teoretická část dále rozlišuje mezi predikčním a deskriptivním dolováním dat s hlavním zaměřením na predikční metody. Kromě toho jsou představeny metriky hodnotící úspěšnost modelu.

Druhá část práce se věnuje praktickému zpracování dat s využitím metodiky CRISP-DM. Analýza začíná počátečním porozuměním doméně a směřuje k vyhodnocení modelu. Implementační fáze byla vynechána z důvodu její náročnosti. Avšak, tato fáze nabízí potenciální směr pro budoucí rozšíření a prohloubení této práce. Klíčovým aspektem praktické části je aplikace teoretických znalostí s využitím odpovídajících technologií a nástrojů. Jako hlavní nástroj byl vybrán programovací jazyk Python, doplněný o specifické knihovny pro datovou vědu. V počátečních fázích analýzy byl kladen důraz na seznámení s datovým souborem a identifikaci vazeb mezi atributy a jejich vztahem k cílové proměnné. Následně byla data upravena a optimalizována pro modelování. V modelovací fázi byly vybrány dva přístupy: logistická regrese pro lineární klasifikaci a náhodné lesy pro nelineární klasifikaci. Ačkoli oba algoritmy byly vybrány s ohledem na specifika dat, logistická regrese se ukázala jako nevhodný prediktor. Naopak náhodný les přinesl uspokojivé výsledky. Přesto zde stále existuje prostor pro optimalizaci. V budoucích výzkumech by bylo vhodné zvážit další modely stromových algoritmů, zvláště vzhledem k výsledkům dosaženým s náhodným lesem.

Během zpracování diplomové práce bylo kromě hlavního cíle stanoveno několik dílčích cílů, jež napomohly lépe porozumět charakteristikám dat. Jedním z těchto

zjištění bylo, že odvětví „Oil & Energy“ má nejvyšší počet podvodných nabídek. V rámci pracovních pozic největší počet podvodných nabídek bylo v oblasti administrativy následované účetnictvím a auditem. Dalším dílčím cílem bylo identifikovat časté slovní fráze v inzerátech. Zajímavé bylo zjištění, že slova „project“ a „oil gas“ jsou typická pro podvodné nabídky, zatímco v pravdivých inzerátech dominují slova „information“ a „technology“. Taková zjištění ukazují možnosti pro další zkoumání. Hlubší analýza těchto souvislostí by mohla odhalit další vzory, jež by mohly významně pomoci při zdokonalení a rozšíření predikčního modelu.

Závěrem lze shrnout, že hlavní i dílčí cíle této diplomové práce byly naplněny. Práce nabízí komplexní pohled na problematiku dobývání znalostí z databází, kombinuje teoretický rámec s praktickým provedením a přináší výsledky zaměřené na predikční modelování. Ukázala význam a potenciál dolování dat v identifikaci podvodných nabídek. Výsledné modely jsou univerzální v kontextu daných atributů, to znamená možnost aplikace na jiné datové sady, pokud obsahují shodné atributy. Tato univerzálnost modelů otevírá mnoho možností pro budoucí výzkum, ať už v implementačních aspektech, nebo v rozšiřování a zdokonalování modelů.

Obor dolování dat je neustále se vyvíjející disciplínou. Průlomové technologie, jako jsou kvantové počítače nebo pokročilá umělá inteligence, mohou otevřít nové možnosti v analýze a interpretaci dat. V kontextu vzrůstající digitalizace společnosti bude role dobývání znalostí z databází stále významnější. Budoucí trendy, jako internet věcí či autonomní systémy, povedou k ještě většímu množství dostupných dat. V tomto prostředí bude zásadní nejen efektivní zpracování dat, ale i jejich hluboká interpretace. Výzkumníci v oboru tak budou čelit novým výzvám, ale zároveň mít příležitost k inovacím a pokroku v dobývaní znalostí z databází.

8 Seznam použité literatury

[1] ŞIMŞEK, Hazal. Top 50 Big Data Statistics in ’23: Market Size & Benefits [online]. AIMultiple, 2023 [cit. 2023-08-26]. Dostupné z: https://research.aimultiple.com/data-mining/.

[2] ROGALEWICZ, Michal a SIKA, Robert. Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering. Management and Production Engineering Review, 2016, roč. 7, č. 4.

[3] Developer Survey 2023 [online]. Stack Overflow, 2023 [cit. 2023-08-25]. Dostupné z: https://survey.stackoverflow.co/2023/#technology-most-popular-technologies.

[4] SKLENÁK, Vilém. Data, informace, znalosti a internet. Praha: C.H. Beck pro praxi, 2001. ISBN 80-7179-409-0.

[5] ALVARO, Felix. SQL: Easy SQL Programming & Database Management For Beginners, Your Step-By-Step Guide To Learning The SQL Database. Místo neznámé: CreateSpace Independent Publishing Platform, 2016. ISBN 978-1539916055.

[6] FAYYAD, U. M. a SMYTH, P. Advances in knowledge discovery and data mining. California: MIT Press, 1996. ISBN 0-262-56097-6.

[7] HAN, Jiawei a KAMBER, Micheline. Data mining: concepts and techniques. 2. vyd. San Francisco: Morgan Kaufmann Publishers, 2006. ISBN 978-1-55860-901-3.

[8] MARISCAL, Gonzalo, MARBÁN, Óscar a FERNÁNDEZ, Covadonga. A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 2010, roč. 25, č. 2. ISSN 0269-8889.

[9] NISBET, Robert, ELDER, John F. a MINER, Gary. Handbook of statistical analysis and data mining applications. Amsterdam: Elsevier, 2009. ISBN 978-0-12-374-765-5.

[10] CHAPMAN, Pete a kol. CRISP-DM 1.0: Step-by-step data mining guide [online]. Semanticscholar.org, 2000 [cit. 2023-01-12]. Dostupné z: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf.

[11] Getting Started with Data Science (Edition 2021) [online]. OpenSAP, 2021 [cit. 2023-01-21]. Dostupné z: https://open.sap.com/courses/ds3.

[12] HAND, David, MANNILA, Heikki a SMYTH, Padhraic. Principles of data mining. Massachusetts: MIT Press, 2001. ISBN 0-262-08290-x.
[13] SALUJA, Chhavi. Data Preparation — A crucial step in Data Mining [online]. Medium, 2018 [cit. 2023-03-05].
Dostupné z: https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281.
[14] LUNA, Zipporah. CRISP-DM Phase 3: Data Preparation [online]. Medium, 2021 [cit. 2023-03-14]. Dostupné z: https://medium.com/analytics-vidhya/crisp-dm-phase-3-data-preparation-faf5ee8dc38e.
[15] RUBIN, Roderick J., LITTLE, Donald B. a A. Statistical Analysis with Missing Data. New York: John Wiley & Sons, Inc., 2002. ISBN 0471183865.
[16] MATTHES, Jörg, DAVIS, Christine S. a POTTER, Robert F. The International Encyclopedia of Communication Research Methods. Místo neznámé: Wiley, 2017. ISBN 9781118901762.
[17] EEKHOUT, Iris. Missing Data – MCAR, MAR, NMAR mechanisms [online]. Don't Miss Out, [cit. 2023-03-14]. Dostupné z: https://www.missingdata.nl/missing-data/missing-data-mechanisms/.
[18] PECÁKOVÁ, Iva. Problém chybějících dat v dotazníkových šetřeních. Prague: Prague University of Economics and Business, 2014, roč. 2014, č. 6. ISSN 0572-3043.
[19] ENDERS, Amandou N. Baraldi a CRAIGEM K. An introduction to modern missing data analyses. Journal of School Psychology, 2010, roč. 48, č. 1. ISSN 00224405.
[20] DURRANT, Gabriele Beissel. Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. International Journal of Social Research Methodology, 2009, roč. 12, č. 4, s. 293-304.
[21] SCIENCE, ODSC – Open Data. Data Imputation: Beyond Mean, Median, and Mode [online]. Medium, 2020 [cit. 2023-03-21]. Dostupné z: https://odsc.medium.com/data-imputation-beyond-mean-median-and-mode-6c798f3212e3.
[22] SPARK, Cambridge. Tutorial: Introduction to Missing Data Imputation [online]. Medium, 2019 [cit. 2023-03-20]. Dostupné z:

https://medium.com/@Cambridge_Spark/tutorial-introduction-to-missing-data-imputation-4912b51c34eb.
[23] EEKHOUT, Iris. Single imputation methods. [online]. Don't miss out, [cit. 2023-03-21]. Dostupné z: https://www.missingdata.nl/missing-data/missing-data-methods/imputation-methods/.
[24] RIDZUAN, Fakhitah a WAN ZAINON, Wan Mohd Nazmee. A Review on Data Cleansing Methods for Big Data. Amsterdam: Elsevier, 2019, roč. 161. ISSN 1877-0509.
[25] AGARWAL, Malay. Pythonic Data Cleaning With pandas and NumPy. [online]. Real Python, [cit. 2023-04-12]. Dostupné z: https://realpython.com/python-data-cleaning-numpy-pandas/.
[26] RODRIGUEZ, Edgar Acuna a CAROLINE. A meta analysis study of outlier detection methods in classification [online]. ResearchGate, 2004 [cit. 2023-04-18]. Dostupné z: https://www.researchgate.net/publication/228728761_A_meta_analysis_study_of_outlier_detection_methods_in_classification.
[27] TUFFÉRY, Stéphane. Data Mining and Statistics for Decision Making. Hoboken, New Jersey, USA: Wiley, 2011. ISBN 978-0-470-97749-2.
[28] TRIPATHI, Himanshu. Different Type of Feature Engineering Encoding Techniques for Categorical Variable Encoding [online]. Medium, 2019 [cit. 2023-08-31]. Dostupné z: https://medium.com/analytics-vidhya/different-type-of-feature-engineering-encoding-techniques-for-categorical-variable-encoding-214363a016fb.
[29] BERKA, Petr. Dobývání znalostí z databází. Praha: Academia, 2003. ISBN 80-200-1062-9.
[30] PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011, vol. 12, pp. 2825-2830.
[31] ZHANG, Jia. Supervised vs. unsupervised learning [online]. IBM Corporation, 2021 [cit. 2023-02-22]. Dostupné z: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.
[32] BRAMER, Max. Principles of Data Mining. London: Springer, 2013. ISBN 978-1-4471-4883-3.

[33] BOX, George P. a DRAPER, Norman R. Empirical model-building and response surfaces. New York: Wiley, 1987. ISBN 0-471-81033-9.
[34] VIADINUGROHO, Raden Aurelius Andhika. Understanding Evaluation Metrics in Classification Modeling [online]. Medium, 2021 [cit. 2023-08-29]. Dostupné z: https://towardsdatascience.com/understanding-evaluation-metrics-in-classification-modeling-6cc197950f01.
[35] SAINI, Anshul. Conceptual Understanding of Logistic Regression for Data Science Beginners [online]. Analytics Vidhya, 2021 [cit. 2023-08-27]. Dostupné z: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/.
[36] AGRAWAL, R., IMIELINSKI, T. a SWAMI, A. Mining associations between sets of items in massive databases. In: ACM Press. Washington D.C., USA, 1993, s. 207-216. DOI: 10.1145/170035.170072.
[37] NERD, Normalized. Decision Tree Classification Clearly Explained! [video]. YouTube, 2021.
[38] HASTIE, Trevor, TIBSHIRANI, Robert a FRIEDMAN, J. H. The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2009. ISBN 0387848576.
[39] DONGES, Niklas. Random Forest: A Complete Guide for Machine Learning [online]. Builtin, 2023 [cit. 2023-08-28]. Dostupné z: https://builtin.com/data-science/random-forest-algorithm.
[40] INSTITUT BIOSTATISTIKY A ANALÝZ LÉKAŘSKÉ FAKULTY MASARYKOVY UNIVERZITY. Koncept umělé neuronové sítě [online]. Matematická biologie: e-learningová učebnice, [cit. 2023-06-20]. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--umela-inteligence--neuronove-site-jednotlivy-neuron--uvod-do-neuronovych-siti--koncept-umele-neuronove-site.
[41] HAMID, Kaiser. How To Use Scala for Data Science [online]. Knowledgehut, 2023 [cit. 2023-07-25]. Dostupné z: https://www.knowledgehut.com/blog/data-science/how-to-use-scala-for-data-science.

[42] SAYETH, Ahamed Lebbe. Comparative Analysis of Various Tools for Data Mining and Big Data Mining. International Journal of Engineering Research. 2018, roč. V7, č. 11.

[43] CAREY, Scott. What is SaaS? Software as a service defined [online]. Infoworld, 2022 [cit. 2023-06-21]. Dostupné z: https://www.infoworld.com/article/3226386/what-is-saas-software-as-a-service-defined.html.

[44] KUHLMAN, Dave. A Python Book: Beginning Python, Advanced Python, and Python Exercises. Platypus Global Media, 2011. ISBN 0984221239.

[45] EMERITUS. Here's Why Use Python for Data Science [online]. 2023 [cit. 2023-06-27]. Dostupné z: https://emeritus.org/in/learn/heres-why-use-python-for-data-science/.

[46] ENACHE, Maria Christina. Data Analysis with Pandas. 2019, roč. 25, č. 2. ISSN 15840409.

[47] NUMPY DEVELOPMENT TEAM. NumPy: the absolute basics for beginners [online]. NumPy, [cit. 2023-06-26]. Dostupné z: https://numpy.org/doc/stable/user/absolute_beginners.html.

[48] VACHIYATWALA, Rupal Snehkunj a KHUSHBOO. Data Analysis Using Pandas Library of Python. Acta Scientific Computer Sciences. 2022, roč. 4, č. 3.

[49] ANACONDA, Inc. All the Best Tools in One Platform [online]. Anaconda, Inc., [cit. 2023-06-26]. Dostupné z: https://www.anaconda.com/why-anaconda.

[50] SCIENCE, ODSC – Open Data. Why You Should be Using Jupyter Notebooks [online]. Medium, 2020 [cit. 2023-06-26]. Dostupné z: https://odsc.medium.com/why-you-should-be-using-jupyter-notebooks-ea2e568c59f2.

[51] KELLY, Jack. Fake Job Scams Are Becoming More Common—Here's How To Protect Yourself [online]. Forbes, 2023 [cit. 2023-08-14]. Dostupné z: https://www.forbes.com/sites/justinbirnbaum/2023/08/11/how-eddie-hearn-built-matchroom-sport-into-a-boxing-heavyweight/.

Přílohy

Obsah přiloženého CD

Součástí této práce je datový nosič s následujícím obsahem:

- src – adresář obsahující jupyter notebook se všemi zdrojovými kódy
- data – adresář obsahující všechna zdrojová data
- readme.txt – soubor s informacemi o jupyter notebooku

Příloha 2 Oskenované zadání práce

Univerzita Hradec Králové
Fakulta informatiky a managementu

Zadání diplomové práce

Autor: **Bc. Sharon Moscato**

Studium: I2000087

Studijní program: N0688A140001 Informační management

Studijní obor: Informační management

Název diplomové práce: **Dolování dat z vybrané datové sady pro predikci podvodných pracovních nabídek**

Název diplomové práce AJ: Data mining from a selected dataset for predicting fraudulent job offers.

Cíl, metody, literatura, předpoklady:

Cílem této diplomové práce je prozkoumat problematiku dobývání znalostí z databází a demonstrovat aplikaci predikčního modelu na reálných datech. Práce se zaměřuje na analýzu datového souboru obsahujícího pracovní nabídky, které jsou rozděleny do dvou kategorií: podvodné a pravdivé. Hlavním cílem je vytvořit predikční model, který bude schopen identifikovat podvodné nabídky práce. Mezi dílčí cíle této práce patří zodpovězení následujících analytických otázek:

- Jaký průmysl vykazuje největší počet podvodných pracovních nabídek?
- Která pracovní pozice je nejčastěji spojována s podvodnými nabídkami?
- Jaká slova se nejčastěji objevují v textech podvodných a nepodvodných pracovních nabídek?

Pro dosažení těchto cílů bude provedena podrobná analýza datového souboru pracovních nabídek s použitím metodiky CRISP-DM a metody pro dolování dat. Výsledkem práce bude model, který bude schopen klasifikovat pracovní nabídky jako pravdivé nebo podvodné s vysokou přesností. Diplomová práce poskytne náhled do problematiky dobývání znalostí z databází včetně jeho praktického využití na reálném datovém souboru.

1. SKLENÁK, Vilém. *Data, informace, znalosti a Internet.* Praha : C.H. Beck pro praxi, 2001. 80-7179-409-0.

2. FAYYAD, U. M. a P. SMYTH. *Advances in knowledge discovery and data mining.* California : MIT Press, 1996. 0-262-56097-6.

3. HAN, Jiawei a Micheline KAMBER. *Data mining: concepts and techniques. 2nd ed.* San Francisco : Morgan Kaufmann Publishers, 2006. 978-1-55860-901-3.

4. NISBET, Robert, John F. ELDER a Gary MINER. *Handbook of statistical analysis and data mining applications.* Amsterdam : Elsevier, 2009. 978-0-12-374-765-5.

5. CHAPMAN, Pete, et al. CRISP-DM 1.0: Step-by-step data mining guide. *semanticscholar.org.* [Online] 2000. [Citace: 12. 01 2023.] https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf.

Zadávající pracoviště: Katedra informačních technologií, Fakulta informatiky a managementu

Vedoucí práce: Ing. Tereza Otčenášková, BA, Ph.D.

Oponent: Ing. Patrik Urbaník

Datum zadání závěrečné práce: 21.1.2021