VYSOKÁ ŠKOLA POLYTECHNICKÁ JIHLAVA

Aplikovaná informatika

DATOVÁ VĚDA: KLASIFIKACE ŽADATELŮ O ÚVĚR

Bakalářská práce

Autor práce: Karolína Kočařová

Vedoucí práce: doc. Dr. Ing. Jan Voráček, CSc.

Jihlava 2026

Vysoká škola polytechnická Jihlava

Tolstého 16, 586 01 Jihlava

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Autor práce:	**Karolína Kočařová**
Studijní program:	Aplikovaná informatika
Garant studijního programu:	doc. Ing. Lenka Kuklišová Pavelková, Ph.D.

Název práce:	**Datová věda: klasifikace žadatelů o úvěr**
Vedoucí práce:	doc. Dr. Ing. Jan Voráček, CSc.
Cíl práce:	Cílem práce je aplikace úplného životního cyklu datové vědy na problematiku hodnocení bonity žadatelů o úvěr. Bude zjišťováno, u jakého typu klientů je vysoká pravděpodobnost úplného splacení dluhu v rámci dohodnutého kalendáře. Naplnění cíle zahrnuje identifikaci problému, rešerši souvisejících zdrojů, návrh a optimalizaci vhodných experimentů, čištění a konsolidaci relevantních dat, jejich explorativní analýzy a induktivní modelování, srozumitelnou vizualizaci výsledků a diskusi možných řešení definovaného problému na základě shromážděných, vyhodnocených a řádně diskutovaných podkladů. Základní datový korpus bude převzat z veřejných zdrojů a student ho může dle vlastních úvah dále modifikovat. Experimenty budou realizovány a distribuovány formou interaktivního online notebooku, využívajícího problémově orientované knihovny jazyka Python. Při vývoji finálního řešení bude využita agilní metodika Rapid Application Development s podporou umělé inteligence.

Abstrakt

Cílem bakalářské práce je aplikovat úplný životní cyklus datové vědy na úlohu hodnocení bonity žadatelů o úvěr. Práce se zaměřuje na rozlišení rizikových a nerizikových klientů na základě veřejně dostupných úvěrových dat, přičemž důraz je kladen nejen na predikční výkonnost modelů, ale také na metodickou korektnost postupu a interpretaci výsledků.

Součástí řešení je rešerše odborných východisek, definice cílové proměnné, příprava a čištění dat, explorativní a doplňková neřízená analýza, návrh experimentů, porovnání a ladění klasifikačních modelů a výběr rozhodovacího prahu. Praktická část byla realizována formou interaktivního notebooku v jazyce Python, který zachycuje reprodukovatelný analytický postup a umožňuje jak predikci pro jednotlivého žadatele, tak dávkové zpracování vstupního souboru. Vývoj řešení probíhal iterativně podle principů Rapid Application Development.

Z hlediska prahově nezávislých metrik dosáhl nejlepších výsledků naladěný model Gradient Boosting, zatímco jako finální provozní scénář byl s ohledem na záchyt rizikové třídy zvolen plný model Gradient Boosting s vážením tříd pomocí parametru class_weight, tedy varianta orientovaná na vysoký recall rizikové třídy.

Klíčová slova

datová věda; strojové učení; kreditní skóre; binární klasifikace; úvěrové riziko

Abstract

The aim of this bachelor's thesis is to apply the full data science lifecycle to the task of assessing the creditworthiness of loan applicants. The thesis focuses on distinguishing between risky and non-risky clients based on publicly available loan data, with emphasis placed not only on the predictive performance of the models but also on the methodological correctness of the procedure and the interpretation of the results.

The solution includes a review of the theoretical background, the definition of the target variable, data preparation and cleaning, exploratory and supplementary unsupervised analysis, experimental design, comparison and tuning of classification models, and the selection of a decision threshold. The practical part was implemented in the form of an interactive notebook in Python, which captures a reproducible analytical workflow and enables both prediction for an individual applicant and batch processing of an input file. The development of the solution proceeded iteratively according to the principles of Rapid Application Development.

In terms of threshold-independent metrics, the tuned Gradient Boosting model achieved the best results, whereas the full Gradient Boosting model trained with class weighting (class_weight) was selected as the final operational scenario because it prioritizes a high recall of the risky class.

Keywords

data science; machine learning; credit scoring; binary classification; credit risk

Prohlašuji, že předložená bakalářská práce je původní a zpracovala jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem v práci neporušila autorská práva (ve smyslu zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, v platném znění, dále též „**AZ**“).

Byla jsem seznámena s tím, že na mou bakalářskou práci se plně vztahuje **AZ**, zejména § 60 (školní dílo).

Podle § 47b zákona o vysokých školách souhlasím se zveřejněním své práce podle Směrnice pro vedení, vypracování a zveřejňování závěrečných prací na VŠPJ, a to bez ohledu na výsledek obhajoby.

Beru na vědomí, že VŠPJ má právo na uzavření licenční smlouvy o užití mé bakalářské práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé bakalářské práce (prodej, zapůjčení apod.).

Jsem si vědoma toho, že užít své bakalářské práce či poskytnout licenci k jejímu využití mohu jen se souhlasem VŠPJ, která má právo ode mě požadovat přiměřený příspěvek na úhradu nákladů, vynaložených vysokou školou na vytvoření díla (až do jejich skutečné výše), z výdělku dosaženého v souvislosti s užitím díla či poskytnutím licence.

V Jihlavě dne 15. dubna 2026

…….

Podpis studentky

Poděkování

Mé největší poděkování náleží vedoucímu mé bakalářské práce panu doc. Dr. Ing. Janu Voráčkovi, CSc., za cenné rady, vstřícnost, trpělivost a odborné vedení po celou dobu řešení projektu.

Dále děkuji rodině a příteli za neustálou motivaci, psychickou podporu a vytvoření klidného zázemí pro práci.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

AP	Average Precision
AUC	Area Under the Curve
CSV	Comma-Separated Values
CV	Cross-Validation
DTI	Debt-to-Income ratio
EAD	Exposure at Default
EDA	Exploratory Data Analysis
EL	Expected Loss
FAMD	Factor Analysis of Mixed Data
FICO	Fair Isaac Corporation
FN	false negative
FP	false positive
GB	Gradient Boosting
GDPR	General Data Protection Regulation
IV	Information Value
LGD	Loss Given Default
LR	logistic regression
ML	Machine Learning
PD	Probability of Default
PR	Precision–Recall
QQ	Quantile-Quantile graph
RAD	Rapid Application Development
RF	Random Forest
RFECV	Recursive Feature Elimination with Cross-Validation
ROC	Receiver Operating Characteristic
TN	True Negative
TP	True Positive
UMAP	Uniform Manifold Approximation and Projection
WoE	Weight of Evidence
XAI	Explainable Artificial Intelligence

1 Úvod

Hodnocení bonity žadatelů o úvěr představuje typickou rozhodovací úlohu, v níž je nutné současně zohlednit predikční výkon, interpretovatelnost i asymetrii dopadů chybných rozhodnutí. V praxi totiž nestačí model s vysokou souhrnnou přesností, podstatné je také to, jak se chová při konkrétním rozhodovacím prahu a zda dokáže zachytit skutečně rizikové případy.

Předkládaná práce proto chápe kreditní skóre jako binární klasifikační úlohu nad tabulkovými daty a řeší ji v celém životním cyklu datové vědy. Pozornost je věnována nejen samotným modelům, ale také kvalitě vstupních dat, prevenci úniku informací, nevyváženosti tříd, volbě vhodných vyhodnocovacích metrik a srozumitelné interpretaci výsledků.

Cílem práce je navrhnout a experimentálně ověřit reprodukovatelný pracovní postup, který na základě veřejně dostupných úvěrových dat rozliší rizikové a nerizikové žadatele, porovná vybrané klasifikační modely a umožní transparentně diskutovat kompromis mezi záchytem rizikové třídy a celkovou výkonností. Praktickým výstupem je interaktivní notebook v jazyce Python, který vedle analytické části obsahuje také demonstrační predikci pro jednotlivého žadatele a dávkové zpracování vstupního souboru.

Text práce je členěn do šesti hlavních částí. Po úvodním vymezení problému a teoretických východiscích následuje metodika popisující datový zdroj, přípravu dat, návrh experimentů a způsob vyhodnocení. Na ni navazuje kapitola výsledků, diskuse interpretující dosažená zjištění a závěr shrnující hlavní přínosy, míru naplnění cílů i směry dalšího rozvoje.

1.1 Charakteristika a kontext zkoumaného problému

Hodnocení bonity klientů představuje klíčovou součást řízení úvěrového rizika ve finančním sektoru. V praxi jde o rozhodování, které musí současně respektovat obchodní cíle instituce, regulatorní požadavky i potřebu rozumně omezovat finanční ztráty spojené se selháním klientů.

Současný rozvoj datově orientovaných přístupů a metod strojového učení umožňuje pracovat s rozsáhlejšími a různorodějšími klientskými daty než dříve. Zároveň však zvyšuje nároky na metodickou správnost celého postupu, zejména na kvalitu dat, interpretovatelnost modelu a korektní vyhodnocení výkonu při nevyváženém rozložení tříd.

Explicitně formulovaným problémem práce je navrhnout a experimentálně ověřit reprodukovatelný postup datové vědy, který na veřejně dostupných úvěrových datech rozliší rizikové a nerizikové žadatele, umožní transparentně volit kompromis mezi záchytem rizikové třídy a celkovou přesností a současně ukáže, které charakteristiky žadatelů mají na výsledek největší vliv. Nejde tedy pouze o hledání jednoho „nejlepšího modelu“, ale o metodicky obhajitelný návrh celého postupu od přípravy dat až po interpretaci a demonstrační využití výsledného řešení.

1.2 Cíl a výzkumné otázky práce

Hlavním cílem bakalářské práce je aplikovat životní cyklus datové vědy na problematiku hodnocení bonity žadatelů o úvěr a na tomto základě navrhnout prediktivní klasifikační řešení využitelné pro analýzu úvěrového rizika.

Pro dosažení hlavního záměru byly stanoveny tyto dílčí cíle:

1. analyzovat odborná východiska a přístupy používané při hodnocení bonity a klasifikaci úvěrového rizika,
2. připravit a metodicky zdůvodnit zpracování veřejně dostupného úvěrového datového souboru, včetně definice cílové proměnné a prevence úniku informací,
3. provést explorativní a doplňkovou neřízenou analýzu dat a identifikovat klíčové faktory související s rizikem selhání,
4. navrhnout, implementovat, porovnat a naladit vybrané klasifikační modely v prostředí Python, včetně volby metrik a rozhodovacího prahu,
5. vyhodnotit dosažené výsledky, interpretovat jejich praktický význam a vytvořit interaktivní notebookový prototyp umožňující demonstrační predikci.

Na základě stanovených cílů byly formulovány následující výzkumné otázky:

1. Které charakteristiky klientů nejvíce ovlivňují pravděpodobnost řádného splacení úvěru?
2. Jaké jsou přínosy využití metod datové vědy pro hodnocení bonity klientů?
3. Který z vybraných klasifikačních modelů poskytuje v daném kontextu nejlepší výsledky?
4. Jak lze výsledky modelu interpretovat a využít v praxi při rozhodování o úvěrových žádostech?
5. Jaký dopad mají redukce příznaků a vyvážení tříd na výkon a praktickou použitelnost modelu?

2 Teoretická východiska

Kapitola slouží k vymezení teoretického rámce, na němž stojí praktická část práce. Nejprve je stručně představena datová věda a životní cyklus projektu, který v této práci slouží jako metodická struktura celého řešení. Následně jsou vysvětleny základní pojmy úvěrového rizika, komponenty PD, LGD a EAD a principy kreditního skóre. Další část shrnuje používané techniky a nástroje datové vědy, včetně knihoven jazyka Python, a představuje vybrané klasifikační metody využité v praktické části práce, konkrétně logistickou regresi, Random Forest a Gradient Boosting. Závěr kapitoly se věnuje kvalitě dat, jejich předzpracování a přehledu dosavadních studií relevantních pro predikci selhání při splácení úvěru.

2.1 Datová věda a její charakteristika

Datová věda představuje analyticky orientovanou práci, která vyžaduje vysokou úroveň dovedností v oblasti softwarového inženýrství. Na rozdíl od tradiční statistiky se nezaměřuje pouze na výpočty či modelování, ale zahrnuje i přípravu, čištění a technické zpracování rozsáhlých a různorodých dat. Datový vědec tak spojuje dovednosti analytika, programátora i statistika a usiluje o získávání prakticky využitelných poznatků z dat *(Cady, 2024, s. 2)*.

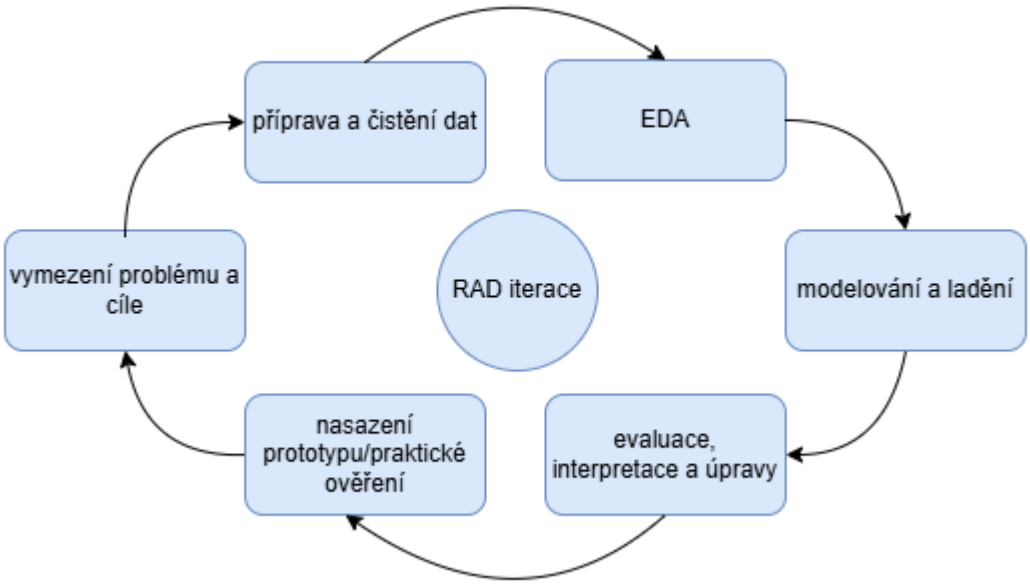
2.2 Životní cyklus datové vědy

Fáze řešení projektu vymezují, co je v jednotlivých etapách řešeno, zatímco metodika Rapid Application Development stanovuje, jakým způsobem je postup realizován. RAD je agilní metodický rámec, který upřednostňuje iterativní a inkrementální dodávání řešení před striktně sekvenčním „vodopádovým“ postupem *(Martin, 1991).*

V kontextu datové vědy umožňuje RAD rychlou tvorbu a testování prvních funkčních prototypů modelů. Pokud se při hodnocení ukáže, že model neodpovídá požadovanému cíli, je možné se vrátit k předchozím fázím, například k přípravě dat nebo k tvorbě odvozených příznaků, provést úpravy a cyklus opakovat. RAD postup zkracuje dobu mezi návrhem a ověřením a zároveň podporuje průběžné zpřesňování řešení.

Současné vývojové prostředí je navíc často spojeno s využitím nástrojů podporujících rychlé prototypování, automatizaci a práci s AI. Jejich využití však vyžaduje jasné vymezení lidských rolí a odpovědností, dokumentované procesy dohledu a průběžné vyhodnocování výstupů v kontextu konkrétního použití *(Tabassi, 2023).*

Iterativní logiku řešení v práci shrnuje obrázek 1. Zachycuje opakovaný cyklus vymezení problému, přípravy dat, explorativní analýzy, modelování, evaluace a následných úprav, který odpovídá způsobu práce využité v praktické části.



Obrázek 1: Iterativní cyklus metodiky RAD
zdroj: vlastní zpracování podle Martin (1991), vytvořeno v programu draw.io.

2.3 Teorie úvěrového rizika a hodnocení bonity

Úvěrové riziko je jádrem veškerého finančního rozhodování, neboť představuje nejistotu budoucího výsledku transakce. Lze jej jednoduše vysvětlit jako kritické rozhodnutí o tom, zda dlužník splní slib uhradit dluh (žádoucí výsledek) nebo zda selže. Rozhodnutí o půjčce proto vždy vyžaduje úsudek o nejistém budoucím výsledku, přičemž musí být podložen buď expertním odhadem, nebo formálním statistickým modelem. Úvěrové riziko představuje nebezpečí ztráty, které pro věřitele (banku nebo jinou finanční instituci) vyplývá z neschopnosti dlužníka dostát svým smluvním závazkům (tj. splatit úvěr nebo úrok včas a v plné výši). Jde o klíčový typ rizika v bankovnictví a je základním předpokladem pro cenotvorbu úvěrových produktů *(Brown a Moles, 2014).*

2.3.1 Komponenty úvěrového rizika

Řízení úvěrového rizika se tradičně opírá o tři základní parametry, které tvoří klíčové vstupy do výpočtu očekávané ztráty (Expected Loss, EL) *(Basel Committee on Banking Supervision, 2017):*

1. Pravděpodobnost selhání (Probability of Default, PD) představuje pravděpodobnost, že dlužník nesplní své závazky ve stanoveném časovém horizontu, obvykle do jednoho roku.
2. Ztráta při selhání (Loss Given Default, LGD) vyjadřuje očekávaný poměr ztráty z celkové expozice v případě selhání dlužníka, po zohlednění zajištění a nákladů na vymáhání.
3. Expozice při selhání (Exposure at Default, EAD) pak určuje očekávanou výši nesplacené částky, kterou má věřitel vystavenou v okamžiku selhání.

Uvedené komponenty jsou základními stavebními kameny interních modelů pro měření úvěrového rizika podle regulačních rámců Basel II a Basel III *(Basel Committee on Banking Supervision, 2017)*.

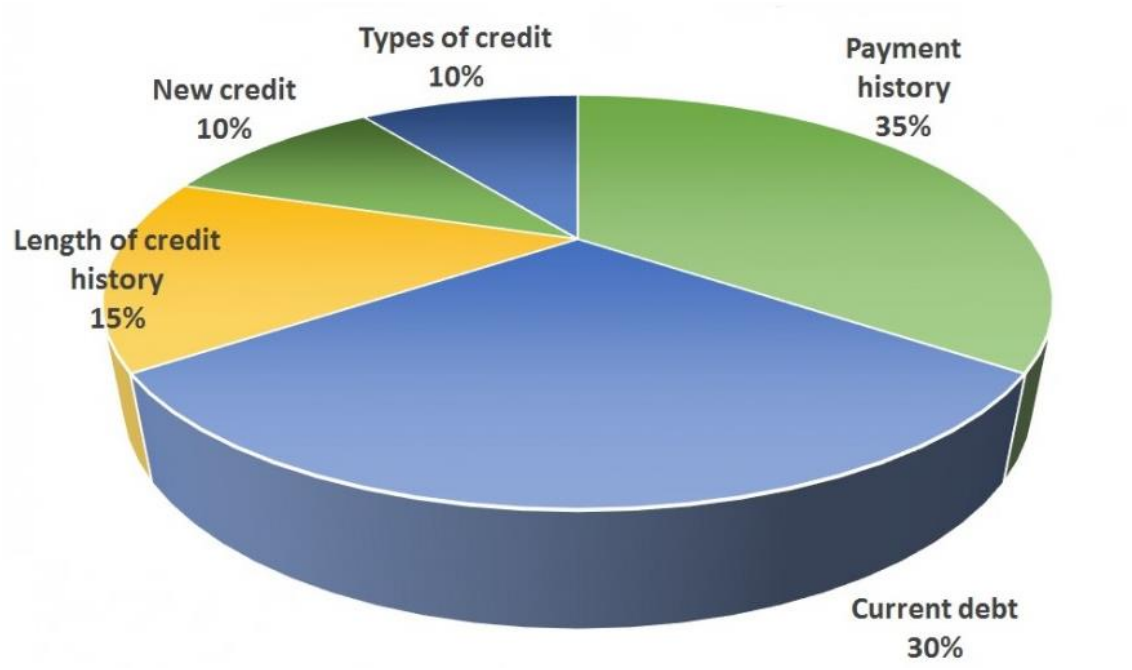
2.3.2 Principy kreditního skóre

Kreditní skóre je statistická technika používaná k automatickému hodnocení úvěruschopnosti žadatele. Systém přiřadí každému klientovi číselné skóre (např. od 1 do 1000), které odpovídá pravděpodobnosti, že klient svůj závazek splatí. Zatímco tradiční přístupy často pracují s kontinuálním skóre, v práci je úloha pojata jako binární klasifikace, tedy rozdělení klientů do dvou tříd podle očekávané schopnosti řádně splácet úvěr *(Thomas, 2009).*

Historicky se používají následující přístupy *(Thomas, 2009):*

1. Application Scoring – používá se pro posouzení nových žadatelů. Vychází z informací poskytnutých v žádosti o úvěr (příjem, zaměstnání, věk) a z dat o bonitě.
2. Behavioural Scoring – používá se pro posouzení stávajících klientů a jejich chování po schválení úvěru (např. pravidelnost splátek, využívání kreditních limitů).

Princip kreditového skórování je schematicky znázorněn na obrázku 2. Obrázek zachycuje obecný proces převodu vstupních charakteristik klienta, jako jsou demografické údaje, finanční situace a platební historie, na výsledné skóre prostřednictvím statistického modelu nebo modelu strojového učení. V praxi existují také komerční skóringové systémy, jako je například skóre FICO používané ve Spojených státech amerických, které pracují s předem definovanými vahami jednotlivých faktorů *(Fair Isaac Corporation (FICO), 2016)*. Tyto váhy však nejsou univerzální a nelze je považovat za obecně platné pro bankovní skóringové modely. V rámci práce je proto kreditní skóring chápán v obecném smyslu jako statistický model převádějící vstupní charakteristiky klienta na skóre, které je následně využito pro binární rozhodování o bonitě klienta.



Obrázek 2: Příklad komerčního skóringového systému FICO s váhovým zastoupením
zdroj: Fair Isaac Corporation (FICO), 2016

2.3.3 Úvěrové riziko jako klasifikační problém

Z pohledu datové analýzy lze hodnocení bonity klienta chápat jako úlohu binární klasifikace, jejímž cílem je rozdělit klienty do dvou skupin – bonitní a rizikoví podle očekávaného splácení závazku *(Brown a Mues, 2012).* Modely kreditního skórování se snaží na základě historických dat odhadnout pravděpodobnost selhání klienta, přičemž výsledné skóre je následně porovnáváno s rozhodovacím prahem, který určuje schválení či zamítnutí úvěru. Uvedený přístup umožňuje aplikaci metod strojového učení, které dokážou efektivně pracovat s větším množstvím proměnných a komplexními vztahy v datech *(Thomas, 2009).*

V dalším textu práce se opakovaně používají čtyři základní pracovní pojmy. Selhání klienta označuje situaci, kdy dlužník řádně nesplácí svůj závazek. Rozhodovací práh představuje hranici, od níž se modelový odhad pravděpodobnosti převádí na konečné zařazení do příslušné třídy. Nevyváženost tříd znamená, že rizikových případů je v datech podstatně méně než případů nerizikových. Podvzorkování většinové třídy (undersampling) a převzorkování menšinové třídy (oversampling) jsou techniky, které tuto nerovnováhu upravují pro potřeby trénování modelů *(He a Garcia, 2009).*

2.3.4 Náklady chybných rozhodnutí

Při hodnocení bonity klientů je důležité zohlednit také rozdílné dopady chybných rozhodnutí. Zatímco falešně pozitivní klasifikace (schválení rizikového klienta) může vést k finanční ztrátě, falešně negativní (zamítnutí bonitního klienta) znamená ztrátu obchodní příležitosti. Z tohoto důvodu nelze kvalitu modelů posuzovat pouze na základě celkové přesnosti, ale je nutné využívat více hodnotících metrik, které zohledňují charakter chyb klasifikace *(Thomas, 2009).* Právě uvedená asymetrie nákladů se v praxi neprojevuje pouze ve volbě hodnoticích metrik a rozhodovacího prahu, ale může vést také k použití nákladově citlivého učení nebo k vážení tříd během trénování modelu.

2.4 Techniky a nástroje používané v datové vědě

V práci byly techniky datové vědy chápány jako navazující pracovní postup směřující k řešení klasifikační úlohy v oblasti úvěrového rizika. Postup proto nebyl vystavěn jako obecný přehled izolovaných kroků, ale jako propojený celek zahrnující přípravu dat, definici cílové proměnné, explorativní analýzu, doplňkovou neřízenou analýzu, konstrukci modelů, jejich ladění, volbu rozhodovacího prahu a interpretaci výsledků. Z metodického hlediska byl důraz kladen zejména na reprodukovatelnost, práci s nevyváženými třídami a omezení rizika úniku informací mezi datovými podmnožinami.

Explorativní analýza dat slouží k počátečnímu porozumění datům, jejich struktuře a případným problémům v datech *(Komorowski, 2016).* Praktická realizace byla postavena na jazyce Python a na knihovnách odpovídajících charakteru tabulkových úvěrových dat. Pandas a NumPy byly využity pro čištění dat, transformační operace a tvorbu odvozených ukazatelů. Matplotlib a Seaborn sloužily k průzkumné analýze a prezentaci výsledků. Knihovna scikit-learn poskytla jednotný rámec pro předzpracování, pipeline, modelování, ladění i vyhodnocení, což bylo důležité pro zachování srovnatelného experimentálního režimu napříč modely.

Jednotlivé techniky nebyly vybírány samoúčelně, ale podle role v řešeném problému. V práci se kombinovala řízená a neřízená analýza. Řízená část pracovala s cílovou proměnnou a soustředila se zejména na imputaci chybějících hodnot, škálování numerických proměnných pro lineární modely, kódování kategoriálních atributů, výběr a redukci příznaků, porovnání více klasifikačních přístupů a interpretaci významu proměnných. Neřízená část naopak zkoumala vnitřní strukturu dat bez přímého využití cílové proměnné a sloužila k lepšímu popisu podobností a rozdílů mezi pozorováními *(Kuhn a Johnson, 2013).*

Explorativní část využila histogramy, boxploty a QQ grafy pro posouzení rozdělení proměnných, šikmosti a přítomnosti odlehlých hodnot. Uvedené nástroje sloužily k základní diagnostice dat a k posouzení, zda je vhodné některé proměnné dále transformačně upravit pro navazující analýzu *(Komorowski, 2016).* Vedle toho byla zařazena i doplňková neřízená analýza zaměřená na popis vnitřní struktury dat. Vzhledem ke smíšenému charakteru dat byla nejprve použita metoda FAMD, která umožňuje společné zohlednění numerických i kategoriálních proměnných v redukovaném prostoru *(Pagès, 2004)*. Pro dvourozměrnou vizualizaci podobností mezi pozorováními byla následně využita metoda UMAP *(McInnes a kol., 2018).* Identifikace skupin podobných klientů byla provedena metodou K-Prototypes, vhodnou pro shlukování dat obsahujících současně numerické a kategoriální atributy *(Huang, 1998).* Pro orientační volbu počtu shluků byly současně využity také elbow grafy.

Analýza byla realizována v notebookovém formátu v prostředí Kaggle/Jupyter. Zvolený formát odpovídal experimentálnímu charakteru práce, protože umožnil průběžně dokumentovat jednotlivé kroky, kontrolovat mezivýsledky a současně zachovat dobrou opakovatelnost postupu. Notebookové prostředí navíc usnadnilo propojení analytických částí s demonstrační predikcí pro jednoho žadatele i s dávkovou predikcí ze souboru CSV. Přehled použitých knihoven shrnuje tabulka 1.

Tabulka 1: Použité knihovny jazyka Python

Knihovna	Oblast použití	Účel v rámci práce
NumPy	Numerické výpočty	Práce s vícerozměrnými poli a numerickými operacemi, které tvoří základ datového zpracování.
Pandas	Zpracování dat	Načítání, čištění, transformace a analýza strukturovaných dat ve formě tabulek.
Matplotlib	Vizualizace dat	Tvorba základních grafů a vizualizací pro explorativní analýzu dat a prezentaci výsledků.
Seaborn	Vizualizace dat	Rozšířená vizualizace statistických vztahů mezi proměnnými s důrazem na přehlednost. Vizuálně statistické grafy jako boxploty, histogramy a heatmapy.
scikit-learn	Strojové učení	Implementace klasifikačních modelů, příprava dat, rozdělení datové množiny a vyhodnocení výkonu modelů pomocí metrik.
warnings	Řízení běhu programu	Potlačení nevyžádaných varování pro zvýšení přehlednosti výstupů.
time	Měření času	Měření doby výpočtu vybraných operací a modelů.

zdroj: vlastní zpracování

V úlohách kreditního skórování bývá důležitou součástí pracovního postupu také výběr příznaků, jehož cílem je omezit redundantní nebo málo přínosné proměnné a vytvořit kompaktnější modelovou reprezentaci. V praxi se kombinují filtrační přístupy založené například na míře statistické závislosti nebo informační hodnotě proměnných s postupy modelově řízenými, které posuzují přínos příznaků přímo vzhledem k výkonu zvoleného modelu. Vedle relevance samotných proměnných je současně vhodné sledovat i jejich vzájemnou nadbytečnost, například pomocí korelačních ukazatelů, aby výsledná sada příznaků nebyla zbytečně překryvná.

2.5 Přehled metod strojového učení pro klasifikaci

V práci nebyl přehled klasifikačních metod pojat jako obecný výčet, ale jako zdůvodnění výběru modelů zařazených do hlavního experimentu. Úvěrové hodnocení bylo v práci chápáno jako binární klasifikační úloha, u níž bylo třeba současně zohlednit predikční výkon, výpočetní náročnost, interpretovatelnost a praktickou použitelnost při práci s tabulkovými daty.

Z tohoto důvodu byly zvoleny tři modelové rodiny reprezentující odlišné vlastnosti. Logistická regrese představovala interpretovatelný lineární výchozí model, Random Forest robustní nelineární přístup založený na kombinaci více rozhodovacích stromů a Gradient Boosting model s předpokladem vysokého výkonu na tabulkových datech. Takto zvolená trojice umožnila porovnat, zda se složitější modely skutečně vyplatí oproti jednoduššímu a lépe vysvětlitelnému řešení.

Z hlediska metodologického zařazení se použité metody dále dělí na parametrické a neparametrické. Logistická regrese je považována za parametrickou metodu, protože pracuje s předem daným tvarem modelu a odhaduje omezený počet jeho parametrů. Naopak Random Forest a Gradient Boosting patří mezi neparametrické metody, protože nejsou založeny na jediném předem určeném lineárním vztahu mezi vstupy a výstupem a umožňují pružněji zachytit složitější strukturu dat, včetně nelinearit a vzájemných interakcí proměnných *(Hand a Henley, 1997; Breiman, 2001; Friedman, 2001).*

Výběr těchto metod současně vycházel z podoby navazujících experimentů v praktické části práce i z požadavků základní úlohy. Random Forest byl využit také jako základ pro některé kroky výběru příznaků, zatímco Gradient Boosting byl zařazen jako vhodný kandidát pro výsledné řešení základní úlohy nad tabulkovými daty. Přehled modelů proto neslouží pouze k obecnému srovnání, ale přímo souvisí s návrhem experimentální metodiky, porovnáním modelů a podobou výsledného řešení.

2.5.1 Logistická regrese

Logistická regrese byla do experimentu zařazena především jako srozumitelný výchozí model. V kontextu kreditového skórování má regresní přístup dlouhodobé opodstatnění, protože umožňuje relativně transparentně popsat vztah mezi vstupními proměnnými a odhadovanou pravděpodobností cílové třídy *(Hand a Henley, 1997).* Pro tuto práci byla důležitá zejména jako referenční model, vůči němuž bylo možné posoudit, zda složitější modely skutečně přinášejí prakticky významné zlepšení.

Současně šlo o vhodný nástroj pro ověření, nakolik jsou vztahy v datech zachytitelné lineárně a jak silně se v úloze projevují nelinearity a interakce. Pokud by logistická regrese dosahovala srovnatelných výsledků jako stromové modely, šlo by o argument ve prospěch jednoduššího a lépe interpretovatelného řešení. V opačném případě bylo možné lépe odůvodnit použití komplexnějších metod.

2.5.2 Random Forest

Random Forest byl zařazen jako robustní stromový model schopný zachytit nelineární vztahy a interakce mezi proměnnými bez nutnosti jejich explicitního zadání. To je u úvěrových dat výhodné zejména tehdy, když se vliv jednotlivých charakteristik klienta neprojevuje izolovaně, ale až v kombinaci s dalšími atributy. Agregace většího počtu stromů současně omezuje citlivost na náhodné výkyvy v trénovacích datech a snižuje riziko přeučení *(Breiman, 2001).*

V práci měl Random Forest dvojí roli. Jednak představoval samostatného kandidáta pro klasifikaci, jednak poskytoval užitečný základ pro posouzení významu proměnných a pro kroky redukce příznaků v navazujících experimentech. Ve srovnání s logistickou regresí nabízel vyšší flexibilitu, avšak za cenu nižší transparentnosti jednotlivých rozhodnutí.

2.5.3 Gradient Boosting

Gradient Boosting byl do experimentu zařazen jako model, u něhož bylo možné předpokládat vysoký predikční výkon na tabulkových datech. Jeho princip spočívá v postupném skládání dílčích modelů tak, aby každý další krok korigoval chyby předchozího řešení. Postup často umožňuje zachytit složitější vztahy v datech než jednodušší lineární modely i některé jiné stromové přístupy *(Friedman, 2001).*

Pro účely práce byl Gradient Boosting důležitý zejména jako kandidát na nejlepší celkovou diskriminační schopnost. Jeho zařazení umožnilo ověřit, zda zvolený datový soubor obsahuje vzory, které si žádají flexibilnější nelineární model, a zda případné zvýšení výkonu vyváží vyšší komplexitu výsledného řešení. Výsledky praktické části poté ukázaly, že právě tento model dosahoval z hlediska prahově nezávislých metrik nejlepších hodnot.

2.5.4 Ladění hyperparametrů modelů

Ladění hyperparametrů představuje hledání vhodného nastavení modelu pro danou úlohu. Hyperparametry jsou hodnoty, které se neurčují automaticky učením z dat, ale volí se předem. Ovlivňují přitom, jak bude model při trénování fungovat a jak složitý bude. Patří sem například síla regularizace u logistické regrese, počet a hloubka stromů u Random Forestu nebo learning rate u Gradient Boostingu.

K tomu se obvykle využívá validační množina nebo křížová validace. Mezi běžné postupy patří úplné prohledání předem definovaného prostoru hyperparametrů a náhodný výběr kombinací hyperparametrů *(Yang a Shami, 2020).*

2.6 Přehled dosavadních studií a přístupů

Tradiční přístupy k hodnocení bonity vycházely zejména ze statistických metod, mezi které patří například logistická regrese nebo diskriminační analýza. Tyto přístupy se v oblasti kreditního skórování prosadily především díky své relativní jednoduchosti, výpočetní nenáročnosti a dobré interpretovatelnosti výsledků *(Thomas, 2009).* Jejich praktická výhoda spočívá také v tom, že umožňují poměrně transparentně odhadovat pravděpodobnost selhání klienta na základě proměnných popisujících jeho finanční a demografické charakteristiky. Ve finančním prostředí, kde je často nutné modelová rozhodnutí nejen vytvářet, ale také srozumitelně vysvětlit, představovala tato transparentnost dlouhodobě významnou přednost. Nevýhodou těchto modelů však bývá omezená schopnost zachytit složitější nelineární vztahy a interakce mezi proměnnými, což může snižovat jejich predikční výkonnost u rozsáhlejších a komplexnějších datových souborů.

S rozvojem výpočetních možností a rostoucí dostupností dat se proto do popředí dostaly také metody strojového učení, které umožňují lépe modelovat nelineární vazby mezi proměnnými a v řadě případů dosahují vyšší predikční přesnosti než tradiční statistické metody *(Brown a Mues, 2012)*. V kreditním skórování se tak pozornost postupně přesunula od jednodušších lineárních modelů k širšímu spektru klasifikačních metod, včetně stromových algoritmů, souborových metod a dalších pokročilejších přístupů. Tento posun však neznamená, že by tradiční modely ztratily význam, spíše ukazuje, že volba vhodné metody závisí na charakteru dat, cíli analýzy a požadavcích na interpretovatelnost. Novější systematická rešerše současně ukazuje, že současný výzkum kreditního skóringu se posouvá od prostého porovnávání algoritmů k širším otázkám robustnosti, interpretovatelnosti a validační kvality celého modelového postupu *(Ayari a kol., 2026).*

Konkrétnější pohled nabízí studie Brown a Mues *(2012),* která porovnávala několik klasifikačních metod na pěti reálných datových souborech kreditního skórování se zvláštním důrazem na problém nevyvážených tříd. Autoři se soustředili především na modelovací a vyhodnocovací fázi analytického procesu, tedy na porovnání alternativních klasifikátorů nad již připravenými daty. Pracovali s logistickou regresí, neuronovými sítěmi, rozhodovacími stromy, Gradient Boostingem, least squares, support vector machines a Random Forestem a výkon posuzovali pomocí AUC. Výsledky ukázaly, že Random Forest a Gradient Boosting patřily k nejúspěšnějším metodám a současně si relativně dobře zachovávaly výkonnost i při rostoucí třídní nerovnováze. Studie je důležitá i tím, že upozorňuje na skutečnost, že v kreditním skórování nelze modely posuzovat pouze podle celkové přesnosti, ale je nutné zohlednit i strukturu tříd a praktické dopady chyb klasifikace. Podobný trend potvrzuje i novější empirická studie, podle níž metody strojového učení v úlohách kreditního skóringu často překonávají tradičnější statistické přístupy. *(Mestiri, 2024).*

Ještě širší benchmark publikovali Lessmann a kol. *(2015),* kteří porovnali 41 klasifikátorů na osmi reálných datových souborech kreditního skórování a napříč šesti výkonnostními metrikami. Jejich přístup tak nepokrýval pouze samotné natrénování modelů, ale i systematické srovnání více typů ukazatelů predikční úspěšnosti a statistické testování rozdílů mezi modely. Z hlediska datového cyklu se tedy zaměřovali hlavně na fázi modelování, validace a srovnávací evaluace, nikoli na celý proces od čištění dat až po demonstrační nasazení. Studie ukázala, že několik pokročilejších metod dosahuje statisticky lepších výsledků než tradiční logistická regrese,

přičemž zvláště silně vyšly heterogenní ensemble přístupy (metody založené na kombinaci více různých modelů). Pro oblast kreditního skórování je uvedená práce významná tím, že poskytuje velmi široké empirické srovnání a podporuje závěr, že modernější metody mohou přinést měřitelné zlepšení, pokud jsou korektně validovány.

Novější komparativní studie navíc ukazuje, že boostingové modely typu XGBoost mohou vedle vyšší predikční výkonnosti nabídnout i lepší stabilitu při stresových scénářích *(Vakrani a kol., 2026).*

Přehledový článek Louzada a kol. *(2016)* naopak zachycuje širší vývoj oboru. Autoři provedli systematickou rešerši klasifikačních metod používaných v kreditním skórování a ukázali postupný posun od tradičních statistických metod k širšímu využití metod strojového učení a ensemble přístupů. Uvedená práce nepředstavuje jeden konkrétní experiment, ale spíše shrnutí toho, jaké modely jsou v literatuře používány, jaké typy dat se při kreditním skórování objevují a jaká hodnoticí kritéria bývají v jednotlivých studiích uplatňována. Současně autoři zdůrazňují, že v kreditním skórování nezáleží pouze na samotné predikční výkonnosti, ale také na interpretovatelnosti modelu, dostupnosti vstupních dat a vhodnosti validačního postupu. Přehledová studie tak podporuje pohled, že úspěšné řešení nelze redukovat pouze na volbu jednoho algoritmu, ale že důležitá je i metodická kvalita celého analytického postupu.

V současném výzkumu i praxi tedy není důležitá pouze samotná predikční výkonnost modelu. Ve finančním prostředí hraje významnou roli také interpretovatelnost, transparentnost a možnost obhájit modelová rozhodnutí v regulatorním rámci. Právě u komplexnějších modelů strojového učení může být jejich interpretace obtížnější, což představuje jednu z překážek jejich širšího nasazení v regulovaných oblastech *(Basel Committee on Banking Supervision, 2017).* Z tohoto důvodu se v odborné literatuře stále častěji zdůrazňuje potřeba hledat rovnováhu mezi výkonností modelu a jeho srozumitelností. Jinými slovy, model nemusí být v praxi nejvhodnější pouze proto, že dosáhne nejlepší hodnoty jedné vybrané metriky, důležitá je i jeho robustnost, vysvětlitelnost a soulad s požadavky konkrétního rozhodovacího prostředí.

Pro praktickou část práce byl zvolen veřejně dostupný datový soubor Credit Risk Analysis z platformy Kaggle, jehož struktura odpovídá typickým atributům úvěrových záznamů *(ranadeep, nedatováno).* Vedle odborné literatury byly jako dílčí inspirační zdroje využity také veřejně dostupné notebooky z platformy Kaggle, a to zejména při návrhu některých vizualizací, pomocných screeningových postupů a vybraných technických detailů zpracování dat. Tyto zdroje však zpravidla pokrývají spíše dílčí části analytického procesu, například explorativní analýzu, tvorbu vybraných grafů nebo technickou implementaci modelu, a nenahrazují proto odborné studie ani metodicky ucelený výzkumný postup. Online zdroje v práci neslouží jako náhrada odborné literatury, ale pouze jako doplňkový praktický podklad. Finální analytický postup, volba experimentů, interpretace výsledků i odpovědnost za správnost řešení zůstávají na autorce.

Z dosavadních studií i praktických aplikací vyplývá, že při hodnocení úvěrového rizika je důležitá nejen predikční výkonnost modelu, ale také interpretovatelnost, transparentnost, kvalita vstupních dat a korektní validace. Na tento přístup navazuje i předkládaná práce, která se zaměřuje na porovnání vybraných klasifikačních metod, konkrétně logistické regrese, Random Forestu a Gradient Boostingu, při predikci bonity žadatelů o úvěr na základě veřejně dostupných dat. Oproti řadě citovaných studií se přitom nesoustředí pouze na samotné srovnání modelů, ale snaží se pokrýt širší část datového cyklu, tedy i přípravu dat, prevenci úniku

informací, doplňkovou explorativní a neřízenou analýzu, volbu rozhodovacího prahu a demonstrační podobu výsledného řešení.

2.7 Kvalita dat a jejich předzpracování

Kvalita dat v úlohách úvěrového hodnocení nepředstavuje pouze otázku technické správnosti datového souboru, ale také otázku metodologické správnosti celého analytického postupu *(Pyle, 1999).* I relativně malé chyby v přípravě dat mohou vést ke zkreslenému obrazu o skutečné použitelnosti modelu, a tím i k nadhodnocení jeho výkonnosti *(Kuhn a Johnson, 2013).* Z tohoto důvodu není důležité pouze ošetření chybějících hodnot nebo odlehlých pozorování, ale také posouzení toho, zda vstupní proměnné odpovídají informacím, které by byly reálně dostupné v okamžiku úvěrového rozhodování *(Thomas, 2009).*

V oblasti úvěrového hodnocení je nutné zohlednit také ochranu osobních údajů podle GDPR, protože reálná klientská data mohou obsahovat informace umožňující přímou nebo nepřímou identifikaci fyzických osob. Z tohoto důvodu nelze s takovými daty pro experimentální, demonstrační nebo veřejně sdílené analytické účely pracovat bez jejich předchozí anonymizace nebo bez přijetí odpovídajícího právního základu a souvisejících ochranných opatření. V akademickém prostředí je proto vhodné využívat zejména anonymizovaná data, případně veřejně dostupná, aby nedocházelo k zásahu do soukromí konkrétních klientů. Omezení současně ovlivňují výběr datasetu i podobu navrženého analytického postupu.

2.7.1 Význam kvality dat v datové vědě

Za zvlášť významné riziko je v této souvislosti považován datový únik. Ten nastává tehdy, pokud model při učení využívá informace, které by v reálné rozhodovací situaci nebyly k dispozici nebo které přímo či nepřímo odrážejí budoucí vývoj sledovaného případu *(Kuhn a Johnson, 2013)*. V oblasti kreditového skórování mohou být problematické zejména proměnné vznikající až po poskytnutí úvěru nebo atributy zachycující následný platební vývoj, protože jejich využití může vést k nadhodnocení predikční schopnosti modelu *(Thomas, 2009).*

Důležitou součástí kvality dat je rovněž správné nastavení transformačních kroků v procesu předzpracování. Imputace, škálování i kódování proměnných mají vycházet pouze z trénovacích dat a na ostatní datové podmnožiny mají být následně jen aplikovány, aby bylo zachováno korektní hodnocení modelu a omezen datový únik *(Kuhn a Johnson, 2013).* Kvalita dat se tak neprojevuje pouze na úrovni vstupních hodnot, ale i na úrovni celého postupu zpracování *(Pyle, 1999).*

2.7.2 Chybějící hodnoty v datech

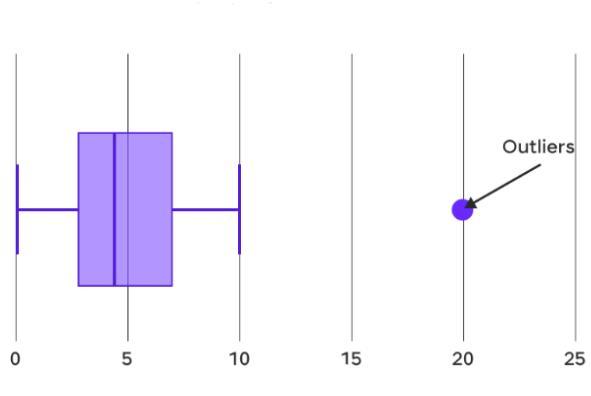
Chybějící hodnoty představují v reálných datech běžný problém, který může ovlivnit jak výslednou podobu datového souboru, tak stabilitu a interpretovatelnost modelu. Jejich ošetření proto nepředstavuje pouze technický krok předzpracování, ale metodologicky významnou součást analytického postupu. Volba způsobu imputace by měla odpovídat charakteru proměnných, rozsahu chybějících údajů i povaze následné analýzy, protože různé přístupy mohou vést k odlišným výsledkům modelování.

2.7.3 Odlehlé hodnoty v datech

Odlehlé hodnoty představují pozorování, která se výrazně odlišují od převládajícího rozdělení dat. V datové analýze však nejsou automaticky chápány jako chybné záznamy, protože mohou vznikat nejen v důsledku měřicí nebo evidenční chyby, ale také jako důsledek skutečné variability sledovaného jevu. Z tohoto důvodu nelze k jejich odstraňování přistupovat mechanicky a jejich význam je nutné posuzovat s ohledem na věcný kontext dat i účel následného modelování.

V oblasti úvěrového hodnocení mohou extrémní hodnoty představovat nejen anomálie, ale i reálně existující, méně časté a potenciálně rizikovější profily klientů. Jejich automatické vyloučení by proto mohlo vést ke ztrátě informace, která je pro odhad úvěrového rizika relevantní. Při práci s takovými proměnnými je proto vhodné zvažovat nejen statistickou neobvyklost hodnot, ale také jejich ekonomický a rozhodovací význam v kontextu úvěrového hodnocení *(Thomas, 2009)*.

Posouzení odlehlých hodnot se obvykle opírá o kombinaci grafických, statistických a transformačních postupů. Jedním z často používaných přístupů je pravidlo založené na interkvartilovém rozpětí (IQR, interquartile range), které je definováno jako rozdíl mezi třetím a prvním kvartilem. Za odlehlá bývají obvykle považována pozorování ležící pod hranicí $Q1 - 1{,}5 \times IQR$ nebo nad hranicí $Q3 + 1{,}5 \times IQR$. Uvedené pravidlo však nelze chápat jako univerzální mechanické kritérium, zejména u silně asymetrických proměnných, kde mohou krajní hodnoty představovat přirozenou součást rozdělení. Boxploty slouží k orientační identifikaci extrémních pozorování v rozdělení dat, která jsou demonstrována na obrázku 3.



Obrázek 3: Identifikace odlehlých hodnot v datech pomocí boxplotu
zdroj: vlastní zpracování

Naopak QQ grafy pomáhají rozlišit, zda je neobvyklé chování proměnné spojeno spíše s odlehlými hodnotami, šikmostí rozdělení nebo odchylkou od předpokládaného tvaru

distribuce. V případech, kdy jsou extrémy spojeny se silnou asymetrií rozdělení, může být vhodnější využít transformační úpravy, například logaritmickou transformaci, než automatické odstranění těchto pozorování *(Kuhn a Johnson, 2013).*

Význam odlehlých hodnot se může lišit podle typu použitého modelu. Lineárnější a obecně citlivější modely bývají na extrémní hodnoty náchylnější, zatímco stromové algoritmy bývají vůči takovým pozorováním zpravidla robustnější. Z tohoto důvodu je vhodné o způsobu zacházení s odlehlými hodnotami rozhodovat s ohledem na charakter konkrétní proměnné i na zamýšlený způsob modelování.

2.7.4 Normalizace a standardizace dat

Škálování numerických proměnných představuje běžnou součást předzpracování dat, avšak význam se liší podle typu použitého modelu. U metod založených na vzdálenostech, optimalizaci parametrů nebo lineární kombinaci vstupních proměnných může být standardizace důležitá, protože omezuje vliv rozdílných rozsahů jednotlivých atributů na proces učení i na výslednou interpretaci modelu. Naopak u stromových metod nebývá škálování zpravidla nezbytné, protože jejich rozhodovací mechanismus není založen na absolutní velikosti vstupních hodnot stejným způsobem jako u lineárních modelů *(James a kol., 2021).*

Z hlediska metodiky proto nelze škálování chápat jako univerzálně nutný krok, ale spíše jako úpravu, jejíž vhodnost závisí na charakteru dat a typu modelu. Smyslem předzpracování není aplikovat všechny dostupné transformace plošně, ale volit takové postupy, které odpovídají analytickému cíli a podporují korektní a srovnatelné modelování.

2.7.5 Kódování kategorických proměnných

Kategorické proměnné je před použitím ve většině klasifikačních algoritmů nutné převést do numerické podoby. U lineárních modelů se často využívá one-hot kódování, které převádí jednotlivé kategorie na samostatné binární indikátory a nevytváří mezi nimi umělé pořadí. Přístup je vhodný zejména tehdy, pokud kategorie nemají přirozenou ordinalitu *(scikit-learn developers, nedatováno b)*.

U některých dalších přístupů může být vhodné také ordinální kódování, zejména pokud je cílem omezit nárůst dimenzionality dat nebo pracovat s proměnnými, které mají přirozené pořadí. Způsob kódování však může ovlivnit chování modelu i interpretaci výsledků, a proto by neměl být chápán pouze jako technický detail, ale jako metodicky významná součást předzpracování dat *(Kuhn a Johnson, 2013).*

Zvláštní pozornost je třeba věnovat proměnným s vysokým počtem unikátních hodnot, zejména textovým polím nebo proměnným obsahujícím volný text. U těchto proměnných nemusí být běžné kódovací postupy vhodné, protože mohou vést k výraznému nárůstu dimenze a výpočetní náročnosti. Jejich zpracování proto často vyžaduje specifický postup, případně jejich vyřazení z hlavního modelovacího rámce *(scikit-learn developers, nedatováno b).*

2.7.6 Rozdělení dat na trénovací, validační a testovací sadu

Rozdělení dat na trénovací, validační a testovací část představuje základní předpoklad korektního vyhodnocení predikčních modelů. Trénovací sada slouží k učení modelu, validační sada k ladění hyperparametrů a výběru vhodného nastavení a testovací sada k závěrečnému nezávislému posouzení generalizační schopnosti. Takové uspořádání umožňuje oddělit jednotlivé fáze modelování a snižuje riziko, že bude výkonnost modelu nadhodnocena v důsledku přizpůsobení se konkrétním datům.

V klasifikačních úlohách s nevyváženým zastoupením tříd bývá vhodné doplnit rozdělení dat také stratifikací podle cílové proměnné, aby byl ve všech podmnožinách zachován podobný poměr jednotlivých tříd. Tím se omezuje náhodné zkreslení výsledků a zvyšuje srovnatelnost modelů při jejich hodnocení.

Vedle samostatné validační sady se v praxi často využívá také křížová validace, která poskytuje stabilnější odhad výkonnosti než jediné rozdělení dat. Kombinace samostatné testovací sady a křížové validace tak představuje vhodný postup pro situace, kdy je cílem současně porovnávat více modelů v kontrolovaném experimentálním režimu a zároveň zachovat nezávislé závěrečné ověření jejich generalizační schopnosti *(Kuhn a Johnson, 2013).*

2.8 Vyhodnocení klasifikačních modelů úvěrového rizika

Při hodnocení klasifikačních modelů se často sleduje více metrik současně, protože každá zachycuje jiný aspekt jejich chování. Recall vyjadřuje podíl správně zachycených pozitivních případů ze všech skutečně pozitivních případů, jak je uvedeno na rovnici:

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

F1 skóre představuje souhrnnou metriku založenou na precision a recall a vypočítá se jako:

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

$$F1 = 2 \cdot \frac{Precision\ \cdot Recall}{Precision + Recall}$$

ROC-AUC vyjadřuje plochu pod ROC křivkou, která porovnává citlivost modelu s mírou falešně pozitivních případů při různých rozhodovacích prazích *(Fawcett, 2006).* Average Precision shrnuje průběh Precision–Recall křivky a v diskrétní podobě ji lze zapsat jako:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

Tyto metriky se využívají zejména tehdy, když nestačí hodnotit model pouze podle jediné souhrnné charakteristiky a kdy je důležité posoudit i kvalitu zachycení sledované třídy *(Powers, 2011), (Saito a Rehmsmeier, 2015).* Často se využívá také balanced accuracy, tedy metrika založená na průměrném recallu obou tříd. Její výhodou je, že omezuje zkreslení způsobené nerovnoměrným zastoupením tříd *(Brodersen a kol., 2010).*

V nevyvážených klasifikačních úlohách však nemusí být jediným řešením pouze změna rozhodovacího prahu nebo fyzická úprava třídního rozložení pomocí podvzorkování či převzorkování. Alternativou je také vážení tříd během učení modelu, při němž jsou chyby

u méně zastoupené třídy penalizovány silněji než chyby u třídy většinové. Zvolený přístup nemění samotná vstupní data, ale ovlivňuje optimalizační proces modelu a může být vhodný zejména tehdy, když je cílem zvýšit citlivost vůči rizikové třídě bez přímého zásahu do trénovací množiny.

Učící křivka slouží k posouzení, jak se mění výkon modelu na trénovacích a validačních datech v závislosti na velikosti trénovací množiny. Umožňuje tak lépe odhadnout, zda model vykazuje známky přeučení, nedostatečného naučení, nebo zda by jeho výkon mohl být zlepšen větším množstvím dat. V kontextu klasifikačních úloh představuje užitečný doplňkový nástroj pro interpretaci dosažených výsledků, protože pomáhá rozlišit, zda je omezení výkonnosti dáno spíše vlastnostmi modelu, nebo rozsahem dostupných trénovacích dat *(scikit-learn developers, nedatováno h).*

Vedle souhrnných metrik a učících křivek lze pro interpretaci výsledného modelu využít také metody odhadu významu proměnných. Jedním z praktických přístupů je permutation importance, která sleduje, jak se změní výkon modelu po náhodném promíchání hodnot vybrané proměnné. Pokud po narušení daného příznaku výkon modelu výrazně klesne, lze usuzovat, že proměnná přispívá k predikci významněji než příznaky, jejichž promíchání má na výkon jen malý dopad. Výhodou této metody je, že ji lze použít i pro složitější modely, u nichž není interpretace vnitřní struktury rozhodování snadná. *(Breiman, 2001)*

Při interpretaci výsledků je současně vhodné sledovat nejen bodové hodnoty metrik z jednoho konkrétního rozdělení dat, ale také jejich variabilitu mezi opakovanými běhy nebo validačními děleními. Intervalové nebo empirické rozpětí metrik pomáhá posoudit, zda je dosažený výkon stabilní, nebo zda výrazně kolísá v závislosti na konkrétním výběru dat *(Kuhn a Johnson, 2013)*. Jedním z vhodných způsobů prezentace takového srovnání je forest plot, který umožňuje přehledně zobrazit centrální odhad metriky i šíři její variability. Takový pohled je užitečný zejména tehdy, když se porovnávají modely s relativně podobnými výsledky a je třeba lépe odlišit skutečný rozdíl výkonu od běžné variability odhadu.

3 Metodika

3.1 Charakteristika a zdroj použitých dat

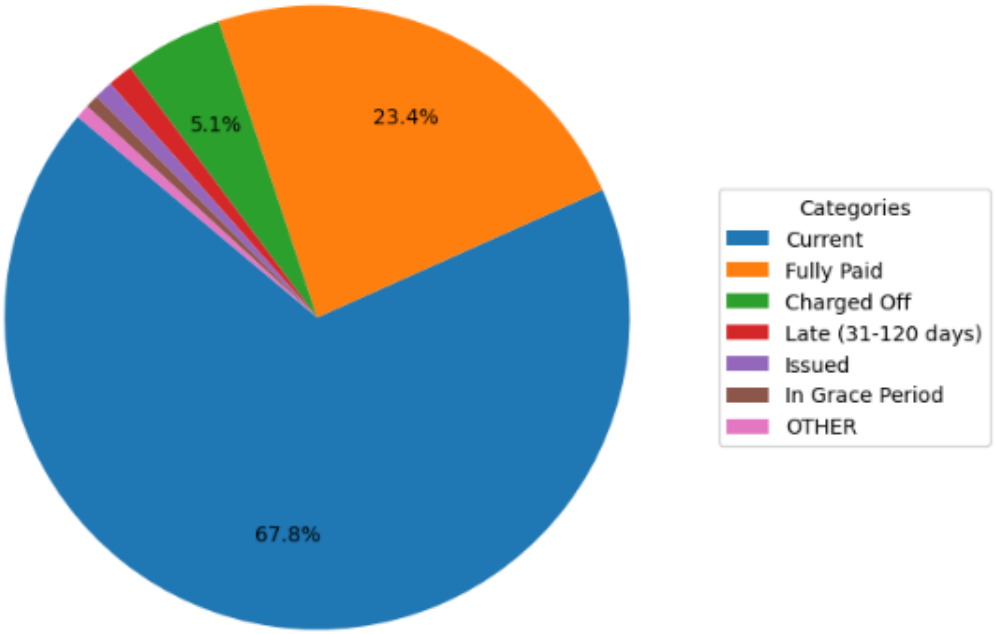
Pro praktickou část práce byl použit veřejně dostupný datový soubor Credit Risk Analysis z platformy Kaggle *(ranadeep, nedatováno),* konkrétně soubor loan.csv, který svým charakterem odpovídá veřejně dostupným úvěrovým datům používaným pro úlohy predikce úvěrového rizika. Datový soubor obsahuje tabulková data popisující úvěry a charakteristiky klientů. Zastoupeny jsou jak numerické proměnné, například výše úvěru, úroková sazba, příjem nebo ukazatele zadlužení, tak kategoriální proměnné, například typ bydlení, účel úvěru, stav ověření nebo délka zaměstnání.

Použitý datový soubor je v dostupném popisu prezentován jako soubor dat určený pro úlohy hodnocení úvěrového rizika. Pro tuto práci byl zvolen proto, že svým charakterem odpovídá problému hodnocení bonity a zároveň umožňuje demonstrovat kompletní pracovní postup datové vědy na otevřených datech. Současně je však nutné upozornit, že jeho úplný původ, způsob sběru i míru autenticity nelze nezávisle ověřit, což představuje jedno z omezení práce. Veřejně dostupná data navíc nemusí plně odpovídat reálnému procesu schvalování úvěrů v konkrétní finanční instituci. Výsledky je proto vhodné chápat především jako metodické ověření zvoleného postupu.

3.1.1 Definice cílové proměnné

Cílová proměnná byla odvozena ze sloupce loan_status pomocí explicitního přiřazení původních stavů úvěru do tříd **GOOD**, **BAD** a **IGNORE**. Do nerizikové třídy 1 (good) byly zařazeny stavy *Fully Paid* a *Does not meet the credit policy, Status: Fully Paid*. Do rizikové třídy 0 (bad) byly zařazeny stavy *Charged Off, Default a Does not meet the credit policy. Status: Charged Off*. Stavy *Current*, *Issued*, In *Grace Period, Late (16–30 days)* a *Late (31–120 days*) byly z dalšího modelování vyloučeny, protože nepředstavují jednoznačně uzavřený výsledek úvěru.

Takto definovaná cílová proměnná umožňuje převést původně vícestavový atribut na přehlednou binární klasifikační úlohu, v níž je cílem odlišit rizikové a nerizikové klienty. Převod původních kategorií *loan_status* do výsledných tříd shrnuje obrázek 4.



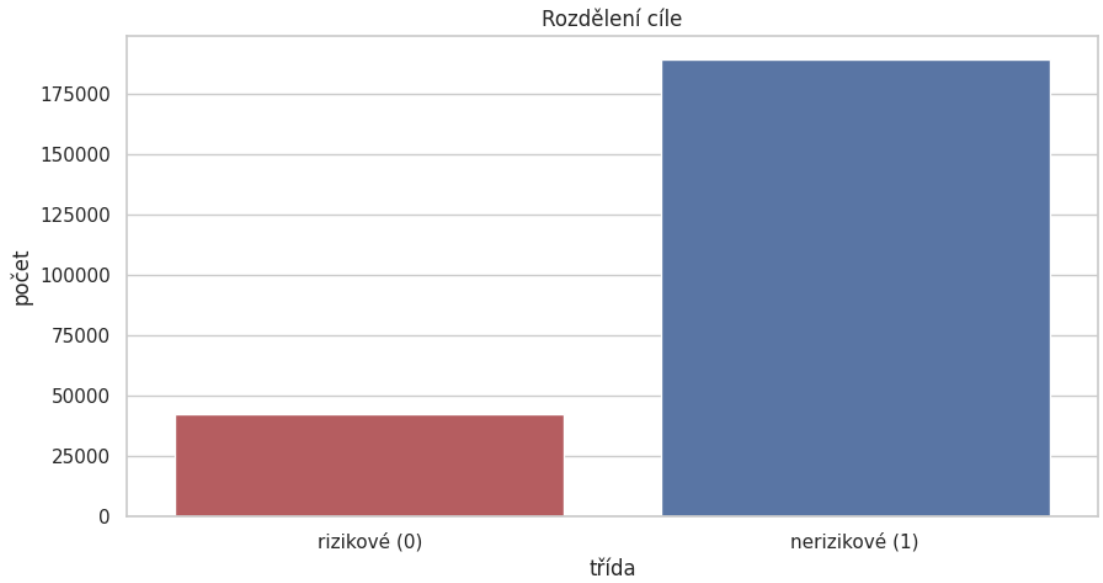
Obrázek 4: Struktura kategorií loan_status
zdroj: vlastní zpracování

Zjednodušená ukázka logiky proměnné *loan_status* v Pythonu zachycuje obrázek 5.

```
TARGET_RAW = "loan_status"

DEFAULT_GOOD = {
    "Fully Paid",
    "Does not meet the credit policy. Status:Fully Paid"
}

DEFAULT_BAD = {
    "Charged Off",
    "Default",
    "Does not meet the credit policy. Status:Charged Off"
}

TARGET_MAP = {s: 1 for s in DEFAULT_GOOD} | {s: 0 for s in DEFAULT_BAD}

y = df[TARGET_RAW].map(TARGET_MAP)
df2 = df.loc[y.notna()].copy()
df2["target"] = y.loc[y.notna()].astype(int)
```

Obrázek 5: Definice cílové proměnné v Pythonu
zdroj: vlastní zpracování

Základní rozdělení výsledné cílové proměnné po vyčištění dat je uvedeno na obrázku 6. Je patrné, že datový soubor je výrazně nevyvážený, protože nerizikových případů je podstatně více než případů rizikových.



Obrázek 6: Rozdělení cíle

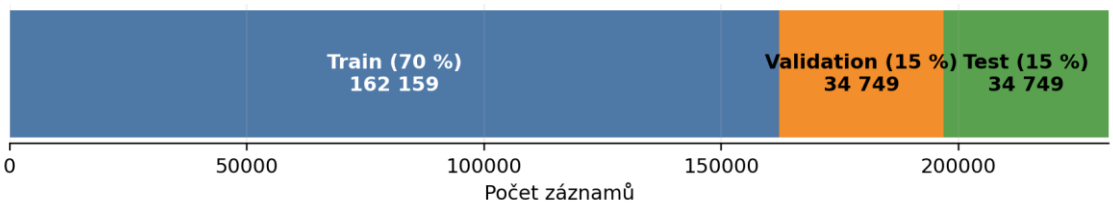
zdroj: vlastní zpracování

3.1.2 Rozdělení dat

Po vytvoření finální vstupní matice příznaků (tabulky vstupních proměnných) a cílové proměnné byla data rozdělena na trénovací, validační a testovací část. Rozdělení bylo provedeno stratifikovaně podle cílové proměnné, aby byl ve všech částech zachován přibližně stejný poměr tříd jako v původním souboru. Uvedený postup je důležitý zejména u klasifikačních úloh s nerovnoměrným zastoupením tříd, protože snižuje riziko zkreslení výsledků při trénování i při následném vyhodnocení modelů. Pokud by rozdělení proběhlo pouze náhodně bez ohledu na zastoupení tříd, mohlo by dojít k tomu, že některé části dat by obsahovaly nepřiměřeně málo rizikových případů, což by negativně ovlivnilo jak učení modelu, tak interpretaci dosažených výsledků *(Kuhn a Johnson, 2013)***.**

Poměr rozdělení byl nastaven na 70 % pro trénovací sadu, 15 % pro validační sadu a 15 % pro testovací sadu. Zvolený poměr umožňuje využít dostatečně velkou část dat pro natrénování modelů a současně ponechat dvě samostatné části dat, které nejsou použity přímo při učení modelu. V aktuálním běhu notebooku popsané rozdělení odpovídá přibližně 162 159 záznamům v trénovací části a 34 749 záznamům ve validační i testovací části.

Pro zajištění reprodukovatelnosti experimentů byl v notebooku použit pevně zvolený seed generátoru náhodných čísel (RANDOM_STATE = 42). Díky tomu při opakovaném spuštění vzniká stejné rozdělení dat i stejné náhodné výběry v krocích, které obsahují prvek náhodnosti. To umožňuje lépe porovnávat výsledky jednotlivých modelů a experimentálních variant bez vlivu nahodilých odchylek. Schéma tohoto rozdělení zachycuje obrázek 7.



Obrázek 7: Rozdělení dat na trénovací, validační a testovací sadu

zdroj: vlastní zpracování

Trénovací sada slouží k vlastnímu natrénování modelů. Současně je využita také při ladění hyperparametrů, a to pomocí stratifikované křížové validace. To znamená, že se trénovací data při ladění opakovaně rozdělují do několika částí, přičemž model je vždy natrénován na většině z nich a průběžně ověřen na části zbývající. Role ověřovací části se přitom střídá, takže výsledný odhad výkonnosti není závislý pouze na jednom konkrétním rozdělení dat. Zvolený postup je vhodnější než ladění hyperparametrů na jediné validační sadě, protože poskytuje stabilnější a metodicky spolehlivější odhad toho, které nastavení modelu funguje dobře obecně, a ne pouze na jednom konkrétním výřezu dat *(scikit-learn developers, nedatováno f).*

Validační sada je oproti tomu samostatná část dat, která se nepoužívá k učení modelu ani k samotnému ladění hyperparametrů. Jejím účelem je průběžně porovnávat již natrénované a naladěné varianty modelů na datech, která během trénování neviděly. V práci validační sada slouží zejména k porovnání výsledků jednotlivých modelových variant, k vyhodnocení jejich chování na dosud neviděných datech a k výběru vhodného rozhodovacího prahu, tedy hranice, od níž je pravděpodobnostní výstup modelu převeden na konečné zařazení do nerizikové nebo rizikové třídy. Samostatná validační sada tak plní jinou roli než křížová validace uvnitř trénovacích dat: zatímco křížová validace pomáhá najít vhodné nastavení modelu, validační sada umožňuje nezávisleji posoudit, jak se již naladěný model chová v praxi.

Testovací sada je ponechána zcela stranou až do závěrečného kroku analýzy. Není využívána ani při ladění hyperparametrů, ani při výběru rozhodovacího prahu. Jejím úkolem je poskytnout co nejobjektivnější finální odhad generalizační schopnosti výsledného řešení. Díky tomu lze lépe posoudit, nakolik by se vybraný model mohl chovat na nových datech mimo rámec samotného experimentu.

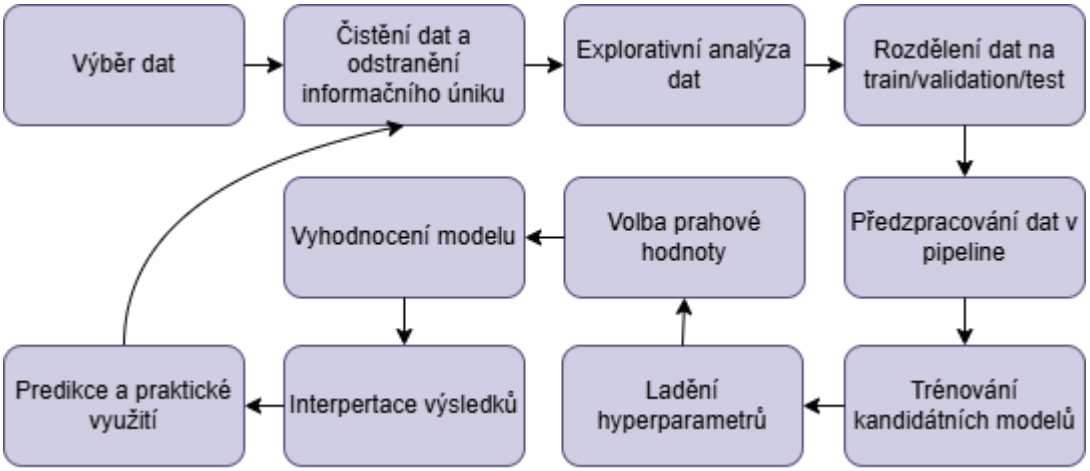
Z metodického hlediska je důležité, že toto rozdělení odděluje tři různé role dat. Trénovací sada slouží k učení modelu, validační sada k průběžnému porovnání naladěných variant a k volbě rozhodovacího prahu a testovací sada k finálnímu nezávislému ověření. Takové uspořádání omezuje riziko přeučení na jedinou část dat a současně zvyšuje důvěryhodnost dosažených výsledků *(scikit-learn developers, nedatováno f)*.

3.2 Zvolený výzkumný přístup a metodologický rámec (RAD)

Praktická část práce je realizována iterativně podle principů metodiky Rapid Application Development (RAD). Místo jednorázového návrhu finálního modelu je postup rozdělen do několika navazujících kroků, v nichž jsou postupně ověřovány různé varianty zpracování dat, modelovacích přístupů a způsobů vyhodnocení.

V souladu se zadáním byla v průběhu vývoje využita také generativní umělá inteligence jako podpůrný nástroj. Sloužila především k návrhům alternativních implementačních postupů, refaktoringu a zpřehlednění kódu, diagnostice chyb, doplňování komentářů a redakčním úpravám textových částí notebooku i práce. Pomocně byla využívána také při dohledávání veřejně dostupných implementačních inspirací. Odborná rozhodnutí o volbě datových transformací, modelů, experimentálního nastavení, interpretaci výsledků i finální podobě řešení však provedla autorka. Všechny podstatné výstupy byly ručně zkontrolovány, upraveny a ověřeny opakovaným spuštěním notebooku a kontrolou metrik.

Generativní umělá inteligence v práci vystupuje jako nástroj podporující rychlejší iteraci v duchu metodiky RAD. Toto vymezení odpovídá důrazu na lidský dohled, validaci výstupů a odpovědnost za konečné rozhodnutí při využívání generativní AI *(Hartman a kol., 2025).* Celkovou návaznost jednotlivých kroků použitých v praktické části shrnuje obrázek 8. Generativní umělá inteligence byla využita také při jazykové a stylistické korektuře textu bakalářské práce. Obsahová správnost i finální podoba textu však zůstaly v odpovědnosti autorky.



Obrázek 8: Schéma pracovního postupu praktické části práce

zdroj: vlastní zpracování v programu draw.io

V první fázi byla ověřena použitelnost zvoleného datového souboru a provedena základní příprava dat, zahrnující jejich čištění, transformaci a úpravu do podoby vhodné pro další analýzu. Následně byla realizována explorativní analýza dat, jejímž cílem bylo lépe porozumět struktuře dat, rozdělení proměnných a vztahům mezi nimi. Součástí této fáze byla také vybraná neřízená analýza, zejména shlukování, které sloužilo k doplňkovému pohledu na vnitřní strukturu dat.

V dalších krocích byly vytvořeny základní modelové varianty a postupně doplňovány další experimentální části, zejména předzpracování dat bez úniku informací, výběr a redukce příznaků, ladění hyperparametrů, volba rozhodovacího prahu, vyvážení tříd a diagnostika chování modelu při učení a orientační posouzení generalizační schopnosti (schopnost modelu dosahovat dobrých výsledků i na dříve neviděných datech). Jednotlivé kroky na sebe navazují, současně však umožňují vracet se k dřívějším rozhodnutím a upravovat je podle průběžně zjištěných výsledků.

Součástí finální verze notebooku je vedle analytických sekcí také formulářová predikce pro jednoho žadatele a dávková predikce ze souboru CSV. Praktická část tak neslouží pouze k vyhodnocení modelů, ale i k demonstraci, jak lze zvolený modelový postup využít v uživatelsky orientovaném prototypu.

Použití notebooku je navrženo v několika navazujících krocích: nejprve se provede příprava a rozdělení dat, následně explorativní analýza a modelové experimenty, poté volba finální konfigurace a rozhodovacího prahu a nakonec demonstrační predikce pro jednoho žadatele nebo dávková predikce ze souboru CSV. Praktická část tedy nevytváří samostatnou produkční aplikaci, ale reprodukovatelný analytický prototyp, na němž lze transparentně ukázat celý pracovní postup.

Autorský přínos práce spočívá především v návrhu metodického postupu, definici cílové proměnné, implementaci a propojení experimentálních větví, volbě vyhodnocovacích metrik, interpretaci výsledků a vytvoření uživatelských predikčních částí notebooku. Převzaté prvky tvoří veřejný datový soubor, standardní implementace použitých knihoven a vybrané inspirační motivy z veřejně dostupných notebooků, které byly v práci dále upraveny a zasazeny do vlastního řešení.

Výhodou tohoto přístupu je zejména možnost rychle porovnávat více variant řešení a postupně přecházet od jednoduchých, základních modelů k metodicky propracovanějším experimentům. Zvolený přístup je vhodný i vzhledem k tomu, že praktická část práce je realizována v interaktivním notebooku, který podporuje opakovatelnost postupu, průběžnou dokumentaci a rychlé ověřování změn.

3.2.1 Reprodukovatelnost, prostředí Kaggle a veřejná dostupnost

Praktická část byla připravena jako interaktivní notebook v prostředí Kaggle/Jupyter. Prostředí podporuje sdílení datových souborů, notebooků a průběžné experimentování nad jednotně definovaným datovým zdrojem. Textová část práce proto shrnuje především klíčové výsledky, zatímco detailní mezivýstupy, doplňkové grafy, technické přepínače a interaktivní prvky zůstávají v přiloženém notebooku.

Z hlediska opakovatelnosti jsou v notebooku explicitně definovány klíčové kroky pracovního postupu: načtení veřejného datového souboru Credit Risk Analysis *(ranadeep, nedatováno),* vytvoření cílové proměnné, stratifikované rozdělení dat, zřetězené předzpracování (pipeline), trénování a ladění modelů, volba rozhodovacího prahu a finální testovací vyhodnocení. Součástí odevzdaných materiálů je také samotný notebook, takže vedoucí i oponent mohou jednotlivé kroky znovu projít a posoudit.

Vedle samotného datového zdroje je důležitý i komunitní aspekt platformy Kaggle: umožňuje porovnávat otevřené notebooky, inspirovat se jejich strukturou a současně transparentně deklarovat, které části pracovního postupu byly převzatou inspirací a které představují vlastní přínos. Použitý princip byl v práci dodržen jak v notebooku, tak v textové části práce.

3.2.2 Rámcový přehled metod a sekcí notebooku

Pro rychlou orientaci shrnuje následující přehled hlavní sekce notebooku a metody, které jsou v nich použity:

• **Příprava dat:** načtení souboru loan.csv, filtrace stavů *loan_status*, vytvoření binární cílové proměnné, převody datových typů, odstraňování proměnných s rizikem úniku informací a tvorba jednoduchých odvozených příznaků (feature engineering).

• **Explorativní analýza**: přehled datových typů, podílů chybějících hodnot a kardinality, distribuce vybraných proměnných, boxploty, IQR pravidlo, bodové grafy, explorativní analýza s ohledem na cílovou proměnnou a kontrola normality pomocí QQ grafů a testů normality.

• **Neřízená explorace**: FAMD, UMAP, K-Prototypes, elbow graf a profilace shluků pro doplňkové porozumění vnitřní struktuře smíšených dat.

• **Předzpracování a pipeline**: SimpleImputer, StandardScaler, OneHotEncoder, OrdinalEncoder, ColumnTransformer a Pipeline s důrazem na prevenci úniku informací.

• **Modelování**: Logistic Regression, Random Forest a HistGradientBoostingClassifier, základní porovnání modelových rodin a následné ladění pomocí RandomizedSearchCV.

• **Výběr příznaků a experimenty**: RFECV, Mutual Information, WoE/IV, redukovaný model s přibližně deseti proměnnými, vyvážení tříd, křivka učení a intervalové hodnocení metrik z křížové validace.

• **Interpretace a prototyp:** permutation importance, validační a testovací reporty, formulářová predikce pro jednoho žadatele a dávková predikce ze souboru CSV.

3.3 Obecné fáze přípravy a zpracování dat

3.3.1 Čištění a úpravy

Příprava dat začíná předběžnou redukcí proměnných, jejichž ponechání by nepřinášelo predikční informaci nebo by výrazně komplikovalo další zpracování. Již v úvodu pracovního postupu byly proto odstraněny identifikátory a sloupce s téměř unikátními hodnotami, zejména *url*, *id* a *member*_id, dále volný text s velmi vysokou variabilitou hodnot, například *desc* a *title*, a také konstantní sloupce bez informační hodnoty, například *policy_code*. Daný krok snižuje paměťovou náročnost, zjednodušuje práci s kategoriálními proměnnými a omezuje riziko přeučení na nerelevantních znacích.

Proměnná *emp_title* představovala ručně vyplněný název pracovní pozice žadatele. Ačkoli by teoreticky mohla nést určitou informační hodnotu pro predikci úvěrového rizika, do výsledného modelu nebyla zařazena. Důvodem byla velmi vysoká variabilita hodnot, která vedla k stovkám tisíc unikátních kategorií. Taková struktura zvyšuje riziko nadměrné dimenzionality a šumu při standardním kódování kategoriálních proměnných a současně výrazně prodlužuje výpočetní čas v prostředí notebooku na platformě Kaggle.

Následně byly provedeny základní převody datových typů. Textově uložené procentuální hodnoty, například *int_rate* a *revol_util*, byly převedeny na numerický formát. Proměnná term byla transformována z textového zápisu typu „36 months“ na číselnou reprezentaci délky úvěru v měsících. U proměnné *emp_length* byl použit převod textového zápisu délky zaměstnání na numerickou aproximaci vyjádřenou v letech. Cílem těchto kroků bylo převést původně heterogenní datové reprezentace do podoby vhodné pro další analytické zpracování.

Vedle základních převodů byly v notebooku vytvořeny také jednoduché odvozené ukazatele, které doplňují původní atributy o lépe interpretovatelné charakteristiky finanční situace žadatele. Jejich smyslem nebylo nahrazovat původní proměnné, ale zachytit některé vztahy mezi výší úvěru, příjmem, pravidelnou splátkou a zadlužeností v kompaktnější podobě. Přehled použitých odvozených proměnných uvádí tabulka 2.

Tabulka 2: Přehled odvozených proměnných použitých v analýze

Proměnná	Přibližný výpočet	Význam
loan_to_income	loan_amnt / annual_inc	Vyjadřuje velikost úvěru vzhledem k ročnímu příjmu žadatele.
installment_to_monthly_income	installment / (annual_inc / 12)	Zachycuje zatížení příjmu pravidelnou měsíční splátkou.
revol_bal_to_limit	revol_bal / total_rev_hi_lim	Aproximuje míru využití revolvingového úvěrového rámce.
balance_to_income	tot_cur_bal / annual_inc	Popisuje celkovou bilanční zadluženost vzhledem k příjmu.

zdroj: vlastní zpracování

Zvolené ukazatele rozšiřují původní datovou reprezentaci o doplňující pohled na finanční zatížení klienta. Mohou tak pomoci lépe vystihnout schopnost žadatele úvěr splácet, protože nezachycují pouze absolutní hodnoty jednotlivých proměnných, ale také jejich vzájemné poměry.

Zvláštní pozornost byla věnována prevenci úniku informací. Z datového souboru byly proto odstraňovány sloupce, které mohou přímo nebo nepřímo odrážet výsledek úvěru po jeho schválení. Jednalo se zejména o proměnné související s platbami, inkasem, vymáháním, zbývající jistinou nebo vypořádáním dluhu. Tyto proměnné by sice mohly uměle zvyšovat výkonnost modelu, avšak při reálném rozhodování o novém žadateli by nebyly k dispozici *(James a kol., 2021).*

V samostatném kroku byly dále konzistentně odstraněny vybrané sloupce ze všech datových částí, tedy z množin X_*train*, X_*valid*, X_*test* i z finální modelové matice X_*final*. Smyslem tohoto kroku bylo jednak omezit únik informací prostřednictvím proměnných, které vznikají až po poskytnutí úvěru nebo přímo odrážejí jeho další průběh, jednak zachovat jednotné a stabilní schéma vstupních dat pro navazující pipeline předzpracování a modelování. V notebooku je uvedený seznam definován konfiguračně v proměnné **DROP_ALWAYS**, takže jej lze v případě potřeby upravit podle konkrétní experimentální větve.

Do této skupiny spadají zejména proměnné časového charakteru nebo proměnné dostupné až po poskytnutí úvěru a další atributy, které by při reálném rozhodování o novém žadateli nebyly v okamžiku predikce k dispozici. Jejich ponechání by mohlo vést k nadhodnocení výkonnosti modelu a zároveň ke zbytečné nekonzistenci mezi jednotlivými datovými částmi. Konzistentní odstranění těchto sloupců proto přispívá jak k metodické správnosti experimentu, tak k technické robustnosti celého pracovního postupu.

Po základních úpravách byla finální množina kandidátních proměnných dále zúžena pomocí jednoduchých filtračních pravidel. V notebooku je tato část řešena interaktivně prostřednictvím prahů pro maximální přípustný podíl chybějících hodnot a maximální kardinalitu kategoriálních proměnných. Ve výchozím nastavení jsou odstraňovány proměnné s podílem chybějících hodnot vyšším než 70 % a kategoriální proměnné s více než 200 unikátními hodnotami. Výsledkem fáze je matice X_*final*, která již obsahuje pouze proměnné určené pro následné modelování.

Je však nutné odlišit datový soubor používaný pro explorativní analýzu od finální modelové matice. Hlavní EDA byla v notebooku prováděna nad celým analyzovaným souborem po základním očištění, sjednocení proměnných a po prvotní redukci sloupců s vysokým podílem chybějících hodnot nebo nevhodnou kardinalitou. Tím bylo možné popsat strukturu dat na úrovni celého souboru pozorování, avšak již nad proměnnými, které byly považovány za vhodné pro další analýzu. Matice X_*final* pak slouží jako navazující datová reprezentace pro modelování a experimenty.

3.3.2 Předzpracování dat a zřetězené zpracování (pipeline)

Pro numerické proměnné byla ve výchozím nastavení použita imputace mediánem a pro kategoriální proměnné imputace nejčastější hodnotou. Jde o základní univariační strategie podporované nástrojem *SimpleImputer*, který pro numerická data umožňuje například průměr nebo medián a pro kategoriální data zejména strategii *most_frequent* nebo constant *(scikit-learn developers, nedatováno a).*

Pro numerické proměnné byl zvolen medián, protože je méně citlivý na extrémní hodnoty než průměr, a je tak vhodný i pro proměnné s vychýleným rozdělením. Ve výchozím nastavení notebooku je u kategoriálních proměnných použita imputace nejčastější hodnotou. Varianta nahrazení samostatnou kategorií, například hodnotou „MISSING“, je v notebooku podporována jako alternativní možnost, ale neodpovídá standardnímu běhu aktuální verze.

Složitější multivariační imputace pomocí *IterativeImputer* nebyla v práci použita. Důvodem je jednak vyšší výpočetní náročnost, jednak skutečnost, že daný nástroj je ve scikit-learn stále veden jako experimentální. Dokumentace současně uvádí, že v predikčním kontextu nemusí být složitější imputace nutně výhodnější a že při použití silnějších modelů mohou jednoduché imputace dosahovat stejně dobrých nebo i lepších výsledků než komplexnější postupy. Zvolená strategie tedy odpovídá snaze o metodicky přiměřené, stabilní a výpočetně rozumné předzpracování dat. Reprezentativní výřez použité pipeline a jejích hlavních transformačních kroků je zachycen na obrázku 9.

```
numeric_pipe_linear = Pipeline(steps=[
    ("imputer", make_numeric_imputer()),
    ("scaler", StandardScaler(with_mean=False))
])

categorical_pipe_ohe = Pipeline(steps=[
    ("imputer", make_categorical_imputer()),
    ("onehot", OneHotEncoder(handle_unknown="ignore", sparse_output=True))
])

preprocess_linear = ColumnTransformer(
    transformers=[
        ("num", numeric_pipe_linear, num_cols),
        ("cat", categorical_pipe_ohe, cat_cols)
    ]
)
```

Obrázek 9: Ukázka pipeline předzpracování

zdroj: vlastní zpracování

Klíčovým metodickým prvkem je využití scikit-learn pipeline. Pipeline propojuje předzpracování dat a vlastní model do jednoho navazujícího postupu. Její význam nespočívá pouze v technickém zpohodlnění práce, ale především v metodické správnosti. Transformace, které se učí parametry z dat, například imputace, škálování nebo vytváření kódování kategorií, jsou v rámci pipeline nastavovány pouze na trénovací sadě. Na validační a testovací data jsou pak již jen aplikovány. Tím je zajištěno, že model nemá během trénování nepřímý přístup k informacím z dalších částí dat a nedochází k úniku informací *(scikit-learn developers, nedatováno d).*

Současně pipeline zvyšuje reprodukovatelnost postupu, protože zajišťuje, že všechny kroky probíhají vždy ve stejném pořadí a stejným způsobem. Snižuje se tím riziko metodických chyb, které by mohly vzniknout při ručním a nekonzistentním zpracování různých částí datového souboru.

V notebooku byly použity dvě hlavní větve preprocessingu. Pro lineární modely je využita pipeline **preprocess_linear**, která kombinuje imputaci numerických proměnných, jejich škálování a one-hot encoding kategoriálních atributů. One-hot encoding je zde ve výchozím nastavení doplněn omezením kardinality pomocí parametrů *min_frequency* = 0,01 a *max_categories* = 50, aby se snížila paměťová náročnost a riziko příliš vysoké dimenze datového prostoru. Parametr *min_frequency* = 0,01 znamená, že samostatně jsou ponechány pouze ty kategorie, které se vyskytují alespoň u 1 % záznamů v trénovacích datech. Méně časté kategorie jsou sloučeny do společné skupiny vzácných hodnot. Parametr *max_categories* současně omezuje maximální počet výsledných kategorií pro jeden atribut. Toto nastavení pomáhá snížit výpočetní náročnost, omezit vliv velmi vzácných kategorií a udržet reprezentaci dat v rozumné velikosti *(scikit-learn developers, nedatováno b).*

Pro stromové modely je využita pipeline **preprocess_tree**, která zahrnuje imputaci chybějících hodnot a ordinal encoding kategoriálních atributů. Dokumentace scikit-learn uvádí, že OrdinalEncoder převádí jednotlivé kategoriální proměnné na ordinální číselné kódy a vytváří tak pro každý atribut jeden číselný sloupec. V rámci této práce byl proto ordinal encoding zvolen s ohledem na kompaktnější reprezentaci dat a nižší výpočetní náročnost při modelování stromovými algoritmy. Současně byl v notebooku OrdinalEncoder nastaven pomocí parametrů *handle_unknown* = "use_encoded_value" a *unknown_value* = -1, takže případné dříve neviděné kategorie při transformaci nevedou k chybě, ale jsou převedeny na zvláštní kód. Uplatněné nastavení odpovídá i oficiální dokumentaci scikit-learn, podle níž parametr *handle_unknown* = "use_encoded_value" umožňuje neznámé kategorie zachytit a parametr *unknown_value* určuje jejich číselné označení. *(scikit-learn developers, nedatováno c)*

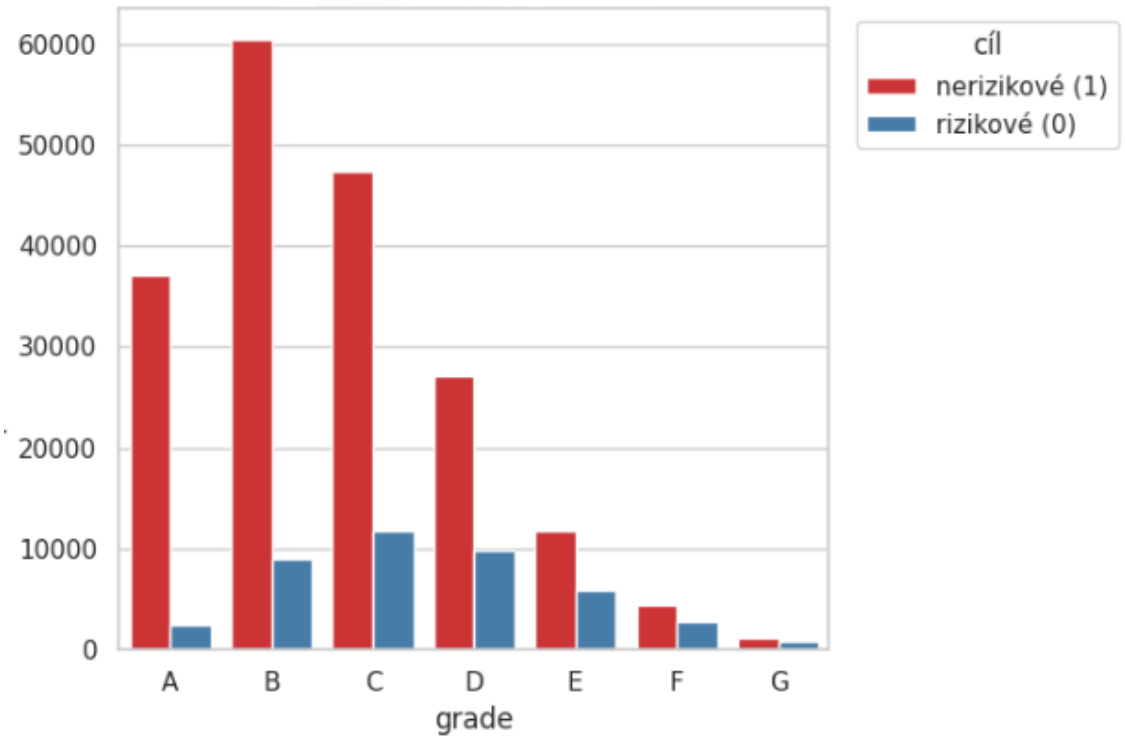
3.3.3 Explorativní a neřízená analýza dat

Explorativní analýza dat (EDA) představuje důležitou fázi zpracování datového souboru, protože umožňuje lépe porozumět jeho struktuře ještě před samotným modelováním. Jejím cílem není pouze popsat data, ale také odhalit možné problémy, které by mohly ovlivnit další analytické kroky, například podíl chybějících hodnot, přítomnost odlehlých pozorování, rozdílné datové typy nebo výskyt kategoriálních proměnných s velmi vysokým počtem unikátních hodnot. Současně EDA poskytuje kontext pro interpretaci pozdějších modelových výsledků, protože pomáhá pochopit, jaké vlastnosti data skutečně mají a jak se od sebe jednotlivé skupiny pozorování liší.

V této práci tedy EDA neznamená pouze popis finální redukované matice X_*final* určené pro modelování. Explorativní a supervizovaná analýza byla vedena nad celým dostupným souborem záznamů po základním vyčištění a odstranění zjevně problematických proměnných nebo proměnných s rizikem úniku informací, aby zjištěné vzorce co nejlépe odpovídaly struktuře celého datového souboru. Přísnější filtrování podle podílu chybějících hodnot a kardinality bylo až následným krokem pro konstrukci kompaktnější modelové matice.

V rámci této práce EDA zahrnovala zejména přehled datových typů, podílů chybějících hodnot, paměťové náročnosti a kardinality kategoriálních proměnných. Součástí analýzy byly také vizualizace distribucí vybraných numerických proměnných, boxploty, bodové grafy a korelační analýza. Zvláštní pozornost byla věnována porovnání skupin podle cílové proměnné, aby bylo možné lépe posoudit, které atributy se mezi rizikovými a nerizikovými případy zřetelně odlišují.

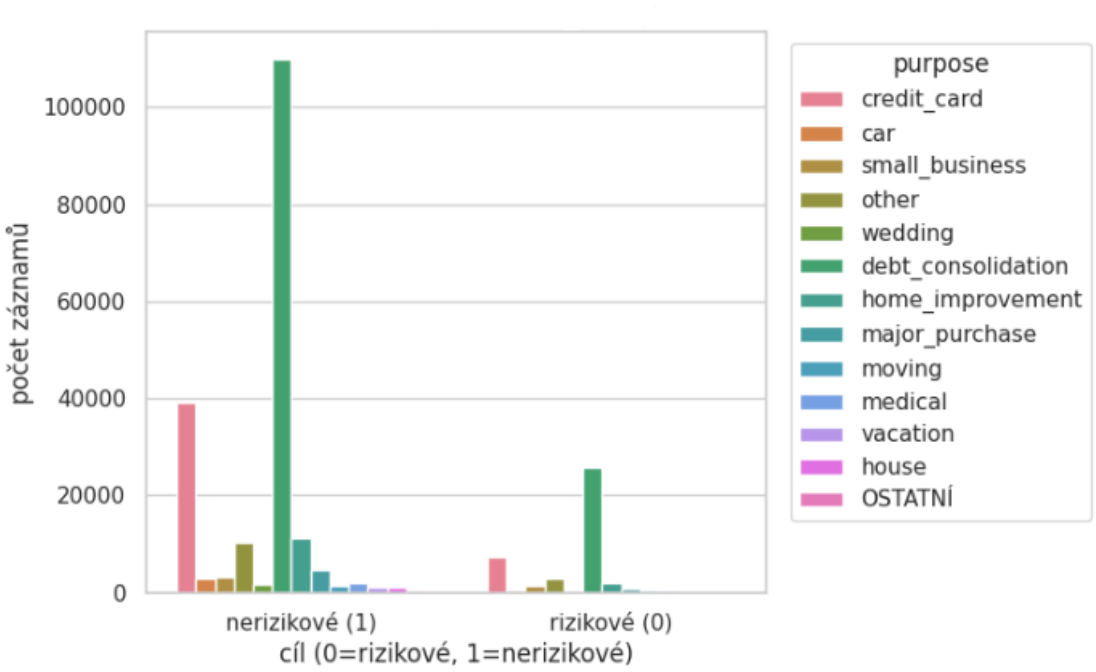
Obrázek 10 zachycuje rozdělení stupně úvěru podle cílové proměnné, obrázek 11 vztah mezi účelem půjčky a cílovou proměnnou, obrázek 12 rozdělení úrokové sazby, obrázek 13 rozdělení výše půjčky a obrázek 14 rozdělení ukazatele DTI. Explorativní analýza tak v této práci nepředstavuje pouze doplňkovou vizualizační část, ale důležitý podklad pro další rozhodnutí o předzpracování dat, výběru proměnných a interpretaci výsledků modelování *(Komorowski, 2016).*



Obrázek 10: Rozdělení stupně úvěru podle cílové proměnné

zdroj: vlastní zpracování

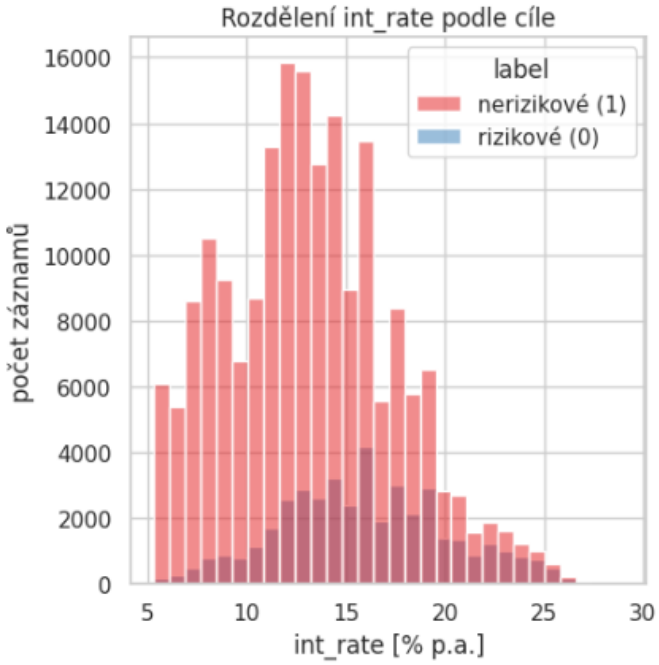
Obrázek 10 ukazuje rozdělení stupně úvěru podle cílové proměnné. Je patrné, že největší část záznamů je soustředěna ve stupních B a C, zatímco stupně E až G jsou zastoupeny výrazně méně. Současně lze pozorovat, že s horším stupněm úvěru relativně roste zastoupení rizikových případů, což odpovídá očekávání, že horší úvěrový profil bývá spojen s vyšší pravděpodobností selhání. Proměnná grade se tak jeví jako věcně významný atribut pro další modelování.



Obrázek 11: Rozdělení cílové proměnné podle účelu půjčky

zdroj: vlastní zpracování

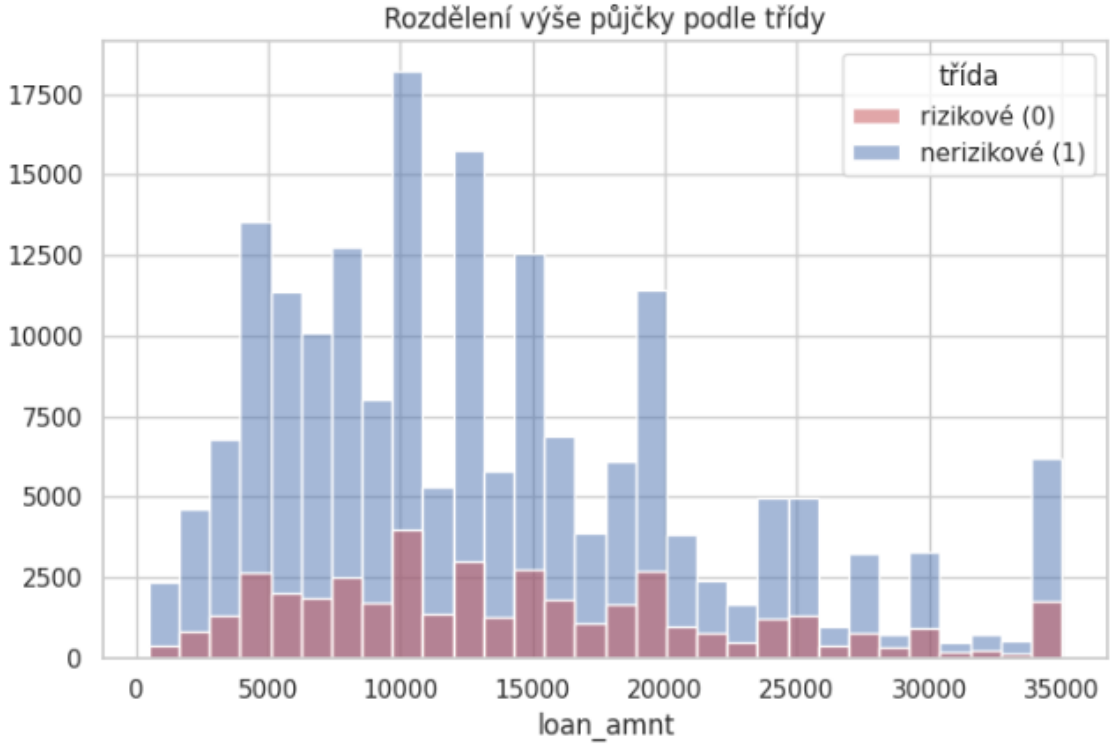
Obrázek 11 zachycuje rozdělení cílové proměnné podle účelu půjčky. Největší zastoupení mají v obou třídách účely *debt_consolidation* a *credit_card*, což odpovídá i jejich celkové četnosti v datovém souboru. Současně je patrné, že některé účely půjčky se mezi třídami liší více než jiné, což naznačuje, že proměnná *purpose* může nést doplňkovou informaci o rizikovosti úvěru. Jejich význam je však vhodné posuzovat v kombinaci s dalšími charakteristikami klienta a úvěru.



Obrázek 12: Rozdělení úrokové sazby podle cílové proměnné

zdroj: vlastní zpracování

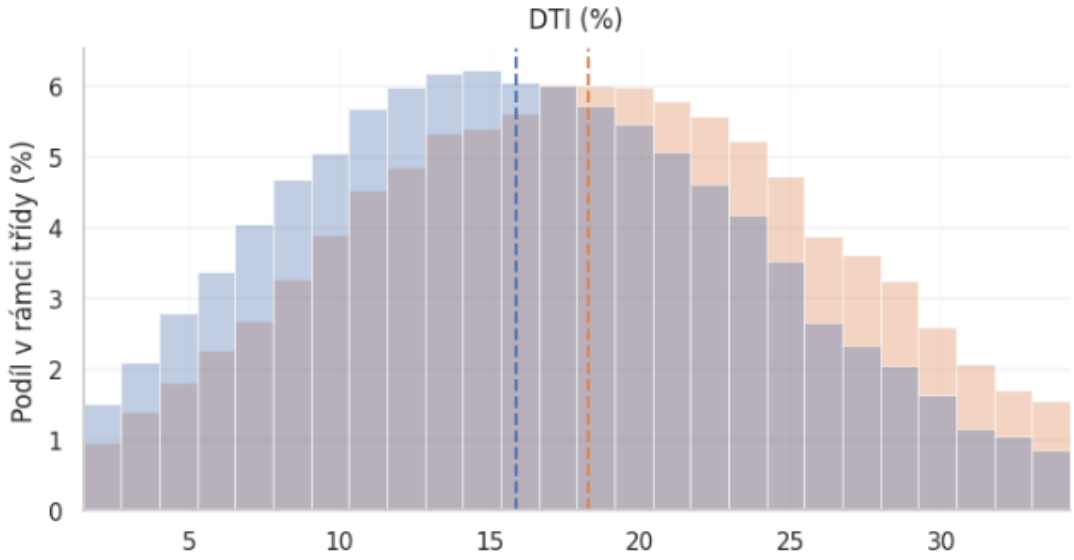
Z obrázku 12 je patrné, že rozdělení úrokové sazby se mezi rizikovými a nerizikovými případy částečně liší. Rizikové případy jsou relativně častěji zastoupeny ve vyšších hodnotách proměnné int_*rate*, což odpovídá očekávání, že úvěry s vyšší úrokovou sazbou bývají spojeny s vyšší vnímanou rizikovostí klienta. Současně je však vidět značný překryv obou tříd, a proměnná proto sama o sobě nestačí k jednoznačnému oddělení rizikových a nerizikových případů. Její význam je nutné posuzovat v kombinaci s dalšími atributy.



Obrázek 13: Rozdělení výše půjčky podle cílové proměnné

zdroj: vlastní zpracování

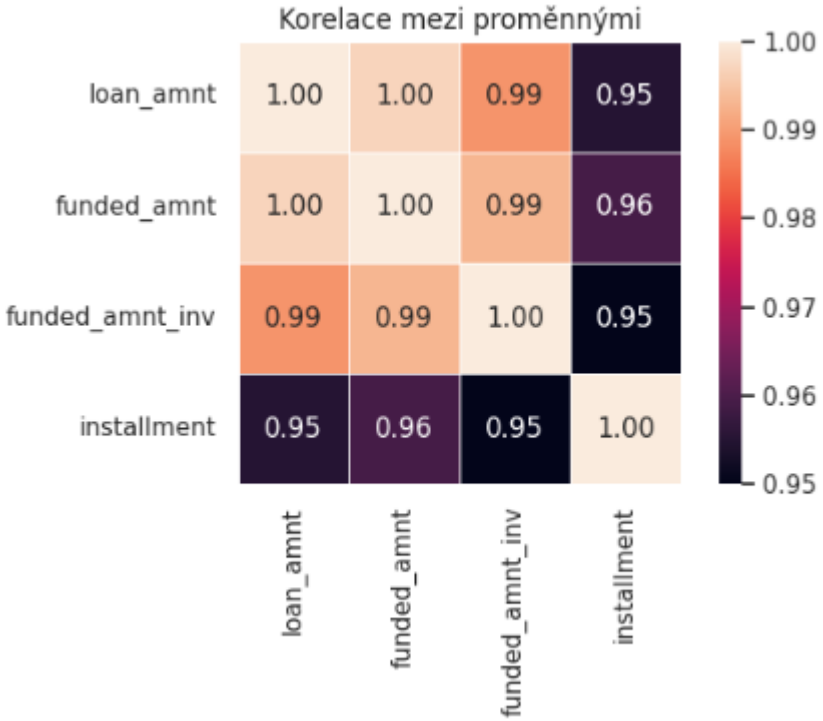
Obrázek 13 zachycuje rozdělení výše půjčky podle cílové proměnné v dolarech. Z porovnání je patrné, že se obě třídy ve značné míře překrývají, takže samotná výše úvěru nepředstavuje dostatečný diskriminační znak pro oddělení rizikových a nerizikových případů. Přesto lze pozorovat dílčí rozdíly v četnostech některých intervalů, což naznačuje, že proměnná *loan_amnt* může nést určitou doplňkovou informaci. V dalších krocích proto dává smysl hodnotit její význam společně s dalšími finančními ukazateli.



Obrázek 14: Rozdělení ukazatele DTI podle cílové proměnné

zdroj: vlastní zpracování

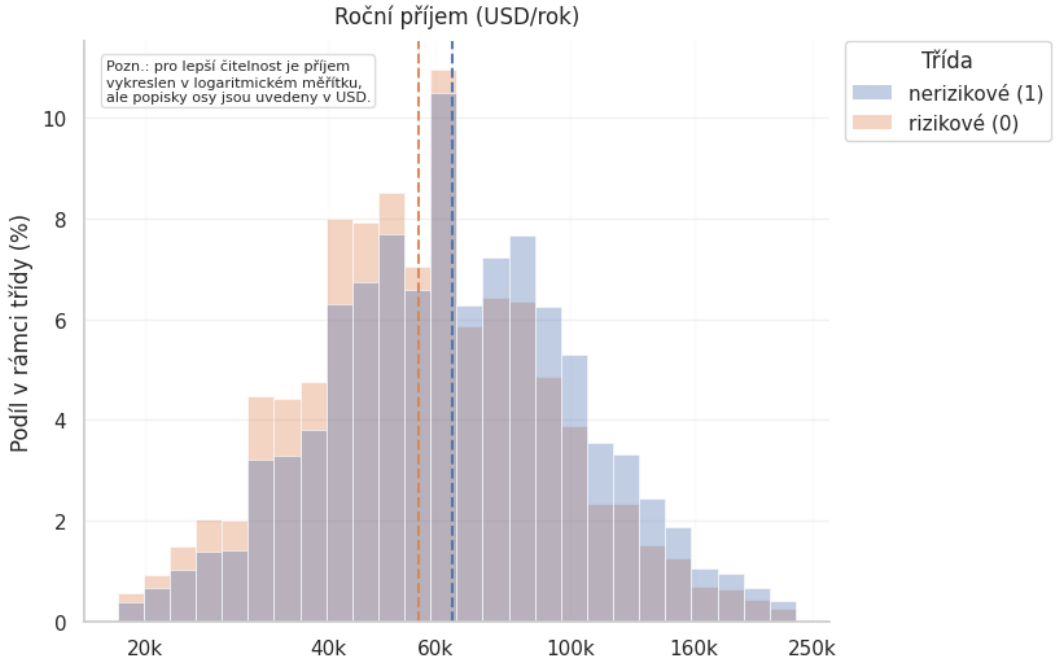
Obrázek 14 zachycuje rozdělení ukazatele DTI (Debt-to-Income ratio) podle cílové proměnné. Z porovnání je patrné, že rizikové případy jsou relativně častěji zastoupeny při vyšších hodnotách DTI, zatímco u nižších hodnot převládají případy nerizikové. To naznačuje, že vyšší zadlužení vůči příjmu může souviset se zvýšenou pravděpodobností problémů se splácením. Současně je však rozložení obou tříd stále výrazně překryté, a proměnná dti proto sama o sobě nepředstavuje dostatečný znak pro jednoznačné oddělení rizikových a nerizikových případů.



Obrázek 15: Korelační heatmapa pro vybrané proměnné

zdroj: vlastní zpracování

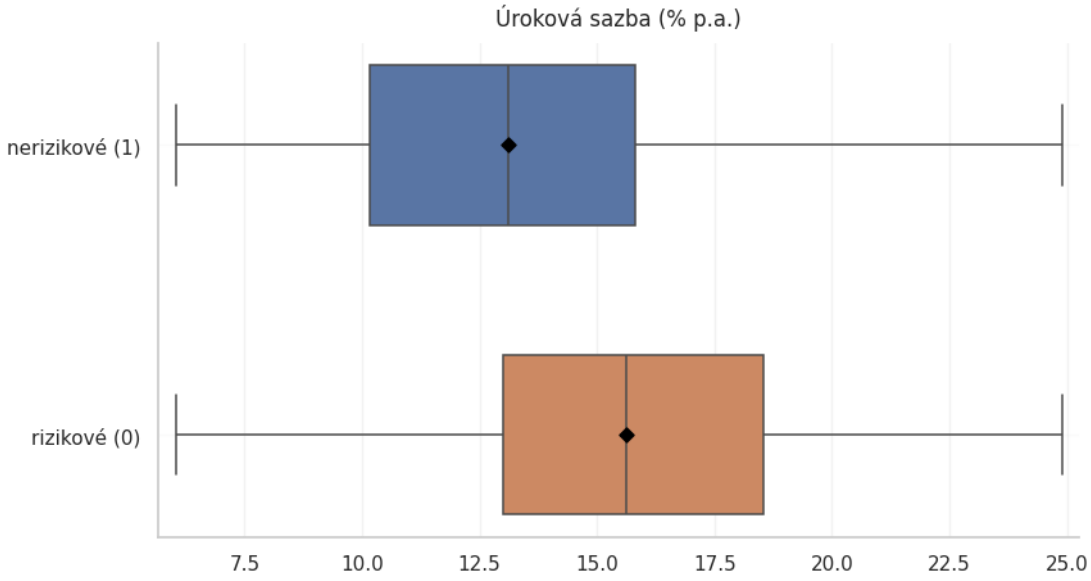
Pro doplnění základní EDA byla provedena také stručná kontrola nadbytečnosti mezi vybranými částkovými proměnnými *loan_amnt*, *funded_amnt*, *funded_amnt_inv* a *installment*. Korelační mapa ukazuje na obrázku 15, že mezi prvními třemi uvedenými proměnnými existuje velmi silná vzájemná souvislost, zatímco proměnná installment je s nimi rovněž výrazně propojena. Výsledek naznačuje, že tyto atributy do značné míry zachycují podobný aspekt úvěru a mezi částkovými ukazateli se objevuje výrazná informační překryvnost. Zvolený krok sloužil pouze jako podpůrná explorativní kontrola a nebyl sám o sobě finálním kritériem pro výběr příznaků.



Obrázek 16: Roční příjem

zdroj: vlastní zpracování

V rámci explorativní analýzy byla pozornost věnována také odlehlým hodnotám a tvaru rozdělení vybraných numerických proměnných, a to zejména pomocí histogramů, boxplotů a posouzení šikmosti. Zvláštní důraz byl kladen na atributy se silnou pravostrannou vychýleností, u nichž bylo porovnáváno rozdělení před a po aplikaci transformace *log1p,* tedy *ln (1 + x).* Tato transformace pomáhá zmírnit asymetrii rozdělení, omezit vliv extrémních hodnot a současně je použitelná i pro proměnné obsahující nulové hodnoty. Obrázek 16 zachycuje rozdělení proměnné ročního příjmu, na němž je patrná výrazná pravostranná vychýlenost i odlišný průběh rozdělení mezi oběma třídami. Na obrázku 17 je pak uveden skupinový boxplot proměnné *int_rate*, z něhož vyplývá, že u rizikových případů dosahuje úroková sazba obecně vyšších hodnot než u případů nerizikových. Výstupy tak naznačují, že některé proměnné se mezi třídami odlišují již v explorativní fázi analýzy a mohou být významné i pro další modelovací kroky.

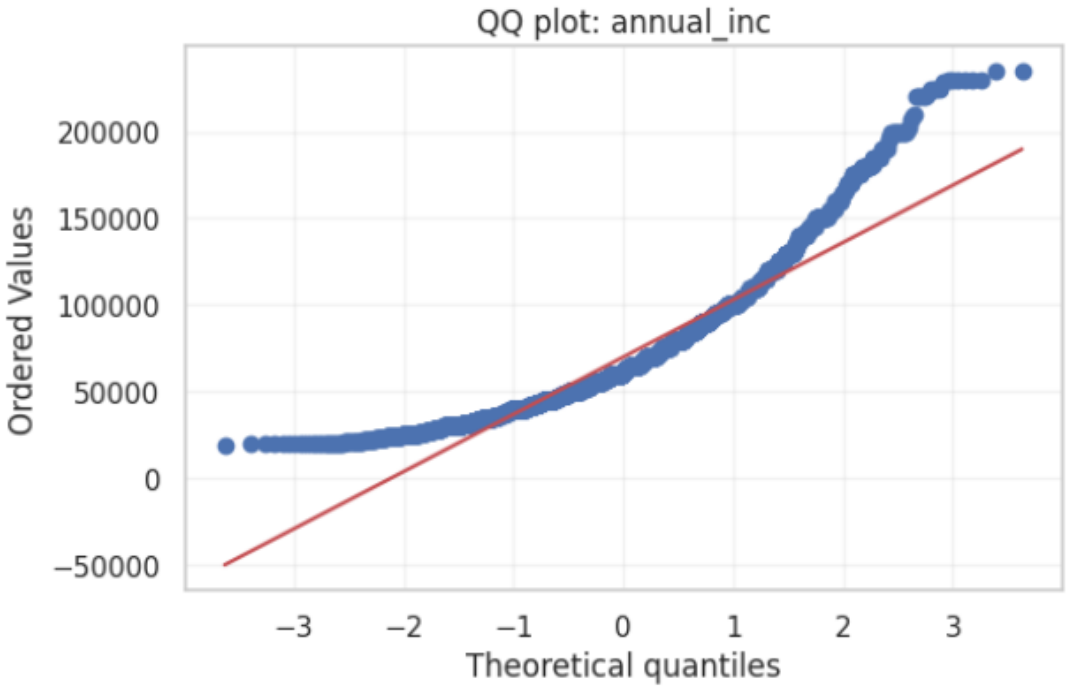


Obrázek 17: Boxplot úrokové sazby podle cílové proměnné

zdroj: vlastní zpracování

Transformace sloužila jednak k posouzení vhodnosti logaritmické škály pro diagnostické a vizualizační účely, jednak jako podklad pro konzervativní filtraci pouze nejextrémnějších hodnot pomocí pravidla založeného na interkvartilovém rozpětí. V praktické implementaci přitom nešlo o využití jediné specializované funkce, ale o explicitní výpočet kvartilů, interkvartilového rozpětí a mezních hodnot nad logaritmicky transformovanými daty u vybraných proměnných. Cílem tohoto kroku nebylo plošné čištění dat ani automatické odstraňování všech neobvyklých pozorování, ale omezení vlivu krajních případů, které by mohly nepřiměřeně deformovat navazující pokročilou EDA, projekce a shlukování. Pro účely některých grafických výstupů byl navíc zachován ořez rozsahu zobrazení, například na úrovni 1. až 99. percentilu, aby byly výsledné vizualizace lépe čitelné. Na rozdíl od čistě vizuálních úprav však tato konzervativní filtrace extrémních hodnot představovala samostatný krok přípravy analytického datasetu, který byl následně použit i v dalších částech analýzy a modelování.

Zvláštní pozornost byla věnována také posouzení normality rozdělení vybraných numerických proměnných. Pro zvolený účel byly v notebooku využity QQ grafy a p-hodnoty testů normality, konkrétně pro proměnné *annual_inc*, *dti*, *int_rate* a *installment*. Jak ukazuje obrázek 18, u vybraných ekonomicky významných atributů nelze předpoklad normálního rozdělení považovat za samozřejmý. Výsledky této kontroly sloužily zejména jako podpůrný argument pro použití neparametrických metod při porovnávání skupin.

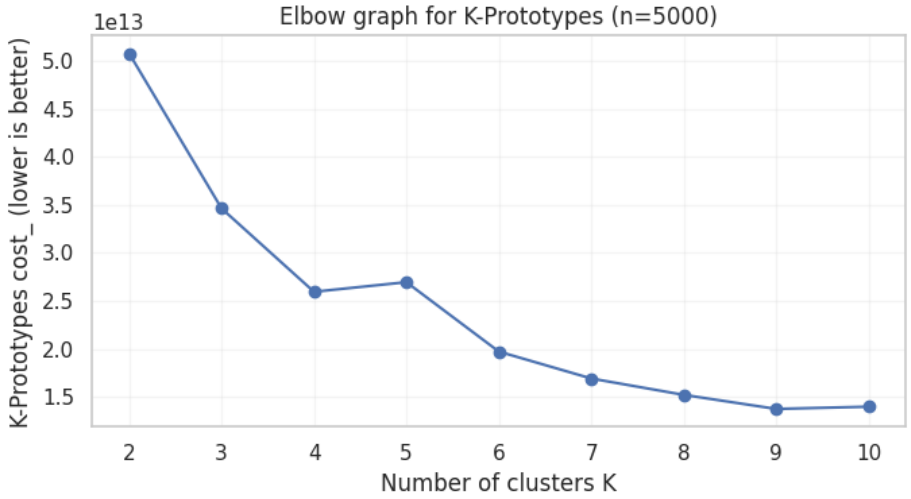


Obrázek 18: QQ graf ročního příjmu

zdroj: vlastní zpracování

QQ grafy ukazují, že vybrané numerické proměnné se od normálního rozdělení více či méně odchylují. Nejvýraznější odchylka je patrná u proměnných *annual_inc* a *installment*, které vykazují zřetelnou asymetrii a odchylku od referenční přímky zejména v krajních částech rozdělení. Také proměnné *dti* a *int_rate* nevykazují ideální normalitu, i když jejich průběh je v centrální části rozdělení bližší přímce. Výsledky tedy naznačují, že předpoklad normality není u těchto proměnných plně splněn, což podporuje použití neparametrických metod.

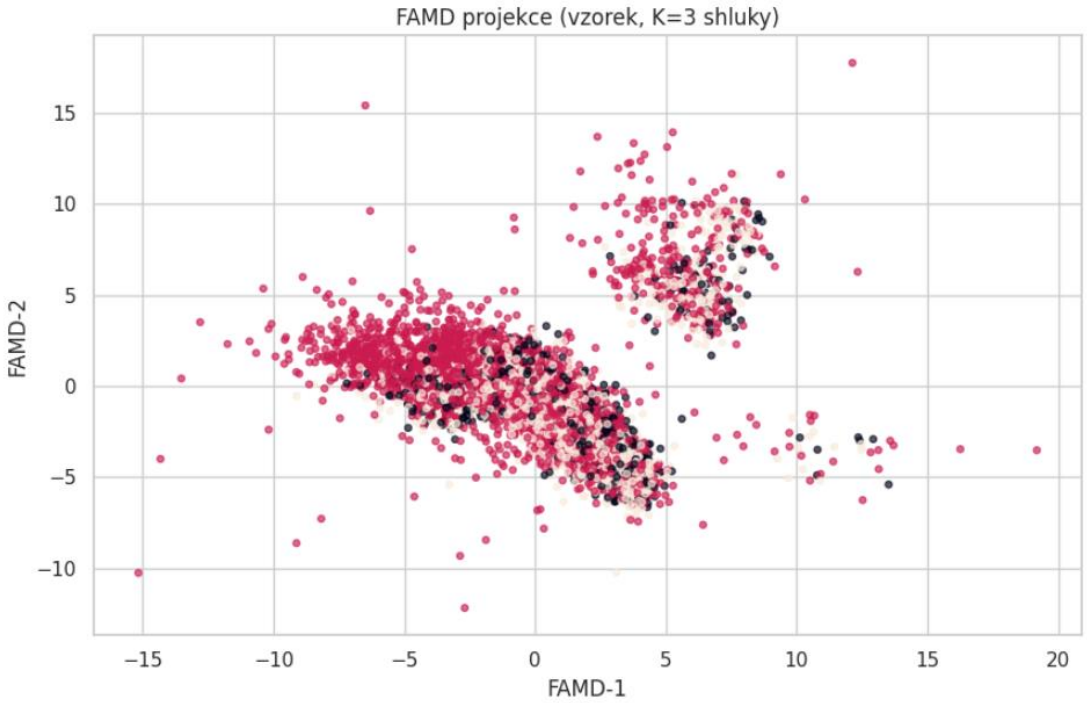
Vedle řízené explorace byla do analýzy zařazena také neřízená explorace dat smíšeného typu. Neřízená analýza znamená, že se v této části nevyužívá cílová proměnná pro přímou predikci, ale zkoumá se vnitřní struktura dat samotných. Jejím cílem tedy nebylo vytvořit další klasifikační model, ale získat lepší představu o tom, zda se v datech přirozeně objevují odlišné profily klientů. Pro redukci dimenze smíšených dat byla využita metoda FAMD, na kterou navazovala dvourozměrná vizualizace pomocí UMAP. Pro samotné shlukování byla na vzorku dat použita metoda K-Prototypes, doplněná orientačním elbow grafem a základní profilací vzniklých shluků. Postup a výslednou vizualizaci této části zachycují obrázky 19 až 22. Tato část analýzy slouží především jako podpůrný interpretační nástroj pro lepší pochopení struktury datového souboru *(McInnes a kol., 2018).*



Obrázek 19: Elbow graf pro volbu počtu shluků metodou K-Prototypes

zdroj: vlastní zpracování

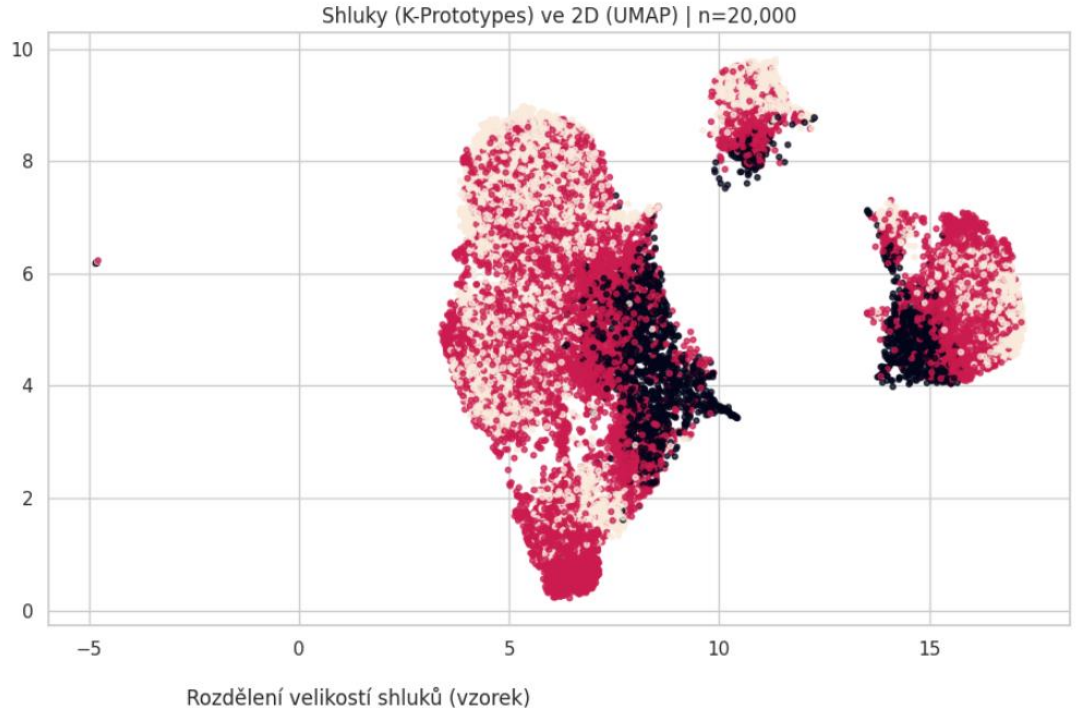
Obrázek 19 zachycuje průběh nákladové funkce metody K-Prototypes při různém počtu shluků. S rostoucím počtem shluků hodnota účelové funkce postupně klesá, což je očekávané, protože jemnější rozdělení dat vede k lepšímu přizpůsobení modelu. Z grafu je patrné, že nejvýraznější pokles nastává přibližně do hodnoty pěti shluků, zatímco při dalším navyšování počtu shluků se tempo zlepšení postupně zpomaluje. Popsaný výsledek podporuje volbu menšího počtu shluků jako rozumného kompromisu mezi jednoduchostí interpretace a zachycením vnitřní struktury dat.



Obrázek 20: Projekce smíšených dat pomocí FAMD

zdroj: vlastní zpracování

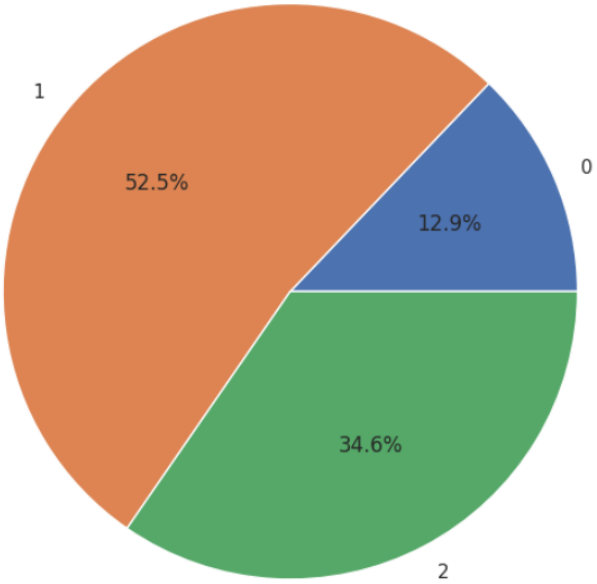
Obrázek 20 zobrazuje dvourozměrnou projekci smíšených dat pomocí metody FAMD, která umožňuje společně zachytit numerické i kategoriální proměnné v redukovaném prostoru. V projekci jsou patrné oblasti s vyšší koncentrací pozorování i postupné přechody mezi jednotlivými částmi prostoru. Rozložení bodů tak naznačuje, že datový soubor vykazuje vnitřní strukturu a určitou míru segmentace, současně však mezi částmi projekce přetrvává i překryv. Výsledek proto nepoukazuje na zcela jednoznačně oddělené skupiny, ale spíše na přítomnost dílčích segmentů s plynulejšími hranicemi. FAMD zde slouží především jako podpůrný nástroj pro orientační posouzení struktury smíšených dat a pro vizuální doplnění navazující shlukové analýzy.



Obrázek 21: Vizualizace shluků v projekci UMAP

zdroj: vlastní zpracování

Obrázek 21 zobrazuje shluky určené metodou K-Prototypes ve dvourozměrné projekci UMAP. Ve zjednodušeném dvourozměrném prostoru lze rozlišit tři hlavní oblasti odpovídající identifikovaným shlukům, ačkoli jejich hranice nejsou ve všech částech zcela ostré. Vizualizace tak podporuje závěr, že datový soubor obsahuje tři odlišné segmenty klientů.



Obrázek 22: Relativní velikost identifikovaných shluků

zdroj: vlastní zpracování

Obrázek 22 ukazuje relativní zastoupení jednotlivých shluků ve vzorku dat. Je patrné, že jeden shluk výrazně převažuje, zatímco některé další shluky jsou zastoupeny podstatně méně. To naznačuje, že struktura datového souboru není mezi skupiny rozložena rovnoměrně a že dominantní část záznamů sdílí podobnější charakteristiky než menší okrajové skupiny. Z hlediska interpretace je proto vhodné věnovat pozornost nejen samotné existenci shluků, ale také jejich velikosti a reprezentativnosti v rámci celého souboru.

Pomocí metody K-Prototypes byly na analyzovaném vzorku identifikovány tři shluky. Jejich interpretace byla následně doplněna základní profilací, která vycházela z porovnání mediánů vybraných numerických charakteristik a z určení nejčastějších kategorií u vybraných kategoriálních proměnných v jednotlivých skupinách. Na základě této profilace se ukázalo, že segmenty se liší především celkovým finančním profilem klientů. První segment zahrnuje klienty s vyšším příjmem, větší finanční rezervou a celkově silnějším ekonomickým zázemím. Druhý segment představuje spíše střední profil, který se ve většině sledovaných charakteristik pohybuje mezi krajními skupinami. Třetí segment naopak odpovídá klientům s nižším příjmem, slabším finančním zázemím a mírně méně příznivými znaky z hlediska zadlužení a ceny úvěru. Tyto shluky však nepředstavují přímo třídy GOOD a BAD, ale doplňkový popis odlišných klientských profilů v datovém souboru.

3.3.4 Modelování a vyhodnocení

Modelovací část byla navržena jako vícekrokový proces, jehož cílem bylo porovnat více klasifikačních přístupů a následně vybrat vhodné řešení pro úlohu úvěrového hodnocení. Do hlavního experimentu byly zařazeny tři modelové rodiny: logistická regrese, Random Forest a Gradient Boosting v implementaci *HistGradientBoostingClassifier*. Zvolená trojice reprezentuje odlišné přístupy k binární klasifikaci. Logistická regrese představuje jednodušší a dobře interpretovatelný lineární model, Random Forest robustní stromovou metodu založenou na kombinaci více rozhodovacích stromů a Gradient Boosting výkonný přístup schopný zachytit i složitější vztahy v datech.

3.4 Návrh experimentů

Experimentální část navazovala na referenční konfiguraci základní úlohy a rozvíjela ji pomocí vybraných variant, jejichž cílem bylo ověřit vliv konkrétních metodických rozhodnutí na dosažené výsledky. V souladu s charakterem řešené klasifikační úlohy a s reálnou implementací v notebooku byly zařazeny pouze experimenty relevantní pro použitý postup. Pozornost byla věnována zejména redukci příznaků, vyvážení tříd a doplňkovému posouzení vlivu velikosti trénovací množiny. Jednotlivé experimentální varianty byly vždy porovnávány se základní úlohou, aby bylo možné posoudit jejich skutečný přínos.

3.4.1 Experiment s redukovanou sadou příznaků

Pro experiment s redukovanou sadou příznaků byl v notebooku zvolen kombinovaný postup, který spojoval tři různé pohledy na relevanci proměnných: modelově řízený výběr pomocí RFECV, míru závislosti vyjádřenou pomocí Mutual Information a screeningový ukazatel Information Value, běžně používaný v úlohách kreditního skóringu *(scikit-learn developers, nedatováno g).* Dílčí pořadí byla následně sloučena do společného žebříčku s vyšším důrazem na RFECV a výsledný výběr byl dále omezen korelačním filtrem podle Spearmanovy korelace. Smyslem tohoto postupu nebylo určit jediný „správný“ ranking proměnných, ale navrhnout rozumnou a kompaktní sadu numerických příznaků pro redukovanou variantu modelu.

Cílem bylo vytvořit kompaktnější variantu modelu s přibližně deseti numerickými příznaky a ověřit, zda lze i při nižší dimenzi dat zachovat přijatelnou predikční schopnost. Důraz na numerické proměnné vycházel ze snahy navrhnout co nejjednodušší a nejlépe interpretovatelnou variantu modelu, která nebude závislá na rozšíření dimenze po kódování kategoriálních atributů.

Redukovaná větev byla v aktuální verzi notebooku realizována pomocí logistické regrese nad vybranou sadou přibližně deseti numerických proměnných. Oproti plné variantě zde nebyla použita imputace chybějících hodnot. Do modelování vstupovaly pouze záznamy s úplnými hodnotami ve všech vybraných příznacích. Experiment tak současně sloužil i jako praktické srovnání kompaktního modelu založeného na úplných datech s plnou variantou základní úlohy, která pracovala s imputací.

3.4.2 Experiment s vyvážením tříd

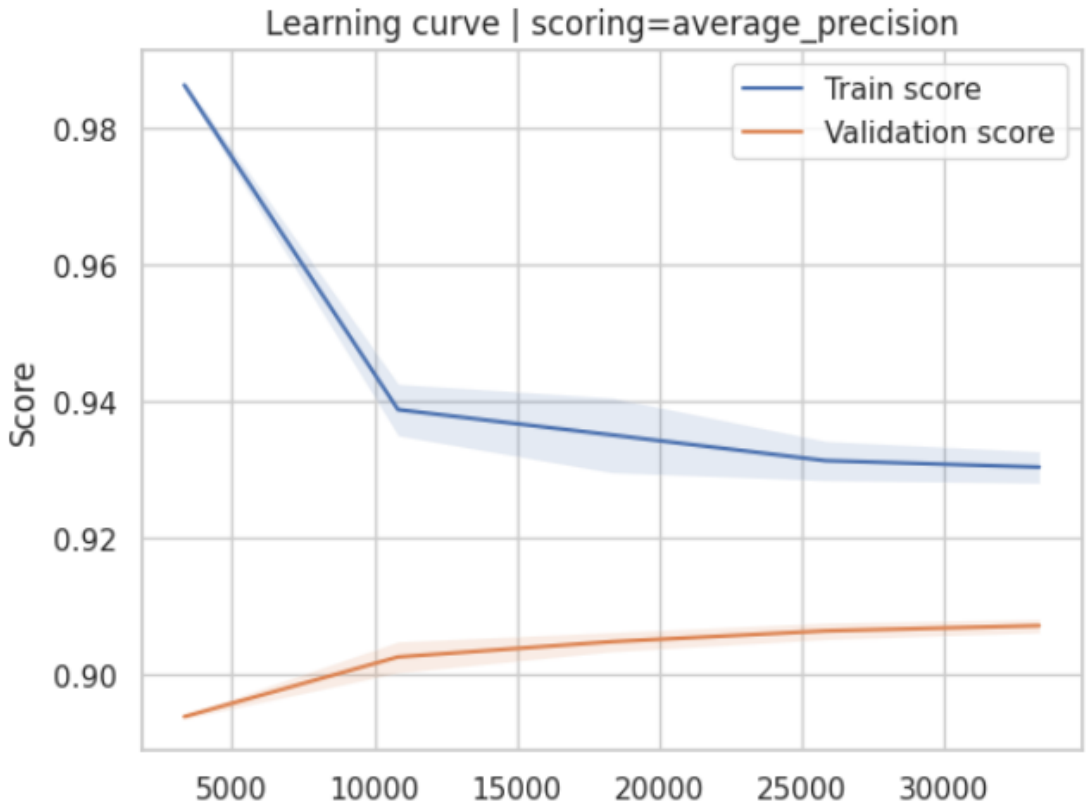
Další experiment byl zaměřen na problém nevyváženosti tříd v datovém souboru. Riziková třída je v datech zastoupena méně než třída neriziková, a proto byl v notebooku porovnáván výkon modelů trénovaných na původním rozložení tříd s výkonem modelů trénovaných na vyvážené trénovací sadě. Cílem bylo ověřit, zda úprava třídního zastoupení povede ke zlepšení záchytu rizikových klientů, zejména z hlediska recall a F1 skóre pro rizikovou třídu.

Notebook je připraven i pro více způsobů práce s nevyvážeností tříd, avšak v aktuálním standardním běhu a v prezentovaných výstupech je hlavní porovnání postaveno především na srovnání původního scénáře s variantou využívající **class_weight**. Interpretace finálního scénáře proto vychází právě z tohoto porovnání a z jeho provozních důsledků pro záchyt rizikové třídy. Zvolený přístup nemění fyzické rozložení trénovacích dat, ale zvyšuje penalizaci chyb u méně zastoupené rizikové třídy. Jeho výsledky jsou v notebooku následně porovnávány s referenčním scénářem bez dodatečného vážení.

3.4.3 Doplňkové posouzení vlivu velikosti trénovací množiny

Součástí experimentální části bylo také doplňkové posouzení vlivu velikosti trénovací množiny na chování preferovaného modelu. Místo samostatného srovnání několika pevných poměrů trénovací a testovací části byla v notebooku využita učící křivka, která sleduje vývoj výkonu modelu na trénovacích a validačních datech při postupně rostoucí velikosti trénovací množiny.

Smyslem kroku nebylo vytvořit další samostatnou modelovou větev, ale doplnit interpretaci výsledků o pohled na to, zda je výkon modelu omezen spíše mírou přeučení, nedostatečným naučením nebo objemem dostupných dat. Učící křivka byla počítána pro preferovaný naladěný model a sloužila jako diagnostický nástroj podporující interpretaci výsledků základní i vyvážené varianty.



Obrázek 23: Učící křivka modelu podle metriky Average Precision

zdroj: vlastní zpracování

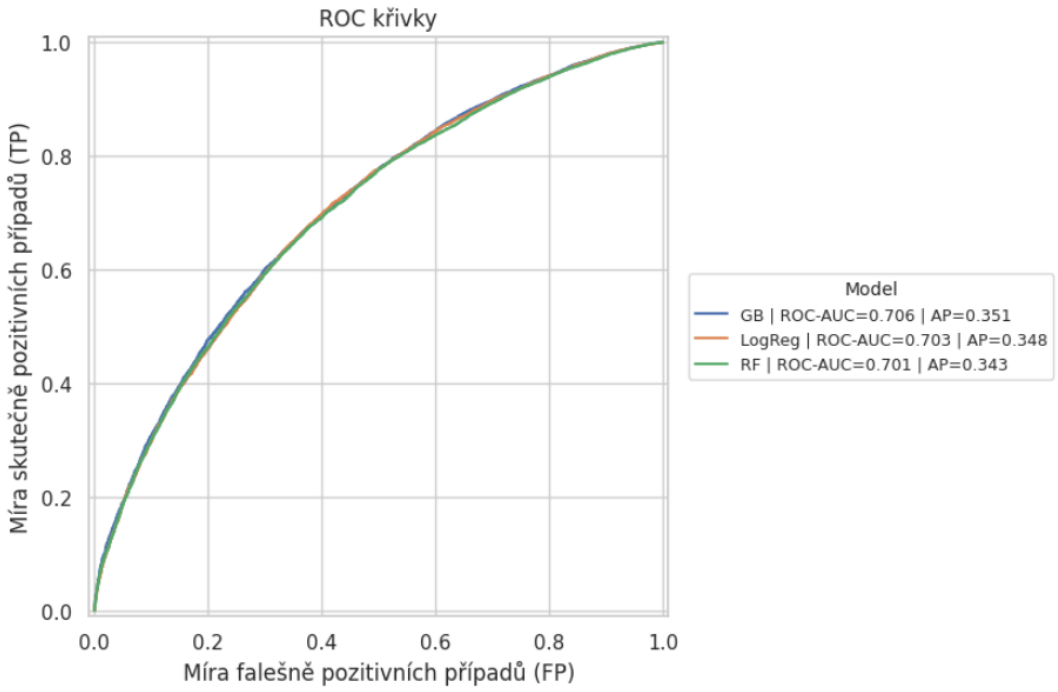
Obrázek 23 zachycuje učící křivku modelu vyhodnocenou pomocí metriky Average Precision. Hodnoty na ose y představují dosaženou úroveň této metriky, tedy kvalitu modelu z hlediska Precision–Recall charakteristiky, čím vyšší hodnota, tím lepší výkon modelu. Modrá křivka odpovídá výkonu na trénovacích datech a oranžová výkonu na validačních datech. S rostoucí velikostí trénovací množiny se rozdíl mezi oběma křivkami zmenšuje, což ukazuje na postupné omezení přeučení. Pokles trénovacího skóre je přitom přirozený, protože při malé trénovací množině se model snáze přizpůsobí konkrétním pozorováním, zatímco při větším a rozmanitějším vzorku více zobecňuje. Validační skóre se současně mírně zlepšuje, avšak později se již spíše stabilizuje. Výsledek proto naznačuje, že model z většího množství dat stále do určité míry těží, ale další zlepšení by již bylo spíše pozvolné.

4 Výsledky analýz

Uvedená kapitola shrnuje hlavní výsledky modelovací části notebooku. Pozornost je věnována zejména porovnání základních modelových rodin, ladění hyperparametrů, redukované variantě modelu, experimentu s vyvážením tříd a doplňkovým diagnostickým výstupům. Na rozdíl od metodické kapitoly se zde již neopakuje technický popis rozdělení dat ani konfigurace pipeline, ale jsou interpretovány dosažené výsledky a jejich význam pro úlohu hodnocení bonity.

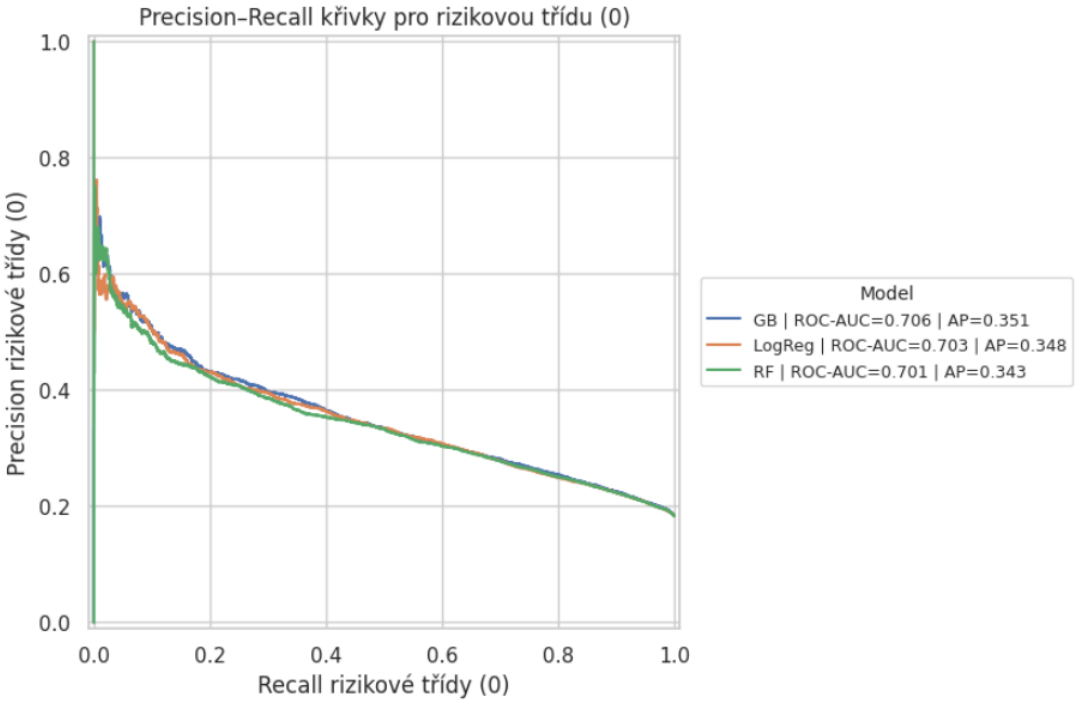
4.1 Porovnání základních modelů a ladění hyperparametrů

Průběžné porovnání modelových rodin v preferovaném notebookovém nastavení ukázalo, že logistická regrese, Random Forest i Gradient Boosting dosahují na validační sadě relativně podobných hodnot prahově nezávislých metrik, přičemž mírně nejlepšího výsledku dosáhl Gradient Boosting. Souhrnné srovnání těchto modelových rodin a jejich ROC/PR křivek zachycují obrázky 24 a 25, numerické hodnoty pak shrnuje tabulka 3. Rozdíly nejsou natolik výrazné, aby bylo možné některou z ostatních modelových rodin již v této fázi označit za jednoznačně nevhodnou. Současně je však vhodné zdůraznit, že dané srovnání představuje průběžný notebookový přehled modelových rodin, nikoli finální samostatné validační vyhodnocení výsledně zvoleného dotrénovaného modelu, které je uvedeno až v následující podkapitole.



Obrázek 24: Souhrnné porovnání naladěných modelových rodin pomocí ROC křivek

zdroj: vlastní zpracování



Obrázek 25: Souhrnné porovnání naladěných modelových rodin pomocí Precision–Recall křivek
zdroj: vlastní zpracování

Tabulka 3: Výkon modelových rodin na validační sadě

Model	Zdroj	Práh na P(y=1)	Accuracy	Precision (třída 1)	Recall (třída 1)	F1 (třída 1)	ROC-AUC	Average Precision (třída 0)
GB	naladěný	0,7950	0,6907	0,8866	0,7128	0,7903	0,7098	0,3490
LogReg	naladěný	0,8025	0,6731	0,8891	0,6855	0,7742	0,7079	0,3471
RF	naladěný	0,8000	0,6727	0,8884	0,6858	0,7741	0,7065	0,3450

zdroj: vlastní zpracování

Na dané základní srovnání navázalo ladění hyperparametrů pomocí RandomizedSearchCV, které bylo v notebooku použito jako výpočetně přiměřený způsob prohledání prostoru parametrů při zachování křížově validačního režimu. Použitý postup odpovídá běžné praxi při optimalizaci modelů, kdy není účelné vyčerpávajícím způsobem procházet všechny kombinace parametrů, ale spíše efektivně testovat omezený počet kandidátních konfigurací *(Yang a Shami, 2020).*

Ladění proběhlo na vzorku 30 000 záznamů z trénovací sady. Uvedené omezení bylo zvoleno z důvodu výpočetní náročnosti, aby bylo možné v prostředí Kaggle efektivně prověřit více kombinací hyperparametrů při zachování přijatelné doby výpočtu, aniž by došlo k úplné ztrátě reprezentativnosti trénovacích dat. Jako hlavní optimalizační metrika byla zvolena ROC-AUC, protože hodnotí schopnost modelu rozlišovat mezi třídami napříč různými rozhodovacími prahy a poskytuje tak vhodné prahově nezávislé kritérium pro ladění hyperparametrů. Použitá konfigurace je shrnuta v tabulce 4.

Tabulka 4: Použitá konfigurace ladění

Parametr	Hodnota	Význam
Počet CV dělení	3	Každá konfigurace byla ověřena ve 3 validačních bězích.
Vzorek pro ladění	30 000 záznamů	Ladění proběhlo na menším vzorku kvůli výpočetní náročnosti.
Počet testovaných konfigurací RF	6	Počet vyzkoušených kombinací hyperparametrů pro Random Forest.
Počet testovaných konfigurací LR	6	Počet vyzkoušených kombinací hyperparametrů pro logistickou regresi.
Počet testovaných konfigurací GB	4	Počet vyzkoušených kombinací hyperparametrů pro Gradient Boosting.
Optimalizační metrika	ROC-AUC	Kritérium, podle kterého byla vybrána nejlepší konfigurace.

zdroj: vlastní zpracování

Uvedený postup umožnil porovnat více kandidátních nastavení bez nutnosti vyčerpávajícího prohledávání všech kombinací parametrů. Zároveň byla validační sada ponechána až pro následné porovnání již naladěných variant modelů.

V rámci ladění pomocí **RandomizedSearchCV** dosáhl nejlepšího výsledku model Gradient Boosting s nejlepší hodnotou CV ROC-AUC 0,7061. V průběžném porovnání modelových rodin na validační sadě daná větev vykázala ROC-AUC 0,7098, zatímco finální samostatné validační vyhodnocení dotrénovaného naladěného modelu, popsané v následující podkapitole, vedlo k hodnotě ROC-AUC 0,7169. Je proto vhodné tyto hodnoty neztotožňovat: první odpovídá křížově validačnímu skóre v procesu ladění, druhá průběžnému validačnímu srovnání modelových rodin a třetí finálnímu validačnímu vyhodnocení výsledně zvoleného modelu. Takové nastavení naznačuje snahu nalézt vyvážený model, který si zachová dostatečnou predikční sílu, ale současně nebude zbytečně složitý a náchylný k přeučení. Hodnota *learning_rate* = 0,1 současně odpovídá spíše konzervativnějšímu průběhu učení, kdy se výsledný model skládá postupně a stabilně. Konkrétní nejlepší nalezené nastavení tohoto modelu shrnuje tabulka 5.

Tabulka 5: Nejlepší nalezená konfigurace modelu GB

Parametr / metrika	Hodnota	Význam v modelu
max_depth	5	Maximální hloubka jednotlivých stromů. Nižší hodnota omezuje složitost stromů a snižuje riziko přeučení.
max_leaf_nodes	31	Maximální počet koncových uzlů listů v jednom stromu. Parametr určuje, jak jemně může strom rozdělovat prostor dat.
min_samples_leaf	50	Minimální počet pozorování v koncovém uzlu. Vyšší hodnota vede k robustnějším, ale méně detailním pravidlům a pomáhá omezit přeučení.
learning_rate	0,1	Rychlost učení, tedy míra, s jakou každý další strom přispívá k výsledné predikci. Nižší hodnota znamená opatrnější a stabilnější učení modelu.

zdroj: vlastní zpracování

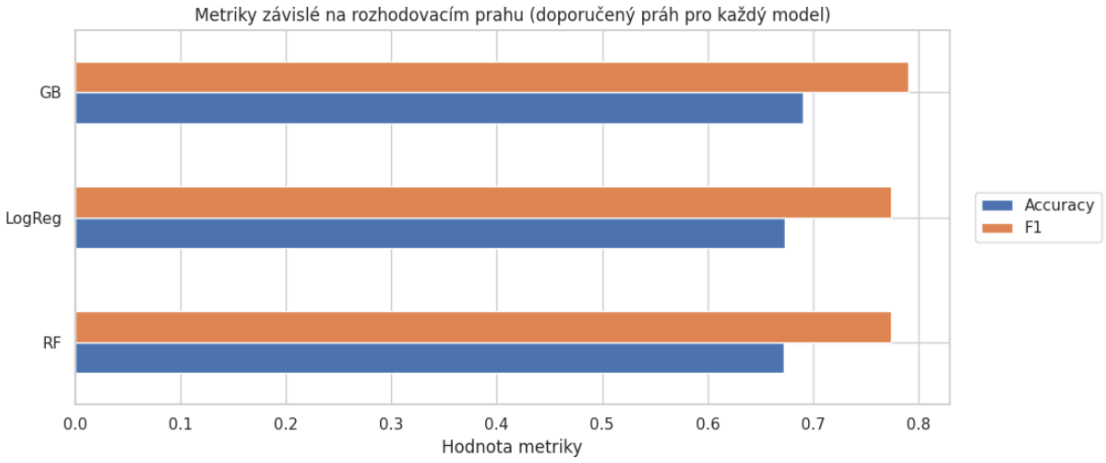
Ladění hyperparametrů bylo provedeno také u logistické regrese a Random Forestu, avšak ani jejich nejlepší nalezené konfigurace nepřekonaly naladěný Gradient Boosting na validační sadě, a proto je dále detailněji interpretován především tento model.

4.2 Finální model a vyhodnocení na validační sadě

Na základě základního porovnání i následného ladění hyperparametrů byl jako nejsilnější kandidát vybrán model Gradient Boosting. Na validační sadě potvrzoval nejlepší prahově nezávislé metriky a představoval proto přirozený referenční bod pro další experimentální větve. Výsledek současně ukázal, že použité úvěrové údaje obsahují nelineární vztahy, které uvedený typ modelu dokáže využít lépe než jednodušší lineární přístup.

Při doporučeném rozhodovacím prahu 0,795 pro nerizikovou třídu (1), zvoleném na validační sadě podle F1 skóre rizikové třídy (0), dosáhl naladěný Gradient Boosting hodnot **precision_0 = 0,3176**, **recall_0 = 0,6083**, **F1_0 = 0,4173** a **balanced accuracy = 0,6581**, současně **ROC-AUC činilo 0,7169**. Souhrnné porovnání accuracy a F1 při doporučeném rozhodovacím prahu na validační sadě zachycuje obrázek 26.

Ve srovnání s konzervativnějším provozním bodem zvoleným při vyšším důrazu na většinovou třídu již toto nastavení nepřehlíží převážnou část rizikových případů, ale zachytí přibližně 61 % skutečně rizikových žadatelů při stále relativně nízké precision rizikové třídy. Již na validační sadě se tak potvrdilo, že vedle ROC-AUC je nutné sledovat také matici záměn a metriky přímo vztažené k rizikové třídě.



Obrázek 26: Porovnání Accuracy a F1 při doporučeném prahu na validační sadě
zdroj: vlastní zpracování

4.3 Redukce příznaků (10 proměnných)

Vedle plného modelu byla vytvořena také redukovaná varianta založená na menší a lépe interpretovatelné sadě příznaků. Výběr proměnných vycházel z kombinovaného rankingu založeného na metodách RFECV, Mutual Information a Information Value, přičemž po kontrole redundancy byla ponechána sada deseti klíčových numerických atributů: *int_rate, dti, installment_to_monthly_income, revol_util, annual_inc, revol_bal, tot_cur_bal, loan_amnt, total_acc a total_rev_hi_lim*.

Výsledek ukázal, že významná část predikční informace je soustředěna v relativně malé skupině ekonomicky smysluplných ukazatelů. Vedle úrokové sazby, zadluženosti vůči příjmu a proměnných spojených s revolvingovým dluhem se do finální desítky dostal také odvozený ukazatel *installment_to_monthly_income*. Právě tento odvozený parametr zachycuje zatížení

měsíčního příjmu pravidelnou splátkou a potvrzuje, že pro odhad rizikovosti nebyly podstatné pouze absolutní charakteristiky úvěru, ale i jejich vzájemné poměry. Přehled finální redukované sady deseti vybraných příznaků uvádí tabulka 6.

Tabulka 6: Finální redukovaná sada deseti numerických příznaků

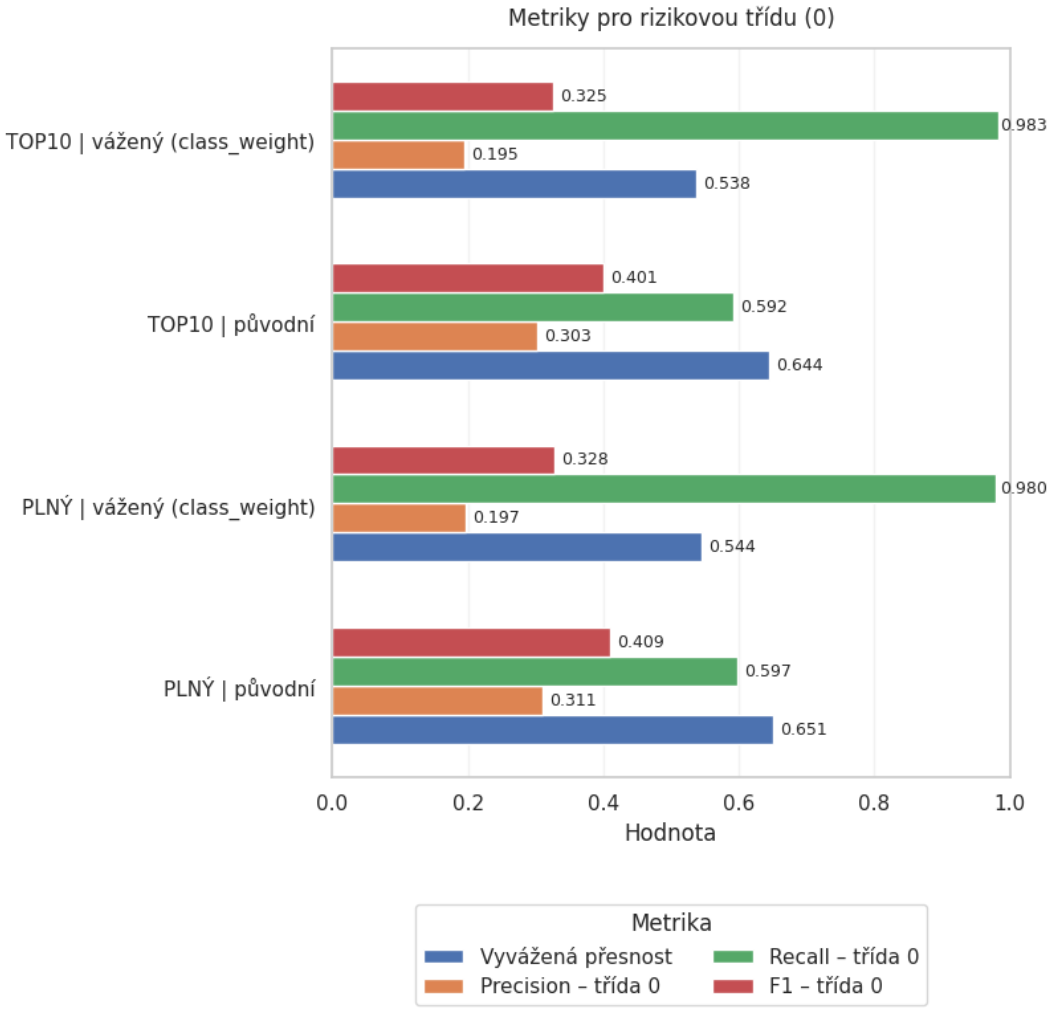
Pořadí	Proměnná	Pořadí	Proměnná
1	int_rate	6	revol_bal
2	dti	7	tot_cur_bal
3	installment_to_monthly_income	8	loan_amnt
4	revol_util	9	total_acc
5	annual_inc	10	total_rev_hi_lim

zdroj: vlastní zpracování

Redukovaná varianta dosáhla jen mírně slabších výsledků než plně naladěný model. Ve srovnání s plnou variantou došlo k poklesu přibližně o 2,1 % u ROC-AUC a o 1,0 % u Average Precision. Výsledek proto podporuje závěr, že podstatnou část rozhodovací informace lze zachytit i kompaktnějším modelem. Současně je však třeba zohlednit, že redukovaná větev byla vyhodnocena pouze na části validační sady po odstranění řádků s chybějícími hodnotami, a její výsledky proto nejsou plně přímo srovnatelné s plnou variantou.

4.4 Experiment vyvážení tříd

Samostatná experimentální větev sledovala, jak změna práce s nevyvážeností tříd ovlivní chování preferovaného modelu Gradient Boosting při zachování plné sady příznaků i redukované varianty TOP10. Ve výchozím nastavení této experimentální sekce byly scénáře na validační sadě porovnávány při prahu 0,80. U plného modelu vedla vážená varianta k velmi výraznému nárůstu recallu rizikové třídy z 0,5969 na 0,9803, tedy k téměř úplnému záchytu rizikových případů. Uplatněný posun byl však vykoupen poklesem precision z 0,3110 na 0,1971 a F1_0 z 0,4089 na 0,3282, zatímco vyvážená přesnost klesla z 0,6507 na 0,5441. Výsledky na obrázku 27 tak neukazují obecně lepší model, ale zřetelný posun k citlivějšímu provoznímu nastavení, které upřednostňuje maximální záchyt rizikové třídy i za cenu vyššího počtu falešně označených žadatelů.



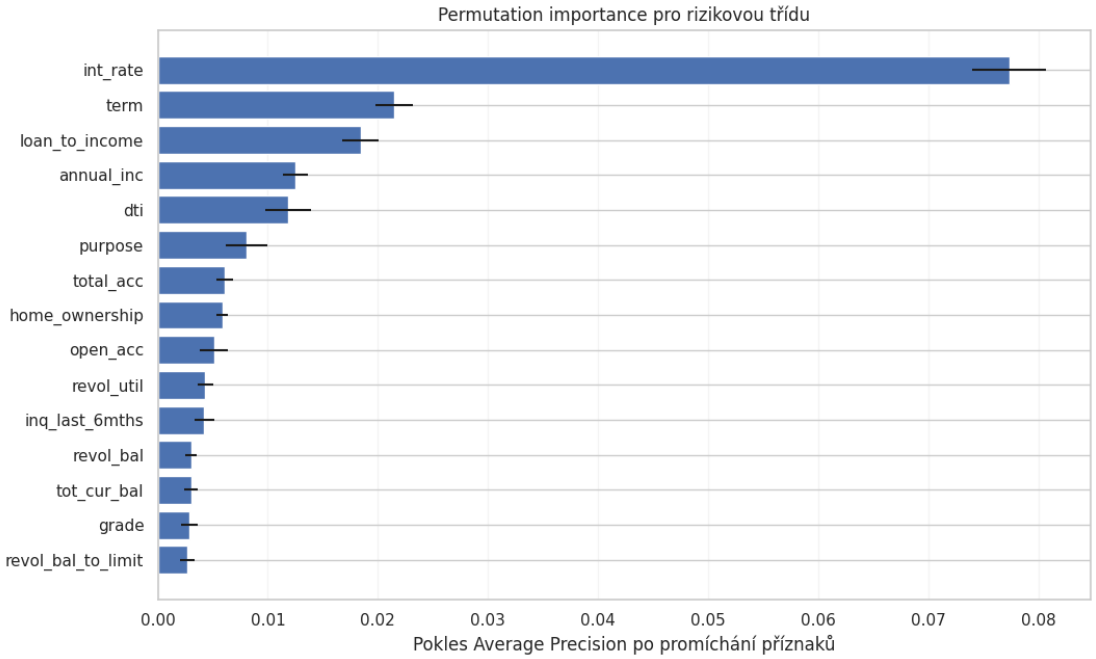
Obrázek 27: Metriky rizikové třídy při experimentu vyvážení tříd pomocí class_weight

zdroj: vlastní zpracování

4.5 Doplňující notebookové vizualizace

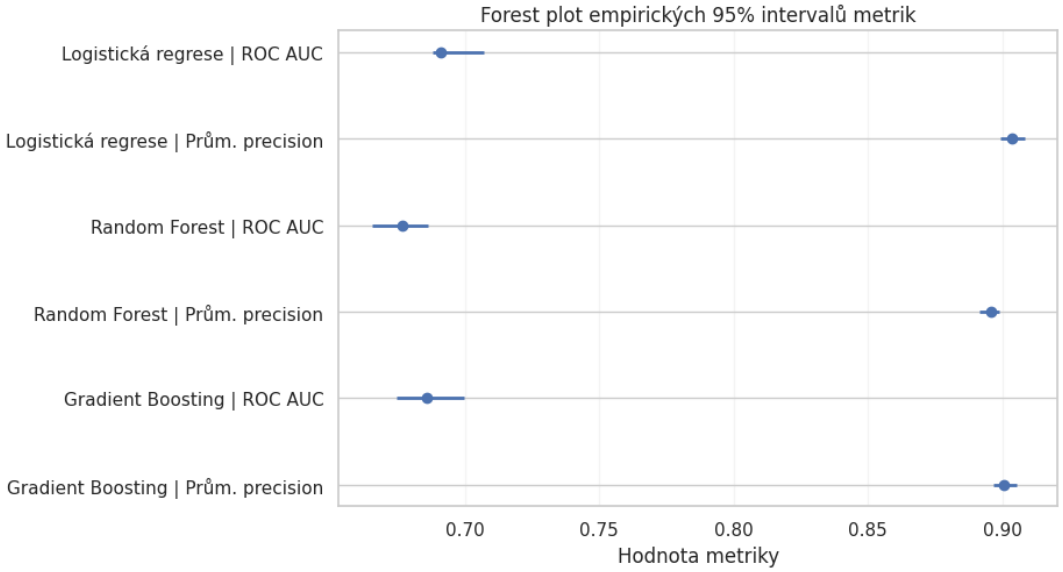
Vedle hlavních výkonnostních tabulek a validačních reportů byl notebook doplněn také o vizualizační výstupy podporující interpretaci. Souhrnné histogramy a boxploty vybraných numerických proměnných podle cílové třídy ukázaly, že mezi rizikovými a nerizikovými klienty se opakovaně liší zejména úroková sazba, zadluženost vůči příjmu, výše revolvingového dluhu a některé další ukazatele finanční zátěže. Tyto vizualizace tak doplnily modelové metriky o srozumitelnější pohled na praktickou podobu rozdílů mezi oběma třídami.

Interpretační část byla dále rozšířena o permutation importance finálního váženého modelu. Pořadí nejdůležitějších proměnných podle permutation importance zobrazuje obrázek 28. Největší vliv na výsledné rozhodování má proměnná *int_rate*, s větším odstupem následují term, *loan_to_income*, *annual_inc* a *dti*. Vedle ceny úvěru a délky splácení se tak mezi důležitými příznaky objevuje i odvozený ukazatel *loan_to_income*, což podporuje závěr, že model reaguje nejen na absolutní hodnoty úvěru a příjmu, ale i na jejich vzájemný poměr. Doplňkové intervalové vyhodnocení pomocí forest plotu pak slouží jako orientační kontrola stability vybraných metrik napříč opakovanou křížovou validací.



Obrázek 28: Permutation importance finálního modelu pro rizikovou třídu

zdroj: vlastní zpracování



Obrázek 29: Forest plot empirických intervalů vybraných metrik

zdroj: vlastní zpracování

Forest plot na obrázku 29 doplňuje bodové výsledky o informaci, jak citlivé jsou jednotlivé modely na změnu rozdělení dat v opakované křížové validaci. Je patrné, že všechny tři modelové rodiny vykazují relativně stabilní chování, protože intervaly variability nejsou široké. Logistická regrese dosahuje nejvyšších průměrných hodnot sledovaných metrik, Gradient Boosting se jí však velmi přibližuje a jejich intervaly se částečně překrývají. To znamená, že z hlediska stability a úrovně výkonu jsou si tyto dvě varianty poměrně blízké. Random Forest vychází v porovnání s nimi slaběji, zejména u ROC-AUC. Forest plot tedy podporuje závěr, že mezi hlavními kandidáty nevznikl dramatický výkonnostní rozdíl a že volba finálního modelu musí vycházet i z dalších hledisek, nejen z tohoto intervalového srovnání.

5 Diskuse výsledků

Dosažené výsledky ukazují, že hodnocení bonity žadatelů o úvěr je na použitých tabulkových datech řešitelné pomocí standardních metod strojového učení. Současně však potvrzují, že pro tuto úlohu nestačí porovnávat modely pouze podle jedné souhrnné metriky. Z metodického hlediska se jako nejdůležitější ukázalo rozlišovat mezi schopností modelu klienty správně seřadit podle rizikovosti a jeho chováním při konkrétním rozhodovacím prahu.

Nejlepších prahově nezávislých výsledků dosáhl naladěný Gradient Boosting, což z něj činí nejsilnějšího kandidáta z hlediska obecné diskriminační schopnosti. Jako finální provozní scénář byl nakonec zvolen plný model s vážením tříd pomocí parametru *class_weight* ne proto, že by nabízel nejlepší celkový kompromis mezi precision a recallem, ale proto, že téměř maximalizuje záchyt rizikové třídy při zachování srovnatelných prahově nezávislých metrik.

5.1 Volba prahu a interpretace metrik

Jedním z hlavních zjištění práce je, že vysoká hodnota accuracy sama o sobě není pro úlohu úvěrového rizika dostatečným kritériem. U naladěného Gradient Boostingu byl na validační sadě jako doporučený provozní bod vybrán práh 0,795 pro třídu 1, a to podle F1 skóre rizikové třídy (0). Při tomto prahu model dosáhl recall u rizikové třídy přibližně 0,6083 a F1_0 = 0,4173. Rozhodovací práh tak výrazně mění praktické chování modelu a potvrzuje, že samotné prahově nezávislé metriky nestačí k posouzení provozní použitelnosti modelu.

Výsledek potvrzuje, že mezi kvalitním pořadím klientů podle rizikovosti a praktickou použitelností modelu existuje podstatný rozdíl. ROC-AUC a Average Precision ukazují, že model umí klienty rozumně rozlišit, avšak samy o sobě neříkají, zda bude při konkrétním prahu dostatečně zachycovat rizikovou třídu. Pro úvěrové rozhodování je proto nezbytné doplňovat souhrnné metriky o matici záměn a o metriky precision, recall a F1 ve vztahu k rizikovým případům.

Rozhodovací práh je v tomto kontextu vhodné chápat jako řídicí parametr, ne jako pevně danou technickou konstantu. Jeho změnou lze posouvat kompromis mezi záchytem rizikových klientů a počtem falešných poplachů. Praktická volba prahu by proto měla vycházet z cíle instituce, tolerance k jednotlivým typům chyb a ideálně i z jejich ekonomických dopadů, ne pouze z maximalizace jedné agregované metriky.

Srovnání s dosavadními studiemi ukazuje, že vlastní výsledky jsou s odbornou literaturou v zásadě konzistentní. Také v této práci se ukázalo, že stromově orientované a boostingové metody dosahují na tabulkových úvěrových datech velmi dobrých výsledků, přičemž naladěný Gradient Boosting vyšel nejlépe v prahově nezávislých metrikách na validační sadě. Tím se výsledek přibližuje závěrům Brown a Mues *(2012),* kteří rovněž identifikovali random forest a Gradient Boosting jako velmi silné kandidáty pro úlohy kreditního skórování, a obecně zapadá i do širšího srovnání Lessmann a kol. *(2015*), podle něhož mohou pokročilejší modely překonávat logistickou regresi.

Současně je však třeba vysvětlit, proč není rozdíl mezi modely v této práci výrazný. V porovnání se studií Lessmann a kol*. (2015)* je zde použit pouze jeden veřejný dataset namísto více reálných datových zdrojů, menší množina modelových rodin a jednodušší experimentální prostor bez heterogenních ensemble metod. Výsledky navíc ovlivňuje i odlišná definice cílové proměnné, rozsah odstranění potenciálně leakujících proměnných, volba metrik a skutečnost, že práce klade důraz nejen na ROC-AUC, ale také na chování modelu při konkrétním rozhodovacím prahu. Proto je logické, že se zde výhoda Gradient Boostingu projevuje spíše jako mírný náskok než jako dramatická dominance.

Orientační srovnání s veřejně dostupnými notebooky na platformě Kaggle ukazuje podobný směr závěrů, i když ne zcela stejnou velikost rozdílů mezi modely. Například veřejný notebook Efficient grading for Lending Club with CatBoost pracuje s detailními daty Lending Club z let 2007 až Q4 2018 a uvádí, že model CatBoost vychází v metrice ROC-AUC mírně lépe než logistická regrese *(karthur10, nedatováno).* Uvedený výsledek je směrově konzistentní i s předloženou prací, v níž naladěný Gradient Boosting dosáhl nejlepších prahově nezávislých metrik. Přímé srovnání absolutních hodnot však není vhodné, protože veřejné Kaggle notebooky často používají odlišnou definici cílové proměnné, jiné filtrování stavů úvěru, odlišný rozsah předzpracování i jiný přístup k proměnným s rizikem úniku informací.

Odlišnost lze pozorovat i v experimentu s vyvážením tříd. Brown a Mues *(2012)* sledovali především robustnost metod při rostoucí třídní nerovnováze, zatímco daná práce ukazuje praktický provozní dopad konkrétního nastavení class_weight na záchyt rizikové třídy. Výsledkem zde nebylo dosažení nejlepšího celkového kompromisu metrik, ale záměrný posun ke scénáři orientovanému na velmi vysoký záchyt rizikové třídy. To je z praktického hlediska důležité, protože v kreditním rozhodování nemusí být cílem maximalizace jedné souhrnné metriky, ale řízený kompromis mezi záchytem rizikových klientů a počtem falešných poplachů.

5.2 Vyvážení tříd

Dalším podstatným zjištěním je silný vliv nevyváženosti tříd na chování modelu. V původních datech výrazně převažuje třída 1 (good), takže model má tendenci preferovat většinovou třídu. V aktuální experimentální větvi proto nebylo použito podvzorkování, ale vážení tříd pomocí parametru class_weight, které zvyšuje penalizaci chyb u rizikové třídy již během učení modelu. Výsledek je v souladu s obecným poznatkem, že nevyváženost tříd výrazně ovlivňuje chování klasifikačního modelu a může měnit kompromis mezi záchytem menšinové třídy a počtem chybných klasifikací. Současná literatura přitom ukazuje, že nevyváženost tříd může zhoršovat také stabilitu modelových vysvětlení, protože interpretace generované metodami LIME a SHAP jsou při rostoucí třídní nerovnováze méně stabilní *(Chen a kol., 2024).*

Výsledky ukázaly, že class_weight vede u plného modelu k téměř maximálnímu recallu rizikové třídy (0,9803 na validaci a 0,9794 na testu), současně však výrazně snižuje precision této třídy a zhoršuje balanced accuracy i F1_0 oproti nevážené variantě. Vyvažování tříd zde proto nepřineslo lepší celkový kompromis, ale záměrně velmi citlivé nastavení preferující záchyt rizikových případů i za cenu vysokého počtu falešně zamítnutých nerizikových žadatelů.

5.3 Redukce příznaků a praktická použitelnost

Experiment s redukcí příznaků ukázal, že významná část predikční informace je soustředěna v relativně malé sadě proměnných. Mezi nejdůležitější atributy patřily zejména úroková sazba, ukazatel zadlužení vůči příjmu, roční příjem a proměnné související s revolvingovým dluhem a celkovou finanční zátěží klienta. Tyto faktory jsou zároveň ekonomicky i věcně dobře interpretovatelné, což zvyšuje důvěryhodnost výsledného řešení. To odpovídá i širší literatuře k vysvětlitelné AI ve financích, která zdůrazňuje, že prakticky využitelný model nemá být posuzován pouze podle výkonu, ale i podle srozumitelnosti rozhodovacích faktorů a možnosti obhájit výstupy v regulovaném prostředí *(Černevičienė a Kabašinskas, 2024).*

Redukovaný model s přibližně deseti proměnnými dosáhl nižší výkonnosti než plná laděná varianta, avšak pokles nebyl natolik výrazný, aby tuto variantu zcela diskvalifikoval. Výsledek naznačuje, že pro základní odhad rizikovosti není vždy nutné využívat velmi širokou množinu vstupních atributů a že jednodušší model může v některých situacích nabídnout přijatelný kompromis mezi výkonem, nároky na data a srozumitelností.

Současně je však třeba zdůraznit, že redukovaná varianta byla vyhodnocena na zúženém souboru po odfiltrování řádků s chybějícími hodnotami, a její výsledky proto nejsou zcela přímo srovnatelné s plným modelem hodnoceným na celé validační sadě. Přesto jde o důležitý experiment, protože ukazuje možnost navrhnout metodicky jednodušší a prakticky lépe nastavitelnou alternativu.

Z hlediska možného nasazení se proto redukovaná varianta jeví jako vhodná například pro předběžný screening klientů, pro situace s omezenou dostupností dat nebo jako jednodušší referenční model. Plný model nabízí vyšší predikční výkon, zatímco redukovaný model lépe ukazuje, že podstata rozhodování je založena na poměrně úzké skupině ekonomicky smysluplných ukazatelů.

Doplňková neřízená analýza navíc naznačila, že datový soubor nepůsobí jako zcela homogenní soubor klientů, ale obsahuje několik částečně se překrývajících profilů. Shluková analýza sice nebyla využita jako přímý predikční nástroj, přesto poskytla užitečný interpretační rámec pro pochopení toho, že úvěrové riziko je výsledkem kombinace více finančních a kategoriálních charakteristik, ne jediné dominantní proměnné.

5.4 Omezení práce

Hlavním omezením práce je použití veřejně dostupného datového souboru z platformy Kaggle. Přestože jeho struktura odpovídá typickým úvěrovým datům, nemusí plně zachycovat reálný rozhodovací proces konkrétní finanční instituce. V akademickém prostředí je navíc přístup k reálným bankovním datům výrazně omezen požadavky na ochranu osobních údajů, nutností anonymizace a interními pravidly poskytovatele. Dosažené výsledky je proto vhodné chápat především jako metodické ověření navrženého pracovního postupu, nikoli jako hotový model určený k přímému nasazení v praxi.

Určité omezení představuje také experimentální nastavení práce. Některé výpočetně náročnější kroky byly realizovány pouze na vzorcích dat, což mohlo částečně ovlivnit stabilitu dílčích výsledků. Současně se v experimentu s vyvážením tříd ukázalo, že vyšší záchyt rizikové třídy je

vykoupen zhoršením jiných metrik. Výsledky tak potvrzují, že v dané úloze neexistuje jedno univerzálně nejlepší nastavení nezávislé na účelu použití, ale vždy je nutné zvažovat konkrétní kompromis mezi typy klasifikačních chyb.

Další omezení spočívá v tom, že do rozhodování nebyly explicitně zahrnuty ekonomické náklady jednotlivých typů chyb. Model je sice vyhodnocován pomocí více klasifikačních metrik, skutečné úvěrové rozhodování však bývá zpravidla založeno na nákladově citlivém přístupu. Pro jeho přesnější zahrnutí by však bylo nutné pracovat s reálnými interními daty finanční instituce a s informacemi o ekonomických dopadech jednotlivých rozhodnutí, které nejsou ve veřejně dostupném datasetu k dispozici. Bez této znalosti nelze rozhodovací práh věrohodně navrhnout podle skutečných finančních nákladů a přínosů. Pro praktičtější využití by proto bylo vhodné doplnit ekonomickou interpretaci rozhodovacího prahu, externí validaci na datech z jiného zdroje a případně také časovou validaci, která by ověřila stabilitu modelu při změnách dat v čase.

6 Závěr

Závěrečná kapitola shrnuje hlavní přínosy práce, explicitně hodnotí míru naplnění stanovených cílů a vymezuje oblasti, které přesahují rozsah bakalářské práce a představují vhodný prostor pro další navazující výzkum.

6.1 Shrnutí hlavních přínosů

Bakalářská práce aplikovala životní cyklus datové vědy na úlohu hodnocení bonity žadatelů o úvěr a vedla k návrhu reprodukovatelného klasifikačního řešení nad veřejně dostupnými úvěrovými daty. Praktická část byla zpracována formou interaktivního notebooku, který pokrývá přípravu dat, explorativní analýzu, modelování, vyhodnocení i interpretaci výsledků. Součástí finální verze je také formulářová predikce pro jednotlivého žadatele a dávková predikce ze souboru CSV, takže práce nepřináší pouze popis experimentů, ale i jednoduchý demonstrační nástroj.

Za podstatný metodický přínos lze považovat důraz na reprodukovatelnost a prevenci úniku informací. Kroky předzpracování, jako jsou imputace, kódování a škálování, byly realizovány v rámci zřetězeného zpracování (pipeline) a učily se pouze na trénovací části dat. Postup byl veden iterativně podle principů RAD (Rapid Application Development), což umožnilo průběžně upravovat experimentální nastavení a zároveň zachovat návaznost jednotlivých analytických kroků.

Z hlediska výsledků bylo prokázáno, že vybrané modelové rodiny poskytují na daných datech použitelnou predikční schopnost, avšak jejich interpretace závisí na zvoleném způsobu vyhodnocení. Nejlepších prahově nezávislých výsledků na validační sadě dosáhl naladěný Gradient Boosting (ROC-AUC = 0,7169). Jako finální provozní scénář byl poté zvolen plný model Gradient Boosting s vážením tříd pomocí parametru class_weight, vyhodnocený na testovací sadě při prahu 0,795 pro třídu 1. Zvolený scénář dosáhl precision_0 = 0,1979, recall_0 = 0,9794, F1_0 = 0,3292 a balanced accuracy = 0,5461. Současně byla ověřena i redukovaná varianta s deseti klíčovými příznaky, která při nižší dimenzi zachovala použitelný výkon.

6.2 Míra naplnění cílů

Hlavní cíl práce byl naplněn. Spočíval v aplikaci metod datové vědy na hodnocení bonity žadatelů o úvěr a v návrhu prediktivního klasifikačního řešení. Práce pokryla celý základní pracovní postup od převzetí a úpravy dat přes definici cílové proměnné, explorativní analýzu, návrh experimentů a porovnání modelů až po výběr finální konfigurace a její vyhodnocení.

Naplněn byl také explicitně formulovaný problém práce: byl navržen a ověřen reprodukovatelný postup datové vědy, který na veřejně dostupných úvěrových datech rozlišuje rizikové a nerizikové žadatele, umožňuje transparentně diskutovat kompromis mezi záchytem rizikové třídy a celkovou výkonností a současně ukazuje, které charakteristiky žadatelů mají na výsledek největší vliv. Výsledkem není hotový produkční skóringový systém, ale metodicky obhajitelné a prakticky demonstrované řešení odpovídající rozsahu bakalářské práce. Finální scénář je vhodné chápat jako provozní variantu orientovanou na vysoký záchyt rizikových případů, ne jako univerzálně nejlepší kompromis všech metrik.

Dílčí cíle byly naplněny rovněž v rovině interpretace a praktické použitelnosti. Explorativní analýza umožnila identifikovat klíčové proměnné související s rizikem selhání a zdůvodnit volbu metod s ohledem na charakter dat, například práci s nevyvážeností tříd, volbu vhodných metrik a využití neparametrických přístupů při porovnávání skupin. Doplňkově byla zařazena i neřízená analýza smíšených dat a redukovaná sada příznaků, která ukázala možnost zachovat použitelnou predikční schopnost při nižších nárocích na vstupní informace.

Použitý datový soubor pochází z veřejně dostupné platformy Kaggle. Jeho struktura odpovídá typickým atributům úvěrových dat, avšak nemusí plně reflektovat specifika reálného rozhodovacího procesu konkrétní instituce. Výsledky je proto vhodné interpretovat především jako metodické a demonstrační ověření navrženého pracovního postupu, ne jako hotové řešení připravené k okamžitému produkčnímu nasazení bez další validace.

6.3 Doporučení pro další výzkum

Prostor pro navazující výzkum spočívá především v externí validaci navrženého řešení na odlišném zdroji dat. Vhodné by bylo ověřit model na jiném úvěrovém datovém souboru, na časově novějších datech nebo na datech pocházejících z jiné instituce. Takový krok by umožnil lépe posoudit generalizační schopnost modelu a ověřit, do jaké míry jsou dosažené výsledky přenositelné i mimo prostředí jednoho konkrétního datového zdroje.

Vedle samotné externí validace bude v navazujícím výzkumu důležité také systematičtější ukotvení vysvětlitelnosti, protože novější přehledové práce ukazují, že v kreditním hodnocení roste význam transparentnosti, důvěry a standardizovaného hodnocení kvality vysvětlení *(Cil a Yildiz, 2025).*

Další vhodné rozšíření představuje doplnění explicitního ekonomického modelu nákladů chybné klasifikace. V prostředí kreditního skóringu totiž nejsou falešně pozitivní a falešně negativní rozhodnutí stejně závažná. Schválení rizikového klienta může vést k přímé finanční ztrátě, zatímco zamítnutí bonitního klienta znamená spíše ušlý zisk a ztrátu obchodní příležitosti. V budoucí práci by proto bylo účelné volit rozhodovací práh nejen podle statistických metrik, ale také podle očekávaných finančních dopadů jednotlivých typů chyb.

Dalším vhodným směrem navazujícího výzkumu je rozšíření srovnání o moderní implementace Gradient boostingu. Vedle již použitého modelu *HistGradientBoostingClassifier*, který je podle dokumentace scikit-learn výrazně rychlejší než klasický *GradientBoostingClassifier* u větších datových souborů, by bylo vhodné zařadit také frameworky *XGBoost*, *LightGBM* a *CatBoost (scikit-learn developers, nedatováno e*). *XGBoost* je v oficiální dokumentaci popsán jako optimalizovaná distribuovaná knihovna Gradient Boostingu navržená s důrazem na efektivitu, flexibilitu a přenositelnost *(XGBoost developers, nedatováno). LightGBM* je prezentován jako efektivní Gradient Boosting framework s rychlejším trénováním, nižší paměťovou náročností a podporou paralelního, distribuovaného i GPU učení *(LightGBM developers, nedatováno). CatBoost* je pak zajímavý zejména přímou podporou kategoriálních proměnných, což je pro úlohy nad tabulárními daty prakticky velmi relevantní *(CatBoost developers, nedatováno).* Budoucí práce by proto mohla ověřit, zda některá z těchto metod nepřinese v úloze kreditního skóringu příznivější kompromis mezi predikční výkonností, stabilitou modelu, nároky na předzpracování a výpočetní náročností.

Seznam použitých zdrojů

AYARI, H., P. R. GUETARI a P. N. KRAÏEM, 2026. Machine learning powered financial credit scoring: a systematic literature review. Artificial Intelligence Review. 59, 13. Dostupné z: doi:10.1007/s10462-025-11416-2.

BASEL COMMITTEE ON BANKING SUPERVISION, 2017. Basel III: Finalising post-crisis reforms [online]. Basel: Bank for International Settlements. [cit. 2026-04-04]. Dostupné z: https://www.bis.org/bcbs/publ/d424.htm

BREIMAN, L., 2001. Random forests. Machine Learning. 45(1), 5–32.

BRODERSEN, K. H., C. S. ONG, K. E. STEPHAN a J. M. BUHMANN, 2010. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. Los Alamitos: IEEE, s. 3121–3124. Dostupné z: doi:10.1109/ICPR.2010.764.

BROWN, I. a C. MUES, 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications. 39(3), 3446–3453.

BROWN, K. a P. MOLES, 2014. Credit risk management. Edinburgh: Edinburgh Business School.

CADY, F., 2024. The data science handbook. Hoboken: John Wiley & Sons.

CATBOOST DEVELOPERS, nedatováno. Categorical features. In: CatBoost Documentation [online]. [cit. 2026-04-06]. Dostupné z: https://catboost.ai/docs/en/features/categorical-features

CIL, A. E. a K. YILDIZ, 2025. A systematic literature review on applications of explainable artificial intelligence in the financial sector. Internet of Things. 33, 101696. Dostupné z: doi:10.1016/j.iot.2025.101696.

ČERNEVIČIENĖ, J. a A. KABAŠINSKAS, 2024. Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review. 57, 216. Dostupné z: doi:10.1007/s10462-024-10854-8.

FAIR ISAAC CORPORATION (FICO), 2016. What's in your FICO® Scores? In: myFICO [online]. [cit. 2026-03-29]. Dostupné z: https://www.myfico.com/credit-education/whats-in-your-credit-score

FAWCETT, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters. 27(8), 861–874. Dostupné z: doi:10.1016/j.patrec.2005.10.010

FRIEDMAN, J. H., 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 29(5), 1189–1232.

HAND, D. J. a W. E. HENLEY, 1997. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A. 160(3), 523–541.

HARTMAN, S., C. S. ONG, J. POWLES a P. KUHNERT, 2025. Position: We Need Responsible, Application-Driven (RAD) AI Research. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research [online]. 267, 81514–81525. [cit. 2026-04-08]. Dostupné z: https://proceedings.mlr.press/v267/hartman25a.html

HE, H. a E. A. GARCIA, 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 21(9), 1263–1284. Dostupné z: doi:10.1109/TKDE.2008.239

HUANG, Z., 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 2, 283–304.

CHEN, Y., R. CALABRESE a B. MARTIN-BARRAGAN, 2024. Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research. 312(1), 357–372. Dostupné z: doi:10.1016/j.ejor.2023.06.036.

JAMES, G., D. WITTEN, T. HASTIE a R. TIBSHIRANI, 2021. An introduction to statistical learning. 2nd ed. New York: Springer.

KARTHUR10, nedatováno. Efficient grading for Lending Club with CatBoost. In: Kaggle [online]. [cit. 2026-04-11]. Dostupné z: https://www.kaggle.com/code/karthur10/efficient-grading-for-lending-club-with-catboost

KOMOROWSKI, M., 2016. Exploratory data analysis. In: Secondary analysis of electronic health records. Cham: Springer, 185–203.

KUHN, M. a K. JOHNSON, 2013. Applied predictive modeling. New York: Springer.

LESSMANN, S., B. BAESENS, H.-V. SEOW a L. C. THOMAS, 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research. 247(1), 124–136.

LIGHTGBM DEVELOPERS, nedatováno. Welcome to LightGBM's documentation! [online]. [cit. 2026-04-06]. Dostupné z: https://lightgbm.readthedocs.io/

LOUZADA, F., A. ARA a G. B. FERNANDES, 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science. 21(2), 117–134.

MARTIN, J., 1991. Rapid application development. New York: Macmillan Publishing Company.

MCINNES, L., J. HEALY a J. MELVILLE, 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426 [online]. [cit. 2026-04-08]. Dostupné z: https://arxiv.org/abs/1802.03426

MESTIRI, S., 2024. Credit scoring using machine learning and deep Learning-Based models. Data Science in Finance and Economics. 4(2), 236–248. Dostupné z: doi:10.3934/DSFE.2024009.

PAGÈS, J., 2004. Analyse factorielle de données mixtes. Revue de Statistique Appliquée [online]. 52(4), 93–111. [cit. 2026-04-08]. Dostupné z: https://www.numdam.org/item/RSA_2004__52_4_93_0/

POWERS, D. M. W., 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies. 2(1), 37–63. Dostupné z: doi:10.9735/2229-3981

PYLE, D., 1999. Data preparation for data mining. San Francisco: Morgan Kaufmann.

RANADEEP, nedatováno. Credit Risk Analysis. In: Kaggle [online]. [cit. 2026-04-04]. Dostupné z: https://www.kaggle.com/datasets/ranadeep/credit-risk-dataset

SAITO, T. a M. REHMSMEIER, 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 10(3), e0118432. Dostupné z: doi:10.1371/journal.pone.0118432

SCIKIT-LEARN DEVELOPERS, nedatováno a. Imputation of missing values [online]. [cit. 2026-03-29]. Dostupné z: https://scikit-learn.org/stable/modules/impute.html

SCIKIT-LEARN DEVELOPERS, nedatováno b. OneHotEncoder [online]. [cit. 2026-03-29]. Dostupné z: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

SCIKIT-LEARN DEVELOPERS, nedatováno c. OrdinalEncoder [online]. [cit. 2026-03-29]. Dostupné z: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html

SCIKIT-LEARN DEVELOPERS, nedatováno d. Pipeline [online]. [cit. 2026-03-29]. Dostupné z: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

SCIKIT-LEARN DEVELOPERS, nedatováno e. HistGradientBoostingClassifier [online]. [cit. 2026-04-06]. Dostupné z: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html

SCIKIT-LEARN DEVELOPERS, nedatováno f. Cross-validation: evaluating estimator performance [online]. [cit. 2026-04-06]. Dostupné z: https://scikit-learn.org/stable/modules/cross_validation.html

SCIKIT-LEARN DEVELOPERS, nedatováno g. RFECV [online]. [cit. 2026-04-06]. Dostupné z: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html

SCIKIT-LEARN DEVELOPERS, nedatováno h. Validation curves: plotting scores to evaluate models [online]. [cit. 2026-04-06]. Dostupné z: https://scikit-learn.org/stable/modules/learning_curve.html

TABASSI, E., 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: National Institute of Standards and Technology. Dostupné z: doi:10.6028/NIST.AI.100-1

THOMAS, L. C., 2009. Consumer credit models: Pricing, profit and portfolios. Oxford: Oxford University Press.

VAKRANI, D. S., P. S. PADHYE, S. K. GUPTA a kol., 2026. Evaluating AI-driven credit scoring models versus traditional statistical techniques. Discover Artificial Intelligence. 6, 72. Dostupné z: doi:10.1007/s44163-025-00772-1.

XGBOOST DEVELOPERS, nedatováno. XGBoost Documentation [online]. [cit. 2026-04-06]. Dostupné z: https://xgboost.readthedocs.io/

YANG, L. a A. SHAMI, 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing. 415, 295–316. Dostupné z: doi:10.1016/j.neucom.2020.07.061

Přílohy

Z důvodu rozsahu podkladových materiálů vzniklých v průběhu zpracování bakalářské práce jsou přílohy zpřístupněny pouze v elektronické podobě. Jako doporučená varianta je určen veřejně dostupný notebook na platformě Kaggle, který představuje hlavní a nejpřímější způsob nahlédnutí do praktické části práce. Notebook je zde uložen v prostředí, pro něž byl primárně navržen, vyvíjen a testován.

Doplňkově jsou související materiály zpřístupněny také prostřednictvím školního cloudového úložiště, kde jsou uloženy doprovodné soubory a stručné pokyny ke spuštění notebooku mimo prostředí Kaggle. V takovém případě je však nutné připravit odpovídající běhové prostředí, nainstalovat potřebné knihovny a zajistit nahrání vstupních dat do správného umístění, aby bylo možné notebook korektně spustit a reprodukovat výsledky.

Pro statické prohlížení notebooku mimo prostředí Kaggle je k dispozici také veřejný repozitář GitHub a odkaz pro službu nbviewer. Tyto alternativy slouží především k prohlížení obsahu notebooku; některé pokročilejší vizuální nebo interaktivní prvky se v nich nemusí zobrazovat plně stejně jako v prostředí Kaggle.

Odkazy na elektronické přílohy:

Kaggle:

https://www.kaggle.com/code/karolnakoaov/credit-score-classification-model-2

GitHub:

https://github.com/karolina-kocarova/bachelor-thesis-credit-scoring/blob/main/credit-score-classification-model.ipynb

nbviewer:

https://nbviewer.org/github/karolina-kocarova/bachelor-thesis-credit-scoring/blob/main/credit-score-classification-model.ipynb

Školní úložiště

https://vspj-my.sharepoint.com/:f:/g/personal/kocaro03_student_vspj_cz/IgA6LnlmfRhUQqvvZ8jw96ICAUCxB8JikPSWV-HxkX62TkY?e=PIKDJm