

ŽILINSKÁ UNIVERZITA V ŽILINE
Fakulta riadenia
a informatiky

Fakulta riadenia a informatiky

Prediktívny model na detekciu podvodných úverových žiadostí

Bakalárska práca

Erik Urban

Študijný program: Informatika

Študijný odbor: Informatika

Školiace pracovisko: Žilinská univerzita v Žiline,

Vedúci bakalárskej práce: Ing. Ivan Škula

Žilina 2024

ŽILINSKÁ UNIVERZITA V ŽILINE, FAKULTA RIADENIA A INFORMATIKY.

ZADANIE TÉMY BAKALÁRSKEJ PRÁCE.

Študijný program: Informatika

Meno a priezvisko

Erik Urban

Osobné číslo

561938

Názov práce v slovenskom aj anglickom jazyku

Prediktívny model na detekciu podvodných úverových žiadostí

A predictive model for detection of fraudulent loan applications

Zadanie úlohy, ciele, pokyny pre vypracovanie

(Ak je málo miesta, použite opačnú stranu)

Cieľ bakalárskej práce:

Finančné inštitúcie z podstaty svojho podnikania boli vždy vystavené rôznych podvodným schémam, ktorých cieľom je získať prostriedky na úkor spoločnosti. Z pohľadu kategorizácie podvodov môžu tieto pokusy o podvod vychádzať zo strany zákazníkov ale aj interných zamestnancov. Jeden z najbežnejších typov podvodov ktorým musia finančné inštitúcie čeliť sú úverove podvody. V minulosti bolo relativne prácne overovať informácie ktoré klienti uvedú do formulára žiadosti. Ešte komplikovanejšie bolo preverovať rôzne priame ale aj nepriame väzby medzi osobami uvedenými v žiadostiach o úver. Väčšina spoločností však už má proces zberu žiadostí plne elektronický a teda je možné realizovať analýzu informácií veľmi efektívne a v dokonca aj v reálnom čase. Cieľom práce je navrhnuť a implementovať algoritmus, ktorý na základe vstupných údajov zo žiadosti o úver odvodí mieru rizika a teda sprostredkuje informáciu, či a do akej miery môže byť uvedená žiadosť podvodom.

Obsah:

- oboznámenie sa modus operandi pri tzv. podovodoch 'application fraud'
- formulácia očakávaných vstupných dát pre algoritmus detekcie
- nájdenie datasetu, ktorý obsahuje relevantné dáta na základe definovaných požiadaviek
- definícia odvozených premenných
- dizajn výsledného datasetu pre modelovanie
- spracovanie datasetu a príprava na dátové modelovanie
- aplikácia metód prediktívneho modelovania a ich porovnanie
- výber najlepšieho algoritmu a jeho popis s výhodami oproti iným metódam (presnosť, vysvetliteľnosť apod.)
- implementácia aplikácie s využitím vybraného algoritmu
- formulácia kritérií alebo krokov, ktoré by mohli viest k zlepšeniu efektivity algoritmu

Meno a pracovisko vedúceho BP: Ing. Ivan Škula, KI, ŽU

Meno a pracovisko tutora BP:

Ivan Urban 31.10.2023

garant štud. programu
(dátum a podpis)

Zadanie zaregistrované dňa 31.10.2023 pod číslom 2459/2023 podpis

Erik Urban

Čestné vyhlásenie

Vyhlasujem, že som zadanú bakalársku prácu vypracoval samostatne, pod odborným vedením vedúceho práce a používal som len literatúru uvedenú v práci.

Žilina 5. marca 2024

podpis

Poděkování

Ďakujem môjmu školiteľovi Ing. Ivanovi Škulovi za jeho odborné vedenie, cenné rady a podporu počas celého procesu písania bakalárskej práce. Vážim si jeho ochotu venovať mi čas, trpeznosť a expertízu, ktorá mi pomohla zlepšiť moju prácu a nadobudnúť cenné poznatky a zručnosti.

Abstrakt

URBAN, Erik: *Prediktívny model na detekciu podvodných úverových žiadostí.*
[Bakalárska práca]. – Žilinská univerzita v Žiline. Fakulta riadenia a informatiky.
Katedra Informatiky. – Školtiel/Vedúci: Ing. Ivan Škula – Stupeň odbornej kvalifikácie:
bakalár. – Žilina: FRI UNIZA, 2024, 97 strán.

Finančné inštitúcie, sú vystavené rôznym podvodným schémam, či už zo strany zákazníkov alebo interných zamestnancov. V tejto náročnej situácii môže byť predikčný model dôležitým nástrojom v boji proti podvodom. Cieľom bakalárskej práce je vytvoriť prediktívny model, ktorý na základe vstupných údajov zo žiadosti o úver a podľa najvhodnejšie zvoleného algoritmu pre vytvorenie daného modelu sprostredkuje informáciu, či a do akej miery môže byť úverová žiadosť podvodom. Práca má ambíciu poskytnúť riešenia, ktoré by dokázali znížiť finančné straty plynúce z podvodných úverových žiadostí a ich včasnym odhalením zvýšiť dôveru zákazníkov. Práca zahŕňa postupnú transformáciu – spracovanie zozbieraných dát od žiadateľa o úver pomocou štatistických a analytických metód, využitím znalostí o finančnej doméne a rozpoznaním vzorov do výslednej formy vhodnej na modelovanie. V práci sú porovnané viaceré prediktívne modely, kde kvalita prediktívneho modelu je určená pomocou metrík vyhodnotenia.

Kľúčové slová: prediktívny model, úverové podvody, dátová analytika, detekcia podvodov

Abstract

URBAN, Erik: *A predictive model for detecting fraudulent loan applications.* [Bachelor thesis]. – University of Žilina. Faculty of Management and Informatics. Department of Computer Science. – Supervisor: Ing. Ivan Škula – Degree of professional qualification: Bachelor's in informatics. – Žilina: FRI UNIZA, 2024, 97 pages.

Financial institutions are exposed to a variety of fraudulent schemes, either by customers or by internal employees. In this challenging situation, a good predictive model becomes an important tool in the fight against fraud. The goal of the bachelor's thesis is to create a predictive model that, based on the input data from a loan application and according to the most appropriate algorithm chosen for the creation of the model, provides the information whether and to what extent a loan application may be a fraud. The bachelor thesis aims to provide solutions that could reduce financial losses resulting from fraudulent loan applications and increase customer trust by detecting them early. The work involves a gradual transformation – processing the collected data from the loan applicant using statistical and analytical methods, leveraging financial domain knowledge and pattern recognition into a final form suitable for modeling. In this paper, several predictive models are compared, where the quality of the predictive model is determined using evaluation metrics.

Keywords: predictive model, loan fraud, data analytics, fraud detection

Obsah

Úvod	12
1 Analýza súčasného stavu	14
1.1 Cieľ úveru a rozdelenie úverov podľa zaistenia	14
1.2 Globálny trh a finančné straty na podvodoch	14
1.3 Modi operandi podvodníkov	15
1.4 Podvodné úverové žiadosti	16
1.5 Ochrana proti podvodmi	16
1.6 Strojové učenie	17
1.6.1 Výhody a Nevýhody strojového učenia	17
1.6.2 Výskum zameraný na aplikovanie strojového učenia na detekciu podvodov	18
2 Peer-to-Peer (P2P) platformy	21
2.1 Výhody P2P a Nevýhody P2P	22
2.2 Aplikovanie strojového učenia na peer-to-peer úvery	23
2.3 P2P platformy vo svete a na Slovensku	24
3 Predstavenie dát a vývojového prostredia	25
3.1 Predstavenie dát a spoločnosť Bondora	25
3.2 Vývojové prostredie	26
3.2.1 Python	26
3.2.2 Miniconda a Conda	26
3.2.3 Numpy a Pandas a Pyarrow	27
3.2.4 Matplotlib a Seaborn	27
3.2.5 Scikit-learn a Jupyter Notebook	27
4 Návrh a Spracovanie dát	28
4.1 Metodika definovania závislej premennej	28
4.2 Výber vhodných nezávislých premenných pre dátovú analýzu	30
4.3 Rozdelenie nezávislých premenných podľa typu	32
4.4 Chýbajúce hodnoty – NA (Not Available) hodnoty	35
4.4.1 Kategorické premenné	37
4.4.2 Numerické premenné	40
4.5 Dátová analýza	41
4.5.1 Korelačná analýza	42
4.5.2 Vizualizácia dát	42
4.6 Škálovanie numerických hodnôt	49

4.7 Kódovanie (encoding) kategorických premenných	49
5 Implementácia modelov	52
5.1 Rozdelenie vstupných dát, Overfitting a Underfitting	52
5.2 Hyperparametre modelu	53
5.3 K-násobná krížová validácia (K-fold cross validation).....	54
5.4 Aplikované algoritmy na tvorbu modelu strojového učenia.....	56
5.4.1 Logistická regresia	56
5.4.2 Rozhodovací strom.....	57
5.4.3 Náhodný les	58
6 Prezentácia výsledkov.....	59
6.1 Vyhodnotenie výsledkov modelov	64
6.1.1 Confusion Matrix (matica zámen).....	65
6.1.2 Klasifikačný report.....	67
6.1.3 ROC krivka	69
6.1.4 Výsledná tabuľka.....	71
6.2 Najdôležitejšie nezávislé premenné.....	72
6.3 Hľadanie najlepšieho modelu na minimálnych vstupných dátach	75
Záver	79

Zoznam obrázkov

Obrázok 1 – Prehľad modi operandi podľa Andersona [6].....	15
Obrázok 2 – Výskyt daného algoritmu vo vedeckých článkoch [10].....	18
Obrázok 3 – Vzťah medzi subjektami pri P2P žiadostiach o úver.....	22
Obrázok 4 – Príklad úverovej žiadosti z platformy Žltý Melón [10].....	24
Obrázok 5 – Počet podvodných žiadostí podľa kroku 3 a algoritmus na jej určenie.....	29
Obrázok 6 – Prehľad kategórií pre každý Rating	31
Obrázok 7 – Mapovanie premennej CreditScoreFiAsiakasTietoRiskGrade.....	39
Obrázok 8 – Rozdelenie žiadosti podľa stavu žiadosti v datasete	43
Obrázok 9 – Podvodné úverové žiadosti podľa pohlavia žiadateľa.....	43
Obrázok 10 – Počet žiadostí v určitú hodinu podľa stavu žiadosti	44
Obrázok 11 – Výška úveru a mesačná splátka podľa stavu žiadosti.....	45
Obrázok 12 – Rozloženie výšky úverov	45
Obrázok 13 – Zlyhanie splácania úveru podľa ratingu	46
Obrázok 14 – Výskyt podvodu podľa Ratingu	46
Obrázok 15 – Rozdelenie žiadateľov podľa veku.....	47
Obrázok 16 – Priemerná výška úveru podľa stavu žiadosti.....	47
Obrázok 17 – Počet pôžičiek podľa pohlavia a krajiny	48
Obrázok 18 – Transformácie hodnôt pomocou StandardScaler a MinMaxScaler [23] .	49
Obrázok 19 – Zakódovanie hodín pomocou sínusu a kosínusu[27].....	51
Obrázok 20 – Rozdelenie dát podľa 10-násobnej krízovej validácie	55
Obrázok 21 – Logisticálna funkcia.....	56
Obrázok 22 – Hierarchia rozhodovacieho stromu.....	57
Obrázok 23 – Náhodný les	58
Obrázok 24 – ROC krivka s AUC v hodnote 0.972	64
Obrázok 25 - Confusion Matrix pre Logisticú regresiu.....	65
Obrázok 26 – Confusion Matrix pre Rozhodovací strom	66
Obrázok 27 – Confusion Matrix pre Náhodný les	66
Obrázok 28 – Klasifikačný report pre Logisticú regresiu.....	67
Obrázok 29 – Klasifikačný report pre Rozhodovací strom	68
Obrázok 30 – Klasifikačný report pre Náhodný les	68
Obrázok 31 – ROC a AUC pre Logisticú regresiu	69
Obrázok 32 - ROC a AUC pre Rozhodovací strom	70
Obrázok 33 - ROC a AUC pre Náhodný les	70
Obrázok 34 – Hodnoty koeficientov Logistickej regresie	73

Obrázok 35 – Najvplyvnejšie nezávislé premenné Rozhodovacieho stromu	74
Obrázok 36 – Najvplyvnejšie nezávislé premenné Náhodného lesa.....	74
Obrázok 37 – Confusion matrix pre Náhodný les s minimálnymi vstupmi	76
Obrázok 38 – Klasifikačný report pre Náhodný les s minimálnymi vstupmi	76
Obrázok 39 – ROC a AUC pre Náhodný les s minimálnymi vstupmi	77
Obrázok 40 – Najvplyvnejšie nezávislé premenné Náhodného lesa s minimálnymi vstupmi	77

Zoznam tabuľiek

Tabuľka 1 – Usporiadanie Ratingov podľa vyhodnoteného rizika modelom [20].....	33
Tabuľka 2 – Pomer podvodných úverových žiadostí ku legitímnym	44
Tabuľka 3 – Kódovanie pomocou OneHotEncoding	50
Tabuľka 4 – Výsledné metriky hodnotenia modelov.....	71
Tabuľka 5 – Výsledné metriky hodnotenia modelov pri rozličnej hraničnej hodnote.....	78

Zoznam skratiek

Skratka	Anglický význam	Slovenský význam
KPI	Key Performance Indicators	Key performance indicators
UK	United Kingdom	Spojené kráľovstvo
USA	United States of America	Spojené štáty americké
P2P	Peer-to-peer	Ludia ľuďom
SMS	Short message service	Krátká textová správa
VPN	Virtual Private Network	Virtuálna privátna sieť
IP	Internet Protocol	Internetový protokol
EU	European Union	Európska únia
GDPR	General Data Protection Regulation	Všeobecné nariadenie o ochrane údajov
NA	Not Available	Chýbajúce hodnoty
NAT	Not A Time	Nie je časová hodnota
TN	True Negative	Pravdivo negatívny
FN	False Negative	Nepravdivo negatívny
FP	False Positive	Nepravdivo pozitívny
TP	True positive	Pravdivo pozitívny
MCC	Matthews Correlation Coefficient	Matthewsov korelačný koeficient
ROC	Receiver Operating Characteristic	Krivka prevádzkových charakteristík
AUC	Area Under the Curve	Plocha pod krivkou

Slovník pojmov

Proxy server - sprostredkovateľský server, ktorý funguje ako medzičlánok medzi zariadením a internetom

Kreditné riziko / úverové riziko – finančné riziko vyplývajúce z možnej neschopnosti alebo neochoty dlžníka splatiť svoje záväzky

Parser engine – to softvérový komponent, ktorý analyzuje a interpretuje text podľa pravidiel.

Modi operandi - typický spôsob páchania trestnej činnosti určitým páchateľom

Data mining - proces analýzy veľkých súborov dát za účelom objavovania vzorcov

Data crawler – softvér ktorý prehľadáva internet alebo databázu za účelom zhromažďovania informácií

Gradient – vektor parciálnych derivácií, ukazuje smer a rýchlosť najstrmšieho stúpania funkcie v danom bode

Default – zlyhanie splácania úveru

ÚVOD

Už pri zdrode finančného sektora a myšlienky princípu úveru – majetnejší človek alebo inštitúcia požičia (veriteľ) svoje finančné prostriedky svojmu náprotivku s nižším množstvom finančných prostriedkov (dlžník), ktoré sú následne využité na získanie aktív alebo pasív dlžníkom, za cenu vopred dohodnutej provízie, respektívne úroku, sa začala vyskytovať skupina osôb na strane dlžníka, ktorá chcela tento systém, založený na určitej dôvere medzi veriteľom a dlžníkom zneužiť na vlastné finančné obohatenie, bez úmyslu splatenia pohľadávky voči veriteľovi.

V súčasnej dobe môže osoba z jednej strany planéty požiadať osobu na druhej o pôžičku na základe vzájomnej dohody, bez toho, aby mali medzi sebou osobný kontakt, dokonca nemusia vedieť ani svoju totožnosť, ani sa nikdy vidieť. Určitá dávka anonymity online sveta dodala podvodníkom potrebný pocit bezpečnosti voči pohľadávkam veriteľa. Rozmachom internetu medzi bežnými ľuďmi a prílivom nových veriteľov aj z bohatších krajín sveta vznikli rôzne peer-to-peer platformy, kde sa môže viacero veriteľov posklaadať určitým podielom žiadateľovi o úver. Značnou výhodou poskytovania úveru vyššie uvedenou formou je absencia sprostredkovateľa, táto platforma má hlavne za úlohu spájať ochotných veriteľov priamo zo žiadateľom na úver, čo má za následok možný vyšší výnos. Na druhej strane majú peer-to-peer platformy menšiu úroveň regulácie, čo môže prispieť k zvýšenej frekvencii podvodov. Rovnako tak aj banky a nebankové spoločnosti nezaostávali za digitalizáciou a začali poskytovať technické vymoženosti, ako je online banking, bezhotovostné platby, žiadosť o úver online vyplnením krátkeho dotazníka, kde v určitých prípadoch stačí iba vedieť rodné číslo a číslo občianskeho preukazu, poskytovanie bezúčelových úverov a iné. Všetky tieto faktory zdôrazňujú potrebu efektívneho a sofistikovaného systému, ktorý pomôže veriteľom (či už sú to individuálni veritelia, peer-to-peer platformy, banky, nebankové spoločnosti) ochrániť sa pred možnými podvodníkmi, prípadne ich informovať o možnom riziku.

Pred tým, ako finančné inštitúcie začali využívať strojové učenie sa tieto inštitúcie spoliehali na postupy, ktoré boli často subjektívne a zdĺhavé, lebo informácie museli byť manuálne spracované a na analýzu klienta, ktorá však dokáže pracovať iba s historickými faktami.

Vďaka strojovému učeniu je možné vyvinúť proaktívny prístup k odhalovaniu finančných podvodov, zvýšila sa prevencia pred podvodmi. Objektivita strojového učenia je priamo závislá od zdrojov dát a dokáže spracovať veľký objem dát za krátky čas. Na základe spracovaných dát dokáže vytvorený model predpovedať možné

budúcnosť. Všetky tieto výhody sú motiváciou pre finančný sektor investovať svoje zdroje do rozvoja a integrácie strojového učenia do svojich procesov. Mnoho týchto modelov je avšak uzavretých, pod rúškom súkromných know-how znalostí, ktoré si prísne strážia, nerady zdieľajú informácie o podvodných úveroch a svoje získané dátu o vlastných zákazníkoch.

Cieľom našej práce je vytvoriť riešenie so zohľadnením najefektívnejšieho prediktívneho modelu určeného na detekciu podvodov, na základe kľúčových ukazovateľov výkonnosti (KPI) a určiť najdôležitejšie vstupné atribúty, ktoré ovplyvňujú kvalitu predikcie modelu. Veríme, že takýto model v prípade zapracovania môže zvýšiť dôveru veriteľov v dlžníkov, čo môže mať za následok zvýšený počet vyhovujúcich žiadateľov o úver. Myslíme si, že takýto projekt je veľmi relevantný pre súčasný finančný sektor, ale aj pre súkromné osoby, ktoré chcú zodpovedne pristupovať k svojim investičným aktivitám.

Model po zapracovaní môže byť alternatívou voči súkromným modelom alebo ako aj ich možný doplnok. Náš model je založený na veľkej vzorke heterogénnych dát o užívateľoch, ktorí pochádzajú z viacerých krajín, z rôznych sociálnoekonomickej vrstiev, kde sú zastúpené rozmanité vekové kategórie.

V kapitole 1 bakalárskej práce, Analýza súčasného stavu, sme popísali stav súčasnej problematiky v oblasti predikcie podvodných žiadostí o úver. Následne, v druhej kapitole sme vysvetlili Platformy peer-to-peer, ktoré poslúžili ako zdroj vstupných dát pre našu bakalársku prácu. V kapitole Predstavenie dát a vývojového prostredia sa nachádzajú informácie o zbere dát, ktoré boli použité na vytvorenie predikčného modelu, ich podrobné vysvetlenie a predstavili sme vývojové prostredie, v ktorom sme prácu realizovali. Štvrtá kapitola, Spracovanie dát je zameraná na transformáciu získaného datasetu na formu vhodnú pre spracovanie algoritmami strojového učenia a na dátovú analýzu, kde sa zaobráme skúmaním vzťahov medzi premennými v datasete, snažíme sa odhaliť ich možné spojitosti a súvislosti. Kapitola Implementácia modelov obsahuje popis tvorby a implementácie vybraných modelov. V Prezentácia výsledkov, ktorá je poslednou kapitolou práce vyhodnocuje metriky vyhodnotenia implementovaných modelov a ich najdôležitejšie vstupné atribúty. V Závere práce sme zhodnotili výsledok našej práce, jej limity a možné smery, do ktorých sa môže posúvať ďalší výskum.

1 ANALÝZA SÚČASNÉHO STAVU

1.1 Cieľ úveru a rozdelenie úverov podľa zaistenia

Cieľom úveru je poskytnutie peňažnej sumy žiadateľovi o úver. Dlžník získa vďaka úveru nutné finančné prostriedky a veriteľ získa po splatení úveru pôvodnú zapožičanú sumu a určitú finančnú províziu z pohľadávky. Tá mu vznikne za poskytnutie úveru vo forme úroku.

Úvery môžu byť zabezpečené a nezabezpečené. Zabezpečené úvery sú podložené niečím hodnotným, napríklad majetkom, ako je nehnuteľnosť, auto. Ak dlžník nedokáže splácať dlh voči veriteľovi, tak veriteľ má nárok vymáhať splatenie dlhu, ktorý mu dlžník spôsobil. Klasickým príkladom zabezpečeného úveru je hypotecký úver. Nezabezpečené úvery sú také úvery, ktoré nie sú podložené majetkom, nehnuteľnosťou, kolaterálom. Vhodným príkladom sú napríklad kreditné karty. Nezabezpečené úvery, už zo svojho princípu sú viac náchylné na podvody. Z dôvodu absencie predmetu zabezpečenia sú nezabezpečené úvery spojené s väčším rizikom podvodov.

1.2 Globálny trh a finančné straty na podvodoch

Globálny trh s pôžičkami a platbami vzrástol z \$8721,16 miliardy v roku 2022 na \$9585,48 miliardy USD v roku 2023 [1]. Takýto obrovský trh prirodzene príťahuje veľké množstvo podvodníkov, ale zároveň podnecuje spoločnosť investovať a skúmať možné opatrenia, ktoré by znížili ich úspešnosť.

Podľa ročného výkazu zameraného na podvody spoločnosti UK Finance, Spojené kráľovstvo Veľkej Británie zaznamenalo viac ako £1,2 miliardové straty na podvodoch v roku 2022 [2]. V Austrálii, podľa výkazu organizácie Australian Competition & Consumer Commission straty na podvodoch činili \$3,1 miliardy za rok 2022 [3]. Federal Trade Commission v USA zaznamenala celkovú stratu na podvodoch v hodnote približne \$2,25 miliárd za 4 kvartály roku 2022 [4]. V európskej únií tvorili straty iba na podvodoch s platobnými kartami €1,53 miliárd v roku 2021 [5]. S týchto údajov je zjavné že podvody predstavujú značnú výzvu pre celý finančný trh.

1.3 Modi operandi podvodníkov

Podvodníci sa zameriavajú na určité činnosti, majú určitú sériu krokov, ktorými chcú dosiahnuť svoj cieľ, ktorým je získanie finančných prostriedkov ilegálnym spôsobom – podvodom .

Tieto Modi operandi podvodníkov sa podľa Andresona[6] (Obrázok 1) delia na:

Zneužitie produktu – cieľom je získať informácie o transakčnom produkte, ako je napríklad kreditná alebo debetná karta, šek.

Vzťah podvodníka k účtu – podvody prvej, druhej a tretej strany.

Zneužitie procesu – k podvodu dochádza pri žiadosti alebo transakcií

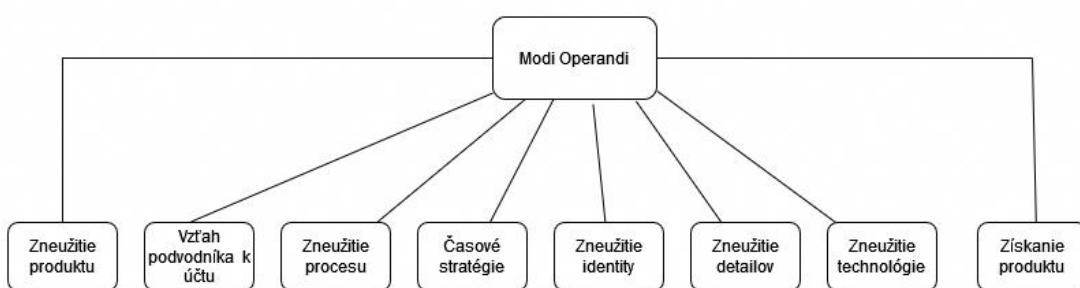
Časové stratégie – krátkodobé a dlhodobé stratégie, kde podvodník vykoná podvod čo najskôr, alebo sa správa určitú dobu ako legitímny zákazník, pred tým než začne vykonávať ilegálnu činnosť.

Zneužitie identity – podvodník mení detaile o svojej vlastnej identite, prípadne sa prezentuje ako niekto iný.

Zneužitie detailov – podvodník vytvára falzifikáty napríklad kreditných kariet, občianskeho preukazu a podobne.

Zneužitie technológie – podvody zamerané na bankomaty, podvody vykonalé prostredníctvom internetu.

Získanie produktu – podvodník získava napríklad kreditnú kartu prostredníctvom krádeže, odchytí poštu, skimming (skopírovanie údajov platobných kariet pomocou zariadení ktoré sú ilegálne nainštalovalené napríklad v bankomatoch, čítačkách kariet).



Obrázok 1 – Prehľad modi operandi podľa Andersona [6]

1.4 Podvodné úverové žiadosti

Naša práca sa zaoberá detekciou podvodných úverových žiadostí. V podvodnej úverovej žiadosti je zámerom dlžníka obohatenie sa na úkor veriteľa, ktorý dlžníkovi poskytuje úver. Dlžník neplánuje úver splatiť.

Tieto podvody môžu byť kategorizované ako podvody prvej, druhej a tretej strany. Podvody prvej strany sú také, kde žiadateľ **poskytne nepravdivé informácie** o sebe za cieľom získania výhodnejšieho úveru. Príkladom môže byť napríklad falošný údaj o príjmoch alebo pracovná **pozícia**. **Podvody druhej strany sú podvody**, kde je využitý koncept takzvanej **peňažnej mulice**, kde podvodník využíva ľudí v zlej situácii, ktorý podvodníkovi poskytnú svoje osobné informácie s cieľom, aby podvodná žiadosť vyzerala legitímne, čo veľmi komplikuje odhalenie nelegálnej činnosti. Podvody tretej strany sú tie najčastejšie, podvodník **ukradne osobe identitu** bez jej vedomia alebo **vytvorí** na základe jeho identity falošnú. Príkladom môže byť krádež totožnosti [7].

Neustály vývoj a inovácia v oblasti detektie podvodných úverových žiadostí je nevyhnutná, pretože podvodníci sa stávajú čoraz sofistikovanejší a veľmi rýchlo sa adaptujú voči ochranným opatreniam a metódam, ktoré vyvíjajú finančné inštitúcie aby im zabránili v čo najväčšej miere v páchaní ilegálnych činností. Momentálne je to bezvýchodisková situácia, v ktorej sa obe strany konfliktu striedajú vo vedúcej pozícii, ktorá má navrch voči druhej.

1.5 Ochrana proti podvodmi

Ochrancu proti podvodmi môžeme rozdeliť na **prevenciu proti podvodu** a **detekciu podvodov**. Do prevencie proti podvodu patrí napríklad šifrovanie údajov o kreditnej karte pri transakcii, číselne sms kódy a tokeny, ktoré treba zadať pri prihlásovaní, overovanie pomocou otlačku alebo tváre, zamedzenie početným pokusom o prihlásenie za krátky čas napríklad formou výberu správnych obrázkov podľa popisu, alebo zadanie správneho textu z obrázka. **Podstatou prevencie** proti podvodom je sťažiť prácu vykonávanú podvodníkmi, ktorú musia vynaložiť pre úspešnú podvodnú aktivitu.

Analýza správania žiadateľa pri aplikačnom procese, analýza a ip adresa zariadenia žiadateľa, z ktorého žiada o úver, používanie VPN alebo proxy serverov, verifikácia prímu a zamestnania, záznamy podvodníkov a podozrivých osôb, frekvencia žiadostí môžu byť kľúčové pre odhalenie podvodných žiadostí o úver [8]. Dôležité sú taktiež interné a externé audity. Toto všetko sú spôsoby **ako detegovať podvod**. Úlohou detektie je zabrániť podvodníkovi spôsobiť škodu. Detekcia podvodov

nastupuje vtedy, keď sa podvodníkovi podarilo prejsť cez preventívne opatrenia. Momentálne sa do popredia dostáva aj **strojové učenie** ako nástroj na odhalovanie podvodov. Použitie strojového učenia môže byť kľúčovým pri prevencii pred podvodmi ale aj pri detekcii podvodov.

1.6 Strojové učenie

Strojové učenie podľa Oracle, jednej z najväčších spoločností na svete je **časť umelej inteligencie, ktorá sa zameriava na spracovanie dát, na základe ktorých sa učí a vylepšuje** [9]. Pri súčasnom stave finančného sektora, kde denne prebiehajú tisíce transakcií za hodinu sa stáva z možnosti zapojenia strojového učenia na odhalenie podvodných pokusov priam povinnosť, ak si daná finančná inštitúcia chce zachovať nie len nadobudnuté finančné zdroje, ale aj svoju reputáciu a zákazníkov. Strojové učenie dokáže pracovať v reálnom čase, čo je kľúčové pre detekciu podvodov.

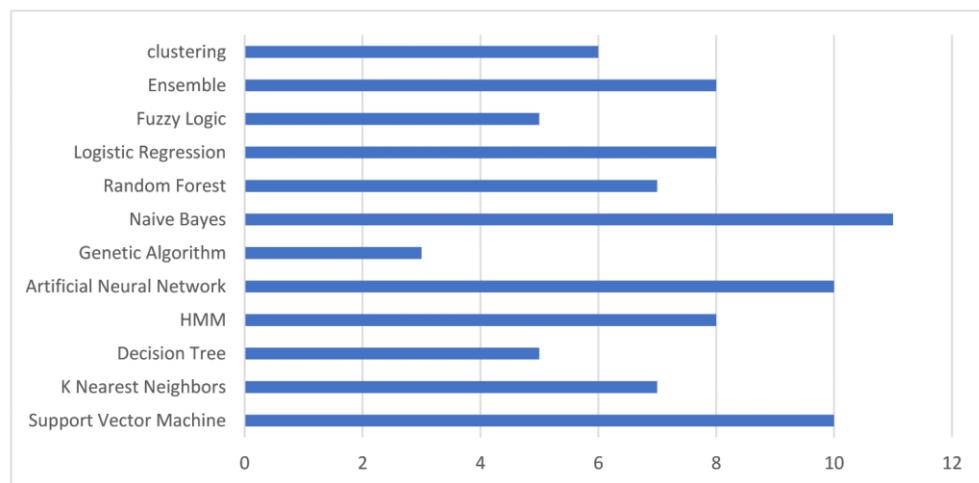
1.6.1 Výhody a Nevýhody strojového učenia

Výhodami strojového učenia je schopnosť spracovať veľké množstvo dát za veľmi krátky čas (ak daný čas porovnáme s časom, ktorý by potreboval človek na ich spracovanie). Taktiež je strojové učenie schopné odhaliť rôzne vzorce, vzťahy, ktoré ani človek, ktorý ma podrobnu znalosť domény nedokáže. Zároveň dokáže podvod odhaliť automaticky, bez externého zásahu. Strojové učenie nie je ovplyvnené externými vplyvmi a emóciami. Má prediktívne schopnosti, dokáže s určitou istotou predpovedať budúce javy, trendy. Vytvorený model je možné prispôsobiť, špecializovať na riešenie konkrétnych problémov, stačí mu iba učiť sa z **vhodných dát**.

Nevýhodami sú napríklad potreba veľkého počtu dát na vytvorenie kvalitného, reprezentatívneho modelu, ktorý bude dosahovať určitú potrebnú úroveň presnosti svojich predikcií. Model ktorý je vytvorený musí byť neustále vyvýjaný, modernizovaný, lebo podvodníci neustále zdokonaľujú svoj proces podvodných schém. To sa aj odzrkadluje na vyšších nákladoch na prevádzku. Na vytvorenie prediktívneho modelu určitej kvality je nutné mať kvalifikovaných pracovníkov, ktorí majú pokročilé znalosti v doméne strojového učenia. Rizikom sú taktiež rôzne etické otázky (príkladom môže byť diskriminácia na základe veku, pracovnej pozície a iné) a problémy so zachovaním súkromia osobných údajov podľa platného zákona (napríklad spĺňanie GDPR pre Európsku úniu). Je potrebné upozorniť na to, že strojové učenie nie je neomylné a stále sa jedná len o určitý druh odhadu na základe dát, ktoré boli poskytnuté na vykonanie predikcie.

1.6.2 Výskum zameraný na aplikovanie strojového učenia na detekciu podvodov

Hlavné rozdelenie strojového učenia je podľa spôsobu učenia : učenie sa s učiteľom alebo bez učiteľa. Pri učení s učiteľom sa v datasete nachádza očakávaná závislá premenná. V prípade určovania podvodnej žiadosti sa jedná o **klasifikačný** problém - závislá premenná je **diskrétna**, má dopredu určené kategórie a v našom konkrétnom prípade je **binárna**, kde **1 znamená podvodná žiadosť a 0 znamená, že žiadosť nie je podvodná (je legitímna)**. Pri učení bez učiteľa sa v datasete **nenachádza závislá premenná**, namiesto toho sa model snaží rozdeliť dátu do skupín, ktoré sú si podobné. Podľa systematického prehľadu literatúry, ktorú vykonali Ali et.al [10] na vzorke 88 vedeckých článkoch vydaných v rokoch 2010-2022, ktorých predmetom bolo odhalovanie finančných podvodov na základe strojového učenia sa medzi **najpopulárnejšie algoritmy** (pozri Obrázok 2) na detekciu finančných podvodov zaraďujú: Podporné vektorové stroje (Support Vector Machine – SVM), K-najbližších susedov (K Nearest Neighbors – KNN), Skrytý Markovov model (Hidden Markov model – HMM), Umelá neurónová sieť (Artificial Neural Network – ANN), Genetický algoritmus (Genetic Algorithm), Naivný Bayes (Naïve Bayes), Náhodný les (Random Forest), Logistická regresia (Logistic Regression), Fuzzy Logika (Fuzzy Logic), Ensemble metódy, Zhlukovanie (Clustering) [10]. Väčšina týchto algoritmov patrí pod kategóriu učenie s učiteľom.



Obrázok 2 – Výskyt daného algoritmu vo vedeckých článkoch [10].

Je nutné podotknúť, že získanie vhodného datasetu na vytvorenie kvalitného modelu je **veľmi náročné**. Finančné inštitúcie musia chrániť zo zákona osobné údaje zákazníka, vďaka ktorým by bola možná jeho priama identifikácia.

Východiskom z tejto komplikovanej situácie môže byť model, ktorý využíva federálne učenie – finančným inštitúciám je zaslaný model, ktorý **lokálne** natrénujú podľa ich dát. Následne namiesto exportovania týchto dát o užívateľovi zašlú natrénovaný model na **vzdialený, zdieľaný server medzi týmito inštitúciami**. Na vzdialenom servery sa tieto modely skombinujú do vylepšeného modelu, ktorý je späť zaslaný finančným inštitúciám. Tento cyklus sa opakuje pokiaľ nebudú dosiahnuté uspokojivé výsledky [11].

Problémom je aj **nevýváženosť datasetu** pri podvodných úverových žiadostiach, v reálnych prípadoch je percentuálne množstvo podvodných úverových žiadostí v **jednotkách percent**. Tento diametrálny rozdiel medzi legitímymi úverovými žiadosťami ktoré predstavujú majoritu žiadostí a podvodných žiadostí, ktoré sú v datasete minoritné spôsobuje **problémy pre väčšinu algoritmov**, ktoré sa používajú na prediktívne modely strojového učenia. Rizikom je aj to, že model bude dosahovať vysokú presnosť iba z takého dôvodu, že bude každú žiadosť predikovať ako legitímnú žiadosť, ktorá je majoritná v datasete. Základnými metódami, ktoré riešia tento problém sú **podvzorkovanie** a **prevzorkovanie**. Pri podvzorkovaní sa náhodne vyberajú prípady, ktoré patrili do väčšiny – ktoré boli legitíme žiadosti a tieto legitíme žiadosti sa odstraňujú, aby došlo k vyváženiu voči podvodným žiadostiam. Naopak, pri prevzorkovaní sa minoritná množina podvodných žiadostí zväčšuje až kým sa nevyrovnaná množina legitímnich žiadostí, čo sa docieli tak, že budú sa náhodne vyberie podvodná žiadosť a vytvorí sa jej kópia, ktorá predstavuje ďalší, nový záznam v datasete, alebo sa môže nová podvodná žiadosť umelo vygenerovať, na základe už existujúcich podvodných žiadostí.

Potenciál na zvýšenie efektivity a robustnosti prediktívneho modelu má aj využitie **multidimenzióvných dát z rôznych zdrojov**. Zhao et. al vo svojom výskume mali prístup nie len k dátam o užívateľovi ako sú vek, pohlavie, príjem a podobne, ale aj k záznamom z aplikácie, ktorú užívateľ využíval na získanie úveru – údaje ako čas vykonania akcie, v ktorej časti aplikácie sa ktorá akcia vykonala a podobne. Zároveň mali dátá o nainštalovaných aplikáciách v telefóne a o tom, aké aplikácie užívateľ odinstaloval a inštaloval [12].

Aplikovanie **data mining techník** má potenciál zvýšiť úspešnosť prediktívnych modelov, tieto techniky dokážu vo veľkom počte historických dát nájsť rôzne skryté vzťahy a vzory, ktoré môžu napomáhať správnej predikcii. Najprv sa **určí** konzistentné správanie užívateľa z historických dát, vytvorí sa profil správania užívateľa a následne sa spustia upozornenia, keď sa správanie užívateľa vymkne z regulárneho správania [13]. Na získanie surových dát, ktoré potom využijú data mining techniky sa

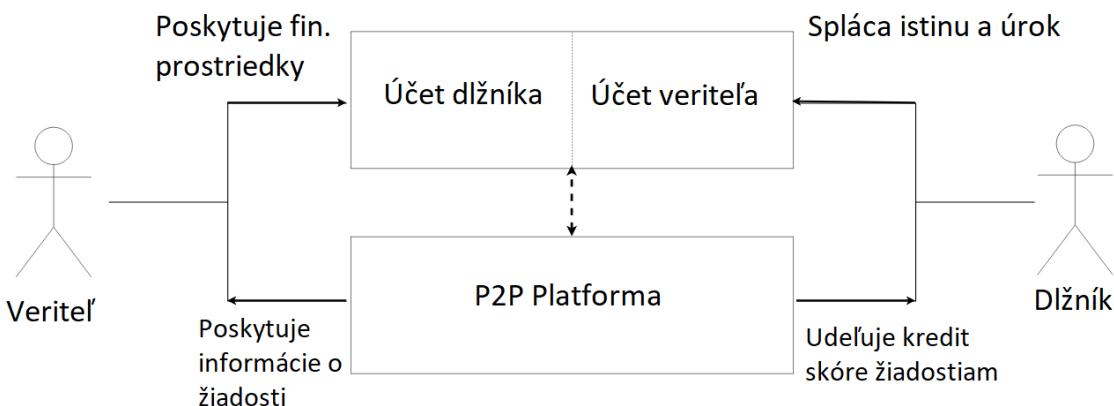
nevyužívajú iba rôzne formuláre, ktoré potrebujú žiadateľove odpovede, ale aj rôzne **data crawler-y**, ktoré sa zameriavajú na informácie ako sú finančné informácie (napríklad príjmy a výdaje žiadateľa, momentálne aktívne úvery), pracovné informácie (napríklad názov spoločnosti v ktorej žiadateľ pracuje, dĺžka spoločnosti na trhu), informácie o transakciách (napríklad či žiadateľ žiadal o nejaký úver predtým, či ho dokázal splatiť) a demografické informácie (napríklad počet súrodencov, miesto narodenia, občianstvo) [14].

2 PEER-TO-PEER (P2P) PLATFORMY

Je viacero dôvodov prečo vznikli **P2P platformy**. Určitý vplyv malo rozšírenie internetu do každého kúta sveta a veľká popularita smartfónov, vďaka ktorým si človek môže jednoducho na P2P platforme podať žiadosť o úver, prípadne investovať ako veriteľ do cudzích úverových žiadostí. Vysoké nároky na žiadateľa od finančných inštitúcií, hlboké, zdĺhavé overovanie údajov o žiadateľovi, vysoké úroky, pocit nepotrebnosti ďalšej strany v záväznom vzťahu a všeobecná averzia voči klasickým finančným inštitúciám ako sú banky zapríčinili veľký rozmach týchto platform po celom svete.

Typická interakcia na P2P platforme zo strany žiadateľa prebieha tak, že si žiadateľ vytvorí žiadosť o úver v určitej sume. Následne je vyrátaný úrok pre danú žiadosť na základe užívateľovi priradeného úverového skóre P2P platformou. Potencionálny veriteľ si dokáže prečítať informácie o žiadosti, ktoré sa rozhodol žiadateľ zverejniť, aby získal investície pre svoju žiadosť. Žiadosť má obmedzenú dĺžku zverejnenia. Ak získa žiadosť potrebné financie, ktoré si vyžiadal žiadateľ, poskytne sa mu úver z peňazí, ktoré investovali veritelia. Zo žiadateľa sa stáva dlžník, ktorý musí splácať mesačne daný úver a veriteľom je z danej splátky vyplácaná čiastka v rovnakom pomere ako zainvestovali do daného úveru. Ak dlžník nesplati mesačnú splátku do určitého obdobia, nastane **zlyhanie splácania úveru (default)**. Celá interakcia na P2P platforme je znázornená na Obrázok 3.

Pri P2P úveroch odpadá tretia strana ktorá sprostredkováva úver, vzťah medzi veriteľom a dlžníkom je priamy. Cieľom žiadateľa o úver je zaujať veriteľa alebo veriteľov, aby investovali svoje finančné prostriedky o ktoré budúci dlžník žiada. Veriteľ očakáva splatenie istiny a taktiež úrok, na ktorom sa dohodli s dlžníkom. Jeden veriteľ môže investovať do viacerých žiadostí o úver a žiadateľ môže získať viacero veriteľov úveru. P2P platforma v tomto vzťahu funguje ako **trhovisko**, ktoré zobrazuje žiadosti o úver možným veriteľom. Účtuje zúčastneným poplatok za svoje služby, poskytuje informácie o žiadosti a udeľuje úverové skóre žiadostiam, ktoré predstavuje riziko nesplácania úveru.



Obrázok 3 – Vzťah medzi subjektami pri P2P žiadostach o úver

2.1 Výhody P2P a Nevýhody P2P

Výhodou pre žiadateľa P2P úveru je to, že pri rozhodovaní veriteľa, či investuje svoje financie do cudzej žiadosti o úver nie sú len tvrdé fakty ako je príjem, vek, vzdelanie ale aj prezentácia žiadateľa, ako je jeho výzor, podrobný popis dôvodu žiadosti o úver, ktorý sa často snaží vyvolať vo veriteľovi emócie a podobne. S toho môže plynúť aj vyššia miera rizika pre veriteľa, čo môže byť negatívum pre veriteľa, ak preferuje bezpečnejšie investície. Zároveň však môže investovať do takých investičných príležitostí ktoré si vyberie a kde by mu to iné finančné inštitúcie nedovolili a dosiahnuť tak vyšší zisk. Výhodou P2P úverov pre veriteľa je aj veľká úroveň diverzifikácie, kde veriteľ môže požičať menšiu sumu viacerým žiadateľom o úver, z rôznych krajín a spoločenských vrstiev. Zároveň veriteľ dosahuje väčšie zúročenie svojich investícií ako keby jeho finančne prostriedky boli uložené v banke a tá by nimi disponovala a namiesto neho poskytovala úvery žiadateľom a fungovala ako tretia strana. **Nevýhodou** sú aj **poplatky**, ktoré platia užívatelia platforme, ktorá sprostredkúva ponuky na financovanie úveru. Veriteľ taktiež nemá možnosť overiť pravdivý dôvod úverovej žiadosti, žiadateľ môže prezentovať dôvod, ktorý je však odlišný od pravého účelu. Výhodou a zároveň nevýhodou môže aj byť anonymita oboch zúčastnených strán. Existuje tu aj určitá nerovnováha informácií, veritelia nemajú prístup k detailným informáciám o žiadateľovi. Detailnejšie informácie majú o žiadateľovi poskytovatelia P2P platform, ktoré sa snažia **zmenšiť riziko**, ktoré podstupujú veriteľa poskytnutím napríklad **udelením úverového skóre pre každú žiadosť**.

2.2 Aplikovanie strojového učenia na peer-to-peer úvery

Súčasný výskum aplikovania strojového učenia na peer-to-peer úvery sa zameriava hlavne na úverové skóringové systémy a modely zamerané na pravdepodobnostné riziko nesplácania úveru alebo skóringové systémy zamerané na zisk [15]. Aj keď úverový podvod začína skoro v každom prípade tak, že žiadateľ o úver prestane úver splácať, treba pripomenúť, že **každé nesplatenie úveru sa automatický neoznačuje za podvod**. Nesplatenie úveru môže mať viacero dôvodov, ktoré môžu byť úplne legítimne.

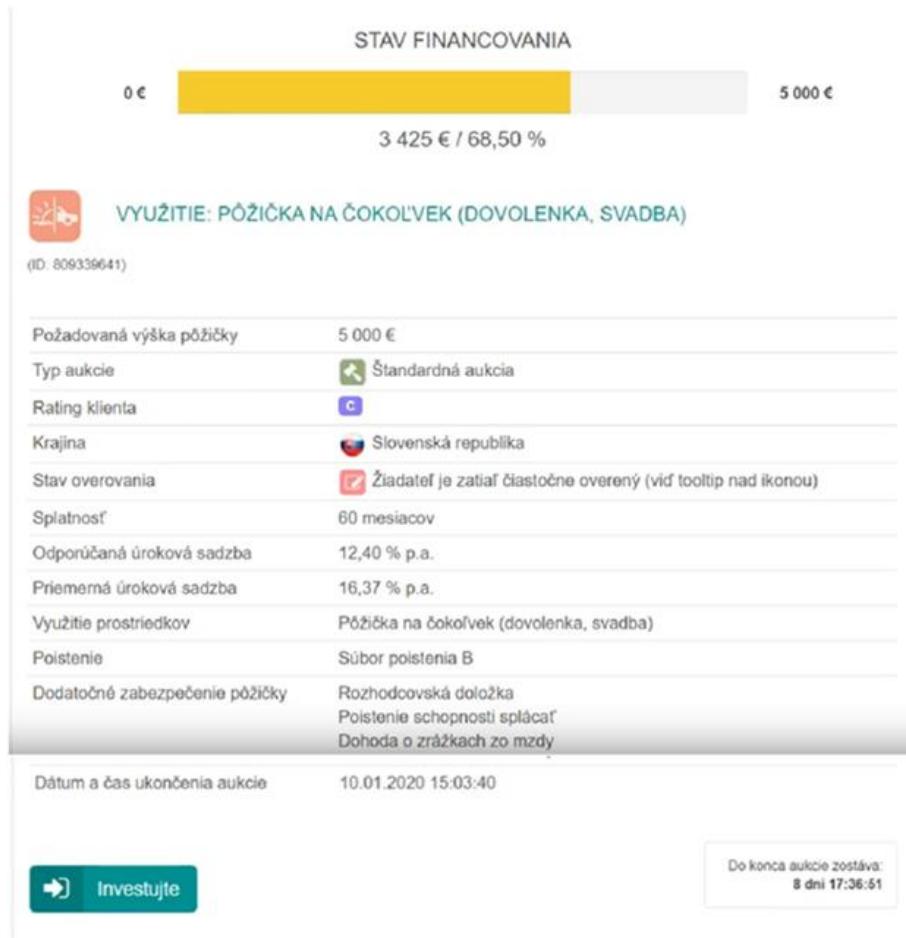
Literatúra zameraná na detekciu podvodov pri peer-to-peer úveroch je hlavne zameraná na čínske peer-to-peer platformy. Hlavným dôvodom podľa autorov [13] je absencia národného kreditového systému a neexistencia komunikácie medzi P2P platformami a tradičnými finančnými inštitúciami. Toto všetko ma za dôsledok zvýšenú aktivitu podvodníkov. Zároveň čínsky P2P úverový trh je najväčší na svete. Autori argumentujú, že dôvodom nižšej iniciatívy vo výskume podvodov zameraných na P2P platformy v rozvinutých krajinách ako USA alebo v krajinách EU je dôkladný systém hodnotenia úverovej spoľahlivosti a prísne zákony proti podvodom, vďaka ktorým je ľahšie klamať o svojej úverovej histórii [13].

Rovnako ako dátá o žiadostiach o úver z klasických finančných inštitúcií, úverové žiadosti na P2P platformách majú **problém v nevyváženosťi** medzi legítimnymi žiadostami a podvodnými. Tieto podvodné žiadosti však dosahujú o niečo **väčšiu frekvenciu** ako v klasických finančných inštitúciách. Zvýšený počet podvodných žiadostí môže mať viacero dôvodov, príkladom môže byť relatívna novota, menšia miera regulácie a overovanie informácií o žiadateľovi, chýbajúca tretia strana medzi veriteľom a dlžníkom.

2.3 P2P platformy vo svete a na Slovensku

Najznámejšou P2P platformou v Európe je Mintos so sídlom v Lotyšsku. Veľmi známa v Európe je aj platforma **Bondora** so sídlom v Estónsku. V Amerike je to zas LendingClub, je to jedna z najstarších a najväčších P2P platforiem na svete, na dátach o žiadostiach o úver, ktoré spoločnosť poskytuje bolo vykonaných niekoľko štúdií zameraných na vytvorenie skóringových modelov. Platforma ma sídlo v USA. V Číne sú populárne platformy Lufax ktorá je súčasťou Alibaba Group a Jingdong Finance, ktorá je súčasťou Jingdong(JD.com) group. Na **Slovensku** je najpopulárnejšia platforma s názvom **Žltý melón**, ktorá je na Slovenskom trhu od roku 2012. Ďalším príkladom je platforma **Finzo** a v **Českej republike** sú to napríklad platformy **Zonky**, ktorá je najväčšou platformou v Českej republike a platforma Bondster.

Na Obrázok 4 je možné vidieť príklad úverovej žiadosti na **platforme Žltý Melón**. Celý príklad úverovej žiadosti z platformy Žltý melón sa nachádza v Príloha A.



Obrázok 4 – Príklad úverovej žiadosti z platformy Žltý Melón [10]

3 PREDSTAVENIE DÁT A VÝVOJOVÉHO PROSTREDIA

3.1 Predstavenie dát a spoločnosť Bondora

Spoločnosť Bondora je spoločnosť, ktorá sa sústredí na poskytovanie peer-to-peer úverov na Európskom trhu od roku 2008. Úverové žiadosti, ktoré ponúka sú najmä zo štátov ako je Fínsko, Španielsko a Estónsko, pričom v Estónsku firma aj sídli a je regulovaná hlavným regulačným orgánom v danej krajine. Tieto úverové žiadosti sprostredkúva investorom zo 40 krajín sveta. Bondora sa zameriava hlavne na úvery s **hodnotou od 500 do 10000 eur**. Doba splatnosti pri úveroch, ktoré Bondora poskytuje investorom je v rozsahu od **3 mesiacoch až do 5 rokov** [17].

Úverové žiadosti sú v Bondore rozdelené na takzvaný **Primárny** a Sekundárny trh. V Primárnom trhu môžu veritelia investovať priamo do novovytvorených žiadostí o pôžičku. Sekundárny trh je zaujímavý tým, že na ňom obchodujú veritelia medzi sebou, predávajú a nakupujú už existujúce pôžičky, na ktoré majú pohľadávky.

Podobne ako Žltý Melón aj Bondora mala prehľad úverovej žiadosti, toto zobrazenie však bolo odstránené v roku 2016 a Primárny trh sa viac nezobrazuje v užívateľskom rozhraní, dáta o úverovej žiadosti sú dostupné iba v API. Argumentom bolo, že ich automatický manažér portfólia je efektívnejší ako manuálne investovanie.

V našej bakalárskej práci sme sa rozhodli použiť datasety **LoanData.csv** a **DebtEvents.csv** od spoločnosti Bondora. Datasety od spoločnosti Bondora sú denne aktualizované a v nich zverejnené dáta spíňajú všetky zákony, ktoré sa vzťahujú na zachovanie ochrany osobných údajov. Veľkou výhodou je aj to, že dáta vychádzajú z reálnych transakcií užívateľov, nejedná sa o syntetické dáta. Každý záznam v datasete s úverovými žiadosťami (LoanData) obsahuje hojný počet vlastností a veľkosť datasetu je dostatočné veľká pre vytvorenie viero hodného prediktívneho modelu. Taktiež na rozdiel napríklad od Žltého Melóna je na Bondore väčšia diverzita, užívateľ si môže podať úverovú žiadosť z viacerých krajín Európy, Žltý Melón sa sústredí na úverové žiadosti iba zo Slovenskej republiky. Datasety sú vo formáte csv, kde ako separátor je použitá čiarka.

Dataset **LoanData** [18] obsahuje 112 stĺpcov – vlastností a 375479 záznamov. Obsahuje úverové žiadosti od roku 2009 do roku 2024. **Legenda** sa nachádza na podstránke Public Reports na doméne bondora.com spoločnosti Bondora [18].

Dataset **DebtEvents** [18] obsahuje záznamy o procese vymáhania pri nesplácaní úveru. Má 5 stĺpcov a 3914009 záznamov, ktoré popisujú každú udalosť v procese vymáhania. **Legenda :**

ReportAsOfEOD – reprezentuje deň (EOD – End Of Day), kedy bol report vygenerovaný

LoanId – jedinečný identifikátor úveru, v danom datasete viaže danú udalosť vymáhania s úverom

CreatedOn – deň a čas, kedy bola vykonaný daná udalosť

Event – vykonaná udalosť v procese vymáhania

Comment – doplnkové vysvetlenie udalosti

3.2 Vývojové prostredie

3.2.1 Python

V našej bakalárskej práci sme si zvolili programovací jazyk Python. V poslednej dobe sa Python teší čoraz väčšej popularite, či už u bežných programátorov, ale aj vo svete strojového učenia. Od jeho popularity sa taktiež odvíja veľká komunita vývojárov a bohatý výber knižníc. Výhodou Pythonu oproti iným jazykom zameraným na strojové učenie, ako je napríklad populárny programátorský jazyk R je jeho relatívna prístupnosť, ktorú zabezpečuje jednoduchá syntax a taktiež jeho všeestrannosť. Python možno použiť nielen pri strojovom učení a štatistike ale aj na riešenie bežných programátorských problémov, webových aplikáciách, jednoduchých počítačových hráčov a podobne.

3.2.2 Miniconda a Conda

Miniconda je minimalistický inštalátor, ktorý obsahuje Python, Condu a základné balíčky. **Conda** je manažér balíčkov a závislosti. Výhodou je, že dokáže vytvoriť ohraničené prostredie, v ktorom sa snaží zabezpečiť kompatibilné verzie jednotlivých knižníc a závislosti. Výhodou je jej kompatibilita medzi platformami, jednoduchosť zdieľania prostredia medzi vývojármami a možnosť viacerých vývojárskych prostredí, ktoré sa môžu odlišovať nielen v knižničach a závislostiach, ktoré využívajú, ale aj v daných verziách.

3.2.3 Numpy a Pandas a Pyarrow

Python knižnica **Numpy** je zameraná na matematické výpočty, manipuláciu s poľami a maticami. Tvorí základ mnohých iných knižníc zameraných na dátovú analytiku, strojové učenie a manipuláciu s číselnými hodnotami.

Pandas je Python knižnica, ktorá sa zaobrá prácou s dátami a dátovými štruktúrami, využíva sa na analýzu dát. Pri svojej práci využíva knižnicu Numpy, ktorá je nutnosťou pre prácu s Pandas, hlavne s číselnými hodnotami.

Knižnicu **Pyarrow** používame pri načítaní csv dát kvôli efektívite, kde pyarrow parser engine je rýchlejší a menej pamäťovo náročný ako python parser engine.

3.2.4 Matplotlib a Seaborn

Matplotlib je Python knižnica na vykreslovanie dát. Je úzko prepojená s Numpy a Pandas, s ktorými spolupracuje a dokáže efektívne spracovať nimi zadefinované dátové typy, ako sú napríklad numpy polia, alebo pandas DataFrame. **Seaborn** je nadstavba nad knižnicou Matplotlib, poskytuje jednoduchšiu syntax, krajšiu vizualizáciu a zjednodušuje prácu s komplexnými prípadmi, ktoré je nutné vizualizovať.

3.2.5 Scikit-learn a Jupyter Notebook

Scikit-learn je Python knižnica, ktorá obsahuje algoritmy zamerané na strojové učenie. Spolupracuje s Numpy, Pandas a Matplotlib. Výhodou tejto knižnice je veľká komunita, rozsiahla dokumentácia a jednoduchosť oproti iným komplexnejším knižniciam, ktoré sa zaobrajú strojovým učením.

Jupyter Notebook je webová aplikácia, vhodná na interagovanie s dokumentami, ktoré obsahujú zdrojové kódy, vizualizácie a výpočty. Výhodou oproti napríklad populárnej cloud služby Google Colab je možnosť lokálneho spustenia a väčšia úroveň súkromia a modifikácie.

4 NÁVRH A SPRACOVANIE DÁT

4.1 Metodika definovania závislej premennej

Ked' sa dlžník oneskorí pri svojej mesačnej splátke, vzniká mu dlh voči veriteľom. Dôvody na oneskorenie môžu byť úplne legitímne ako strata práce alebo náhle a nečakané finančné výdavky. Bondora používa na vymáhanie dlhu voči investorom **3 krokový proces** [19] :

1. Krok je interný pokus o vymáhanie splátky. Dlžníkovi sú zaslané upomienky od firmy Bondora o okamžité splatenie dlhu. Zároveň aj poskytnú informácie lokálnemu registru pohľadávok, ktorý sa taktiež snaží spojiť s dlžníkom. Tento proces je vykonávaný pri oneskorení v rozsahu od 1 až 120 dní od poslednej splátky.
2. Ked' je dĺžka nesplácania dlhšia ako 3 mesiace, žiadosť je preklasifikovaná ako zlyhanie splácania úveru (default). Ak je dĺžka nesplácania dlhšia ako 120 dní, tak informácie a detaily o dlhu sú zaslané na súd a je podaná žaloba. Na verdikt súdu sa čaká pri najlepšom aspoň 4 mesiace.
3. Ak súd vydá pozitívny verdikt a uzna oprávnenosť žaloby, tak sa daný prípad posunie súdnemu exekútorovi, ktorý je oprávnený vymáhať pohľadávku voči dlžníkovi.

Dataset **LoanData** nemá určené, či daná úverová žiadost bola podvodná alebo nie, preto sme museli vytvoriť **vlastnú metodiku**, na základe našich empirických skúseností a znalostí o doméne. Základnou indíciou, že sa jedná o podvod je v našej metodike nesplácanie dlhu viac ako 3 mesiace, čo znamená, podľa pravidel spoločnosti Bondora preklasifikovanie úveru na úver, kde bolo **zlyhané splácanie** (default). Na praktické určenie zlyhania splácania dlhu sme zmenili dátumovú premennú **DefaultDate**, ktorá reprezentuje dátum, kedy došlo k zlyhaniu na binárnu premennú **Default**, ktorá hovorí, či došlo k zlyhaniu splácania alebo nie. Každé zlyhanie splácania dlhu ale neznemená v našej metodike, že došlo k podvodu. Aby došlo k podvodu musí byť splnená jedna s nasledujúcich podmienok:

1. Úverová žiadost sa dostala do stavu zlyhania splácania úveru a dlžník **nesplati ani jedinú splátku**. Takýchto prípadov je **5510**.
2. Úverová žiadost sa dostala do stavu zlyhania splácania úveru a dlžník splatil **menšiu čiastku**, ako sú **dve splátky**. Je ich dokopy **31845**.

Podmienka vyjadrená matematicky:

splatená suma kým sa dostal úver do stavu zlyhania < 2 * mesačná splátka.

3. Úverová žiadosť sa dostala do stavu zlyhania splácania úveru, nachádza sa v datasete **DebtEvents** – rovnaké **LoanID** sa nachádza v oboch datasetoch a v stĺpci **Comment** (komentár k udalostiam vymáhacieho procesu) sa nachádza hodnota:
 - a. Kategória **CriminalCase** – podozrenie na kriminálnu aktivitu, má viacero podkategórií.
 - b. **HopelessCase** – Prípad bol vyhodnotený ako bez šance na úspešné vymáhanie
 - c. **HeavilyIndebted** – Dlžník je nadmieru zadlžený
 - d. **CustomerLivesOutsideSupportedCountries** – Dlžník žije mimo krajinu, kde má Bondora dosah na vymáhací process.
 - e. **LawsuitPresentedAgainstBondora** – Dlžník zažaloval Bondoru v procese vymáhania

K jednotlivým pojmom v datasete DebtEvents nebola poskytnutá legenda a ani popis k hodnotám v datasete, zdôvodnenia hodnôt vyplývajú z **odborného odhadu**. Prehľad hodnôt sa nachádza v Príloha B . Spolu ich je **3643** (pozri Obrázok 5).

Pomocou našej metodiky predpokladáme, že zo všetkých **375479** záznamov žiadosti je **po zjednotení 33515** podvodných, to je takmer **9% (~8.926%)** z celkového počtu všetkých záznamov. Pre každý záznam sme vytvorili dva nove stĺpce, *Fraud* a *Explanation* , kde **Fraud** je závislá premenná, ktorú chceme predikovať pomocou nášho modelu a **Explanation** predstavuje zdôvodnenie prečo sme určili tento záznam ako podvodný podľa našej metodiky.

```
sus = ['CriminalCase', 'HopelessCase', 'HeavilyIndebted',
       'CustomerLivesOutsideSupportedCountries', 'LawsuitPresentedAgainstBondora']
pattern = '|'.join(sus)
dataSusID = dataDebt[dataDebt["Comment"].str.contains(pattern,
                                                       case=False, na=False, regex=True)].LoanId.unique()
print(len(dataLoan.loc[dataLoan.LoanId.isin(dataSusID) & (dataLoan.Fraud != 1)]))
dataLoan.loc[dataLoan.LoanId.isin(dataSusID),
            ['Fraud', 'Explanation']] = [1, '200+ days payment overdue without response and suspicious behaviour']
```

1515

Obrázok 5 – Počet podvodných žiadostí podľa kroku 3 a algoritmus na jej určenie

4.2 Výber vhodných nezávislých premenných pre dátovú analýzu

Náš dataset **LoanData** po zadefinovaní závislej premennej **Fraud** a doplnení nezávislých premenných **Default** a **Explanation** obsahuje 115 stĺpcov.

Podľa legendy, ktorú poskytuje firma Bondora o nezávislých premenných (stĺpce/vlastnosti záznamu) a znalosti domény sme sa rozhodli zredukovať počet stĺpcov. Použili sme nasledujúcu metodiky:

1. **Nezávisle premenné so žiadnou vyplnenou hodnotou** – do tejto kategórie patrili stĺpce, ktoré neobsahovali žiadnu hodnotu. Pomocou jednoduchého algoritmu sme zistili, či hodnota v riadku pre daný stĺpec nie je vyplnená. Ak je nevyplnená, tak algoritmus vráti boolean hodnotu True a naopak ak je vyplnená, tak vráti False. Kedže Python konvertuje True na 1 a False na 0, stačí vypočítať priemer, aby sme zistili, v akej percentuálnej miere nie je stĺpec reprezentujúci nezávislú premennú naplnený. Algoritmus odhalil, že tieto nezávislé premenné nemajú vyplnenú žiadnu hodnotu :

DateOfBirth, County, City a EmploymentPosition.

2. **Dátumové nezávisle premenné:**

Pre takéto premenné sme sa rozhodli vypočítať dĺžku intervalu medzi dvoma časovými úsekmi v dňoch. Párovanie dátumov vzniklo na základe znalosti domény, legendy a potencionálneho vplyvu na určenie podvodnej žiadosti. Boli to tieto premenné:

Pre GracePeriodStart a GracePeriodEnd sme vytvorili stĺpec DeltaGracePeriod, pre MaturityDate_Last a MaturityDate_Original stĺpec DeltaMaturityDate, pre LastPaymentOn a FirstPaymentDate stĺpec DeltaF&LPayment,

pre DebtOccuredOn a BiddingStartedOn stĺpec DeltaBidding&DebtOccured.

Vytvorili sme aj nový stĺpce Rescheduled, ktorý reprezentuje dátumovú premennú RescheduledOn. Rescheduled je binárna premenná. Ak mal dlžník niekedy pridelený nový splátkový kalendár, táto hodnota sa nastavila na logickú hodnotu TRUE, inak sa nastavila na FALSE.

3. Nezávislé premenné, ktoré neobsahovali hodnotu vo viac ako 90% prípadov

Na zistenie nezávislých premenných, ktoré splňajú našu podmienku, sme použili obdobný algoritmus ako v kroku 1. len so zmenenou hraničnou hodnotou. Zistené premenné sú NrOfDependants, WorkExperience, EL_V0, Rating_V0, EL_V1, Rating_V1, Rating_V2, CreditScoreEsEquifaxRisk. Premenné s kategóriou Rating_Vx sú vytvorené ratingovým modelom firmy Bondora, kde označenie za podčiarkovníkom a písmenom V a s číslom x znamená verziu ratingového modelu. Dataset obsahuje taktiež premennú Rating, ktorá však obsahuje iba 2733 nevyplnených hodnôt. Rozhodli sme sa ju doplniť a prepísť kombináciou všetkých stĺpcov Rating_Vx. Týmto postupom sme znížili počet nevyplnených Rating hodnôt na 2717 z pôvodných 2733, ale rôznych hodnôt ako pôvodný Rating, lebo ak bola vyplnená hodnota z kategórie Rating_Vx, tak táto hodnota nahradila hodnotu zo stĺpca Rating. Týchto hodnôt bolo 25455. Vzťah medzi Rating_V0, Rating_V1, Rating_V2 a Rating je taký, že hodnota z vyšej verzie má prednosť pred nižšou. Tento vzťah môžeme zjednodušiť nasledovne:

$$\text{Rating_V2} > \text{Rating_V1} > \text{Rating_V0} > \text{Rating}$$

Táto transformácia je možná lebo každý Rating obsahuje rovnakých 7 kategórií (pozri Obrázok 6).

```
|: ## Vsetky ratingy maju rovnake kategorie hodnotenia
| print(sorted(dataLoan.loc[dataLoan["Rating"].notna(),"Rating"].values.unique().tolist()))
| print(sorted(dataLoan.loc[dataLoan["Rating_V0"].notna(),"Rating_V0"].values.unique().tolist()))
| print(sorted(dataLoan.loc[dataLoan["Rating_V1"].notna(),"Rating_V1"].values.unique().tolist()))
| print(sorted(dataLoan.loc[dataLoan["Rating_V2"].notna(),"Rating_V2"].values.unique().tolist()))
|
| ['A', 'AA', 'B', 'C', 'D', 'E', 'F', 'HR']
| ['A', 'AA', 'B', 'C', 'D', 'E', 'F', 'HR']
| ['A', 'AA', 'B', 'C', 'D', 'E', 'F', 'HR']
| ['A', 'AA', 'B', 'C', 'D', 'E', 'F', 'HR']
```

Obrázok 6 – Prehľad kategórií pre každý Rating

Ďalej sme obdobne zlúčili nezávislé premenné EL_Vx, ktoré predstavujú očakávanú stratu podľa x verzie modelu do nového stĺpca Combined_EL. Zvyšné premenné sme sa rozhodli ponechať a pokračovať v ich spracovaní v podkapitole 4.4.

4. Irrelevantnosť nezávislej premennej pre určenie podvodnej žiadosti

Tieto premenne sme určili ako nerelevantné pre našu prácu. Sú to premenné ako LoanNumber a PartyID, ktoré sú jedinečné identifikátory úveru alebo dlžníka v systéme spoločnosti Bondora, kategória Bids, ktorú ovplyvňujú veritelia a ďalšie, iné premenné, ktoré nemajú pre nás výpovednú hodnotu.

Prehľad všetkých odstránených nezávislých premenných je v Príloha C. Po vykonaní všetkých krokov našej metodiky nám ostalo 87 vhodných nezávislých premenných.

4.3 Rozdelenie nezávislých premenných podľa typu

Nezávislé premenné môžu mať rôzne typy. Každý typ má iný druh spracovania, aby bol vhodný pre spracovanie modelom strojového učenia. Najdôležitejšie typy sú:

1. **Numerické premenné** – v našom datasete sú to premenné ako Age, ktoré sú celočíselné (Python ich typ určuje ako int64) a číselné premenné ako je Interest, ktoré sú vo forme desatinné čísla (v Pythone float64). Numerické hodnoty je vo väčšine modelov strojového učenia vhodné upraviť, tak aby mali **rovnakú škálu**, problémom pri rôznych škálach môže byť napríklad nesprávne priradenie väčšej významnosti nezávislej premennej s väčšími hodnotami pre predikovanie podvodu, aj keď nie je dôležitejšia pri predikcii ako nezávislá premenná s nižšími hodnotami. Vysporiadať sa s číselnými hodnotami, ktoré majú rôznu škálu sa môžeme pomocou **normalizácie** alebo **štandardizácie**. Tento proces je spracovaný v podkapitole 4.6.
2. **Kategorické premenné** – hodnoty týchto premenných dosahujú **konečný počet**, môžu byť vo forme slova alebo zakódované pomocou celých čísel. V našom datasete sú to hodnoty vo forme reťazca, napríklad premenná Country, pre číselne zakódované kategorické hodnoty je to napríklad VerificationType. Kategorické premenné sa ďalej môžu kategorizovať na nominálne, ordinálne. **Ordinálne** sú také, ktoré majú určitú postupnosť, napríklad premenná Rating, kde hodnota 'AA' je najlepšia možná hodnota, ktorá môže byť pridelená dlžníkovi a 'HR' (High Risk – vysoký risk) je najhoršia, celé usporiadanie je v Tabuľka 1. **Nominálne** hodnoty sú také, ktoré sa nedajú zmysluplnie usporiadať do postupnosti. Špeciálnym prípadom sú binárne kategorické premenné, ktoré obsahujú iba dve hodnoty, môžu byť vo forme Pravda/Nepravda (True/False) alebo obdobne ako 1/0. Príkladom v našom datasete je premenná NewCreditCustomer, ktorá symbolizuje, či zákazník

nemal úverovú históriu(True) v spoločnosti Bondora, alebo úverovú históriu už mal (False).

Python kategorické hodnoty často chybne interpretuje ako číselné hodnoty, ak sú zakódované v číselnej forme, je nutná ich transformácia. Môžeme ich transformovať na typ objekt (object), ktorý Python využíva, ak nevie určiť iný typ hodnoty, alebo výhodnejšie na typ kategória (category), ak vieme určiť presnú množinu hodnôt ktoré bude premenná nadobúdať. Typ kategória je a menej pamäťovo náročný ako objekt z dôvodov úplne obmedzeného počtu možných hodnôt na rozdiel od objektu. Taktiež konvertovanie na typ kategória je vhodné pre iné Python knižnice, ktoré vyžadujú takýto typ premennej pred ich spracovaním. Aby nás model strojového učenia dokázal spracovať **kategorické premenne** musí byť **každá zakódovaná** vo forme čísel (takzvaný encoding). Pri ordinálnych hodnotách sa pri zakódovaní **musí zachovať významová postupnosť hodnôt**.

Kódovanie kategorických hodnôt sa nachádza v podkapitole 4.7.

Tabuľka 1 – Usporiadanie Ratingov podľa vyhodnoteného rizika modelom [20]

Očakávaná strata		
MIN	MAX	RATING
0.9%	2%	AA
2%	3%	A
3%	5.5%	B
5.5%	9%	C
9%	13%	D
13%	18%	E
18%	25%	F
25%	∞	HR

3. **Dátumové a časové hodnoty** – hodnoty ktoré zachytávajú informácie o dátume, čase alebo zachytávajú obe informácie v jednej hodnote. Príkladom premennej, ktorá zachytáva dátum a čas je v našom datasete premenná

BiddingStartedOn, ktorá hovorí odkedy mohli veritelia poskytnúť finančné prostriedky na zverejnenú úverovú žiadosť dlžníka. Premenná, ktorá zachytáva iba dátum je DebtOccuredOn, ktorá hovorí, kedy vznikol dlh z dôvodu nesplácania úveru dlžníkom. Premenná ApplicationSignedHour predstavuje časovú hodnotu hodiny v ktorú bola úverová žiadosť podaná žiadateľom o úver.

Dátumových a časových premenných je na rozdiel od kategorických premenných nekonečno veľa. Preto ich **nemožno priamo transformovať** na typ **kategória**. Jednou z možných transformácií je **rozdelenie podľa jednotlivých zložiek** na deň, mesiac, rok, hodiny, minúty a ak je potreba aj na menšie časové zložky a vytvoriť z nich nové nezávislé premenné. Následne ich môžeme buď transformovať na kategorické premenné alebo ich brať ako číselne, oba prístupy majú svoje výhody a nevýhody. Tento prístup je vhodný, ak chceme napríklad zistiť, či podvodníci preferujú určitý čas alebo deň v týždni na podanie podvodného úveru. Príkladom môže byť ApplicationSignedHour alebo aj premenná ApplicationSignedWeekday, ktoré boli s vysokou pravdepodobnosťou transformované spoločnosťou Bondora z dátumu podania úverovej žiadosti. Iný prístup môže byť transformácia dvoch dátumových alebo časových premenných na číselnú tak, že zistíme **rozdiel medzi nimi**. Vhodným príkladom môže byť nami vytvorená premenná DeltaBidding&DebtOccured, ktorá vznikla výpočtom uplynulých dní medzi premennými DebtOccuredOn a BiddingStartedOn. Táto hodnota nám môže poskytnúť informáciu v akej miere má vplyv uplynulá doba medzi týmito významnými udalosťami na to, či úverová žiadosť bude podvodná alebo nie.

Ďalším prístupom na transformáciu môže byť jednoduchá binárna hodnota, ktorá reprezentuje, či udalosť nastala (True) alebo nie (False) ako v prípade tvorby nezávislej premennej Rescheduled z dátumovej premenej RescheduledOn. Oba popisy daných transformácií možno vidieť v podkapitole 4.2 odsek 2.

V Príloha D sa nachádza **prehľad nezávislých premenných**, ktoré boli určené ako kategorické (premenná s názvom columns_to_category) a ako premenné typu dátum a čas (premenná s názvom dates). V kroku 1 sú vybrané všetky kategorické premenné do dátovej štruktúry typu list s názvom columns_to_category. Následne v kroku 2 sa vytvorí dátová štruktúra slovník, kde sa nachádzajú prvky v tvare kľúč:hodnota z týchto hodnôt, kde kľúčom je názov nezávislej premennej a hodnotou je

category, čo je dátový typ kategória. Tento slovník je vstupným argumentom pre parameter dtype, ktorý konvertuje všetky vybrané stĺpce datasetu pri načítaní csv súboru podľa ich názvu na vybraný typ. V kroku 3 sú vybrané všetky premenné, ktoré spĺňajú tvar zhodný s typom dátum a čas do dátovej štruktúry typu list s názvom dates. V kroku 4 sú dátumy v premennej dates konvertované na typ datetime podľa formátu deň.mesiac.rok. Ak nie je možné nejakú hodnotu v danom stĺpci nezávislej premennej konvertovať, je za ňu dosadená hodnota NaT, Not a Time.

4.4 Chýbajúce hodnoty – NA (Not Available) hodnoty

Chýbajúce hodnoty môžu v datasete vzniknúť z rôznych dôvodov. Najčastejším dôvodom je nevyplnenie údaju žiadateľom, ak daná informácia nebola nastavená ako povinná na vyplnenie. Iným dôvodom môže byť to, že sa dátu o tejto vlastnosti začali zbierať neskôr, ako sa začalo prvotné zbieranie dát o žiadostiach. Krajným dôvodom môže byť i technický problém ako chyba v softvéri alebo strata časti databázy.

Tieto hodnoty sú vo svojej podstate problémové hlavne z pohľadu interpretácie, ak dataset obsahuje mnoho chýbajúcich hodnôt, tak nemožno z takýchto dát vyvodiť jasné závery, môže dôjsť k skresleniu predikcie. Preto aj rôzne štatistické metódy a modely strojového učenia vyžadujú, aby sa v datasete nenachádzali žiadne chýbajúce hodnoty.

Najpoužívanejšie spôsoby **ako sa vysporiadáť s chýbajúcimi hodnotami** sú:

1. **Odstránenie úverových žiadostí, ktoré obsahujú neznáme hodnoty.**

Najpriamočiarejší spôsob ako vyriešiť problém chýbajúcich hodnôt. Avšak tento spôsob je nepoužiteľný pri malej nazbieranej vzorke dát alebo ak takýchto žiadostí s nevyplnenými hodnotami je príliš mnoho a ich odstránením môže dôjsť ku skresleniu výsledkov alebo nepresnosti a strate schopnosti predikcie.

2. **Odstránenie celej vlastnosti (stĺpca, nezávislej premennej).** Tento postup je legitímny, ak sa v príliš veľa záznamoch nenachádza daná hodnota. Nastáva tu však riziko, že zachovanie vlastnosti môže signifikantne zvýšiť schopnosť predikcie prediktívneho modelu v prípade, že by sme dokázali inými spôsobmi vysporiadania sa s chýbajúcimi hodnotami a doplniť dané hodnoty.

3. Vyplnenie chýbajúcich hodnôt pomocou štatistických metód aplikovaných na vyplnené hodnoty. Najčastejšie to býva medián, modus, priemer a podobne.

Tento spôsob môže však zmeniť skutočnú výpovednú hodnotu dát, preto treba doplňovať hodnotu až po rozsiahлом zvážení vplyvu metódy na dané dátu.

4. Vyplnenie chýbajúcej hodnoty na základe kvalifikovaného odhadu a znalosti domény. Na základe inej vlastnosti úverovej žiadosti môžeme odhadnúť hodnotu vlastnosti, kde sa nenachádza hodnota, ak sa domnievame že medzi nimi existuje vzájomný vzťah a navzájom sa ovplyvňujú. Príkladom môže byť napríklad vek žiadateľa a pracovná pozícia. Je vysoko pravdepodobné, že človek, ktorý dosiahol dôchodkový vek v danom štáte, bude nezamestnaný, bude na starobnom dôchodku a úver bude aspoň čiastočne hradíť príjomom z dôchodku. Tento prístup však môže rovnako ako vyplnenie hodnoty pomocou štatistických metód ovplyvniť pravdivosť údajov, preto je dôležité podrobne a dôkladne zvážiť možný dopad na doplnenie hodnôt týmto spôsobom na výsledné dátu.

V Príloha E sa nachádza stav vyplnenia všetkých premenných datasetu v tomto bode našej práce. V našom datasete sa nachádza v tomto momente **375 479** záznamov. V stĺpci Non-Null je zobrazené, koľko hodnôt je vyplnených pre danú vlastnosť záznamu. To znamená, ak dané číslo je menšie ako 375 479, tak sa v ňom nachádzajú chýbajúce hodnoty.

Ako prvé sme začali spracovávať nezávislé premenné, kde je viac ako 90% NA hodnôt. Týmito stĺpcami boli NrOfDependants, CreditScoreEsEquifaxRisk, WorkExperience, , Combined_EL.

Napriek zlúčeniu v podkapitole 4.2 Combined_EL stĺpec obsahuje viac než 90% nevyplnených hodnôt. Keďže 90% hodnôt majú tieto vymenované premenné nevyplnené, použiť štatistické metódy ako medián, modus nie je vhodné. Po dôkladnom preskúmaní datasetu sme nenašli inú nezávislú premennú, ktorá by nám pomohla s ich odhadom. Preto jediné zmysluplné riešenie je tieto stĺpce vylúčiť z datasetu.

Následne sme systematicky postupovali v spracovaní ďalších kategorických premenných.

4.4.1 Kategorické premenné

Skúmaním datasetu sme taktiež zistili, že MaritalStatus, OccupationArea a UseOfLoan sú taktiež na 90% tvorené NA hodnotami, ktoré sú však zakódované ako hodnota -1 namiesto prázdných hodnôt. Tieto premenné sme taktiež odstránili.

Premenné ActiveLateCategory, ActiveLateLastPaymentCategory, WorseLateCategory podľa legendy a po preskúmaní datasetu symbolizujú počet dní v rôznych štádiách oneskorenia splátky. Sú reprezentované rozsahmi oneskorených dní (napr. 121-150). Pri týchto premenných predpokladáme, že nevyplnená hodnota znamená, že dlžník sa nedostal do štátia oneskorenia platby, čomu nasvedčuje aj legenda a dát, ktoré neobsahujú hodnotu, ktorá by predstavovala neomeškanie splácania. Preto je hodnota 0 (dní) vhodná ako náhrada za NA hodnotu. RecoveryStage predstavuje štádium v ktorom sa nachádza proces vymáhania dlhu. Táto kategória má celočíselnú reprezentáciu. Podľa legendy neexistuje hodnota určenia úverovej žiadosti, ktorá sa nedostala do tejto fázy, preto aj tu je vhodné vytvoriť kategóriu 0, do ktorej budú spadať žiadosti, ktoré sa do tejto situácie nedostali.

Premenná Education, predstavujúca dosiahnuté vzdelanie žiadateľa. Je celočíselne zakódovaná a obsahuje dve hodnoty, ktoré nie sú zapísané v legende, sú to -1 a 0. Po preskúmaní legendy a dát sme usúdili, že sa jedná o rôzny zápis pre neuvedené vzdelanie. Preto sme tieto kategórie zlúčili do kategórie 0. Okrem tohto problému sa tu nachádzalo 50 chýbajúcich hodnôt, ktoré sme taktiež nahradili kategóriou 0, do úvahy však pripadá aj odstránenie daných záznamov, kde sa NA hodnota v stĺpci Education nachádza. Kategóriu 0 sme vytvorili aj preto, lebo hodnôt pôvodnej kategórie -1 a 0 sa tu nachádza relatívne nízky počet a preto sme nepristúpili k odstraňovaniu záznamov alebo celej vlastnosti.

Podobne ako Education, premenná EmploymentStatus obsahuje dve hodnoty, ktoré nie sú zapísané v legende. Sú to -1 a 0 a premenná je rovnako celočíselne zakódovaná. Postupovali sme rovnako, zlúčením týchto dvoch kategórií. Rozdielom oproti Education je to, že hodnót s v pôvodných kategóriach -1, 0 a NA hodnôt je príliš veľa. Po preskúmaní datasetu sme zhodnotili, že tieto hodnoty môžeme doplniť podľa vlastnosti EmploymentDurationCurrentEmployer, ktorá hovorí o tom, akú dobu žiadateľ pracuje pre súčasného zamestnávateľa. Ak sa nachádzala hodnota v EmploymentDurationCurrentEmployer a nebola to hodnota "Retiree" a "Other" (podľa nášho kvalifikovaného odhadu reprezentujú ľudí, ktorí nemajú v súčasnosti zamestnávateľa), tak v takom prípade nastavíme EmploymentStatus na číselnú hodnotu 3, ktorá reprezentuje podľa legendy trvalý pracovný pomer. Trvalý pracovný pomer je najčastejší pracovný pomer nie len v našom datasete ale aj v reálnom svete,

preto je vhodný kandidát na vyplnenie NA hodnôt, ktoré spĺňajú našu podmienku. Podarilo sa nám doplniť 282556 hodnôt.

Zároveň sme využili EmploymentStatus na doplnenie NA hodnôt v EmploymentDurationCurrentEmployer, kde sme vyplnili tieto hodnoty kategóriou UpTo5Years , ak kategória hodnoty EmploymentStatus nebola 1, čo podľa legendy znamená unemployed – nezamestnaný alebo nebola 0, čo je nami vytvorená kategória pre nevyplnený EmploymentStatus. Doplnili sme takto 715 hodnôt.

Premenná Gender obsahuje 3 kategórie, kde kategória 2 je Bondorou určená pre nedefinované pohlavie. Preto sme vyplnili NA hodnoty kategóriou 2. Gender je zakódovaná celočíselne.

HomeOwnershipType je ďalšia premenná, kde sa nachádza kategória -1, ktorá nie je uvedená v legende. HomeOwnershipType neobsahuje kategóriu pre nevyplnené typy vlastníctva domu a je ich relatívne nízky počet (1660), preto sme sa rozhodli odstrániť záznamy, kde je táto premenná nevyplnená. HomeOwnershipType je tiež zakódovaná celočíselne.

Pri premennej LanguageCode sme zistili výskyt viacerých číselných kódov jazykov, ktoré nemajú oporu v legende datasetu. Ich početnosť je nízka, okrem číselného kódu 19. Po preskúmaní datasetu sme zistili že kód jazyka 19 majú iba žiadateľa z krajiny NL (Holandsko). Preto sme usúdili, že kód 19 je holandčina a hodnotu zachovali. Zvyšné neznáme hodnoty sme odstránili pomocou odstránenia záznamov. Prišli sme o 17 záznamov.

Premenná ModelVersion obsahuje viaceré celočíselné kategórie, ktoré však nie sú popísane v legende od spoločnosti Bondora. Usudzujeme preto, že model kategórie 0 reprezentuje žiadosť, na ktorú neboli použití hodnotiaci systém úverového rizika od spoločnosti Bondora, ako je uvedené v legende. Preto sme všetky NA hodnoty tejto premennej zmenili na túto kategóriu.

V premennej Rating sme sa rozhodli pre jednoduché vylúčenie zvyšných NA hodnôt, ktoré zostali nevyplnené po transformácii v sekcii 3 podkapitole 4.2. Jednalo sa o 1062 záznamov.

30 nevyplnených hodnôt mala premenná MonthlyPaymentDay, rozhodli sme sa pre odstránenie záznamov.

Pri premenných CreditScoreEeMini, CreditScoreFiAsiakasTietoRiskGrade ,CreditScoreEsEquifaxRisk, CreditScoreEsMicroL, , ktoré reprezentujú hodnotenia úverových žiadostí na základe histórie o žiadateľovi zo strany Bondory alebo iných firiem zameraných na udeľovanie tohto hodnotenia (detaľy v legende) [18]. Všetky tieto premenné obsahujú veľké množstvo NA hodnôt, kde by došlo k strate väčšiny

záznamov pri pokuse o ich odstránenie formou odstránenia záznamov. Zároveň vieme, že tieto firmy majú ďaleko rozsiahlejšie, podrobnejšie informácie ako sa nachádzajú v našom verejnom datasete, čo by nám mohlo veľmi pomôcť na dosiahnutie našich cieľov. Rozhodli sme sa ich preto zlúčiť pod spoločnú, novú premennú a pretransformujeme ich individuálne kategórie, ktoré sú všetky odlišné, na jednotné kategórie patriace novej, nezávislej premennej **CreditScoreUnified**. V CreditScoreUnified sme určili nové kategórie. Týmito kategóriami sú : **Very Low, Low, Average, High, Very High**. Aby sme docieli toto zjednotenie, odstránili sme z CreditScoreEeMini všetky nevysvetlené kategórie, ktoré neboli popísane v legende. Následne sme 6 kategórii premennej CreditScoreEeMini namapovali na 5 nových kategórii CreditScoreUnified, ktoré predstavujú riziko spojené so žiadostou. Pri tejto transformácii sme sa rozhodli podľa legendy o zlúčenie kategórie 600 a 500 pod kategóriu Very High. Pri CreditScoreEsMicroL sme zistili, že obsahuje hodnotu 'M', ktorá nevystupuje v legende, preto sme ju zmenili na NA hodnotu. Následne sme 10 kategórii CreditScoreEsMicroL pretransformovali na 5 kategórií CreditScoreUnified zlúčením 2 po sebe nasledujúcich kategórií. Pri CreditScoreEsEquifaxRisk došlo k zlúčeniu kategórie 'AAA' a 'AA' po dôkladnom zhodnení podľa popisu jednotlivých kategórii. Najproblematickejšia bola premenná CreditScoreFiAsiakasTietoRiskGrade, kde sa nachádzalo viacero hodnôt, ktoré nemajú popis. Kedže pod nepopísanými kategóriami sa nachádzalo veľké množstvo záznamov, rozhodli sme sa ich najprv zjednotiť na kategórie premennej CreditScoreFiAsiakasTietoRiskGrade, ktoré boli popísane v legende. Kategóriu 8 sme vylúčili pre jednoduchšie mapovanie. Výsledné mapovanie na kategórie CreditScoreUnified je vidno v Obrázok 7.

Zvyšné NA hodnoty premenných, ktoré sa nám nepodarilo zlúčením doplniť sme vyplnili pod novú kategóriu Not Rated, ktorú sme pridali po výslednej transformácii. Pôvodné nezávislé premenné sme odstránili.

```
##zLucujem 1 a 2 a 6 a 7.... 8micku mazeme, Lebo nevieme |priradit a je ich malo
dataLoan = dataLoan[~dataLoan["CreditScoreFiAsiakasTietoRiskGrade"].isin([8])]
asiakas = {
    'RL1': 'Very Low', '1': 'Very Low', '2': 'Very Low',
    'RL2': 'Low', '3': 'Low',
    'RL3': 'Average', '4': 'Average',
    'RL4': 'High', '5': 'High',
    'RL5': 'Very High', '6': 'Very High', '7': 'Very High'
}
```

Obrázok 7 – Mapovanie premennej CreditScoreFiAsiakasTietoRiskGrade

4.4.2 Numerické premenné

Našim prvým krokom bolo preskúmanie hodnôt numerických premenných. Zistili sme, že premenné PrincipalBalance a InterestAndPenaltyBalance obsahujú zápornú hodnotu, čo by však nemalo nastať, lebo v prípade PrincipalBalance, by to znamenalo, že dlžník zaplatil väčšiu čiastku ako bola pohľadávka úveru. V prípade InterestAndPenaltyBalance by to zas bolo, že dlžníkové nezaplatené úroky a penále sú v zápornej hodnote. Záznamy s týmito chybnými hodnotami sme odstránili. Následne sme odstránili všetky záznamy, kde počet NA hodnôt pre konkrétnu nezávislú premennú bol menší ako 10. Pokračovali sme premennými ktoré reprezentovali aspoň 70% všetkých hodnôt ako NA hodnoty.

Boli to [18]:

1. InterestAndPenaltyWriteOffs a PrincipalWriteOffs – odpisy úroku, penále a dlhovanej sumy pohľadávky. Nahradili sme novou binárnu premennou HadWriteOffs, ktorá hovorí, či úverová žiadosť mala odpisy.
2. DeltaGracePeriod – počet dní takzvanej Grace periódy, zmenili sme na novú binárnu premennú HadGracePeriod, ktorá reprezentuje, či úverová žiadosť mala Grace periód. Grace períoda je doba, kedy môže dôjsť k oneskoreniu platby bez negatívnych dôsledkov.
3. PlannedPrincipalTillDate – výška istiny, ktorú mala investícia získať podľa aktívneho plánu, nemožná approximácia podľa dostupných dát.
4. PreviousEarlyRepaymentsBeforeLoan – suma, ktorá bola splatená na predčasných splátkach dlžníka, ktoré splatiel pred daným poskytnutím úveru. Nemožná približná approximácia, čiastočne stratené dátá poskytuje PreviousEarlyRepaymentsCountBeforeLoan, ktorá hovorí o počte predčasných splátok pred daným poskytnutím úveru.
5. PrincipalDebtServicingCost – cena stratená na vymáhaní požičanej sumy podľa pohľadávky, pre naše potreby nerelevantná.
6. InterestAndPenaltyDebtServicingCost – cena stratená na vymáhaní na vymáhaní úrokov a penále, pre naše potreby nerelevantná.

Všetky tieto nezávislé premenné sme odstránili.

Nezávislé premenné EAD1 a EAD2 predstavujú výpočet očakávanej straty pomocou modelov úverového rizika. Na týchto výpočtoch boli závislé aj premenné PlannedPrincipalPostDefault, PrincipalRecovery, PlannedInterestPostDefault, InterestRecovery. V dôsledku veľkého počtu chýbajúcich hodnôt a nedostatočného

vnútorného pohľadu na proces určovania hodnôt týmito modelmi sme museli tieto premenné odstrániť.

K odstráneniu premennej sme sa uchýlili aj pri InterestAndPenaltyBalance, PreviousRepaymentsBeforeLoan a CurrentDebtDaysPrimary, kde pri PreviousRepaymentsBeforeLoan boli informácie aspoň čiastočne zachované v NoOfPreviousLoansBeforeLoan, ktorá hovorí o počte splatených predchádzajúcich pôžičiek.

Premenná DeltaBidding&DebtOccured bola transformovaná na binárnu hodnotu DebtOccured, ktorá hovorí, či nastal dlh. Hodnoty sa nám podarilo určiť nielen pomocou DeltaBidding&DebtOccured ale aj pomocou premennej Default, lebo k dlhu muselo dôjsť v prípade, že nastalo až zlyhanie splácania úveru.

NrOfScheduledPayments, ktorá obsahuje počet plánovaných splátok sme approximovali v prípade NA hodnôt pomocou premennej LoanDuration, ktorá obsahuje koľko mesiacov sa má splácať úver.

V prípade premennej NextPaymentNr sme zistili, že hodnota 0 nastane v prípade zlyhania splácania úveru (default). Preto sme NA hodnoty approximovali ako medián hodnôt, pričom pri výpočte mediánu sme hodnotu 0 nezahŕňali, aby sme dosiahli relevantnejší počet nasledujúcich platieb pre žiadosti, kde nenastal default. Medián je vhodný ako kompromis, ktorý redukuje vplyv odľahlých hodnôt a vysokú štandardnú odchýlku pri tejto premennej.

V neposlednej rade sme odstránili aj riadky, ktoré obsahovali NA hodnotu v stĺpci PrincipalOverdueBySchedule alebo PlannedInterestTillDate, či DeltaF&LPayment a MonthlyPayment, lebo ich bolo nemožné správne doplniť.

Na záver sme NA hodnotám EmploymentDurationCurrentEmployer priradili kategóriu Other, ktorá môže zahrňovať nevyplnené hodnoty alebo momentálne nepracujúcich u žiadneho zamestnávateľa.

Výsledkom transformácie pôvodného datasetu na dataset, kde sa nenachádzajú NA hodnoty je **349813 záznamov a 66 stĺpcov (premenných)**, pričom podvodných žiadostí je **8.14%** z celkového počtu úverových žiadostí.

4.5 Dátová analýza

Dátová analýza je veľmi užitočný nástroj na pochopenie vzťahov vyplývajúcich z dát. Pomáha nám efektívne preskúmať a identifikovať rozdielne vzorce správania podvodných a legitímnych úverových žiadostí. Správne pochopenie dát pomocou dátovej analýzy je dôležité pre vytvorenie efektívneho prediktívneho modelu. Na

pochopení vzťahov medzi dátami a ich preskúmaním sa veľkým dielom podieľa najmä vhodná vizualizácia.

Dátová analýza nám pomáha aj odhaliť **odľahlé hodnoty – Outliers**, to sú také hodnoty, ktoré sa znateľne odlišujú od zvyšných nameraných hodnôt. Tieto hodnoty môžu byť dôležité pri odhaľovaní podvodných žiadostí, ale môžu aj skresľovať a mať aj za následok chybné natrénovanie modelu a tým znižovať jeho kvalitu. Pomôcť so skresľovaním modelu, ktoré vzniká vplyvom odľahlých hodnôt môže napríklad škálovanie v podkapitole 4.6.

V úvode dátovej analýzy je vhodné, aby sme **zistili základné charakteristiky** datasetu: aritmetický priemer, min, max, smerodajná odchýlka, 1., 2., 3., a 4. kvartil pre numerické hodnoty. Pre kategorické premenné to zas sú početnosti jednotlivých kategórií pre danú premennú.

4.5.1 Korelačná analýza

Korelačná analýza je metóda, ktorá nám hovorí o možných vzťahoch medzi premennými. Možný vzťah medzi premennými je vyjadrený korelačným koeficientom. Kladné hodnoty medzi premennými naznačujú spojitosť, ak sa hodnota jednej premennej zvyšuje, zvyšuje sa hodnota aj tej druhej. Pri záporných hodnotách to znamená, že keď sa jedna zvyšuje, druhá klesá. V Príloha F sa nachádza **korelačná matica** (Pearsonov korelačný koeficient) pre numerické hodnoty [21]. V Príloha G je maticové zobrazenie **Cramerovho V** (kontingenčný koeficient) [22], ktorá hovorí o miere asociácie medzi kategóriami, čím väčšia hodnota (nemôže byť záporná), tým vyššia asociácia medzi kategóriami.

4.5.2 Vizualizácia dát

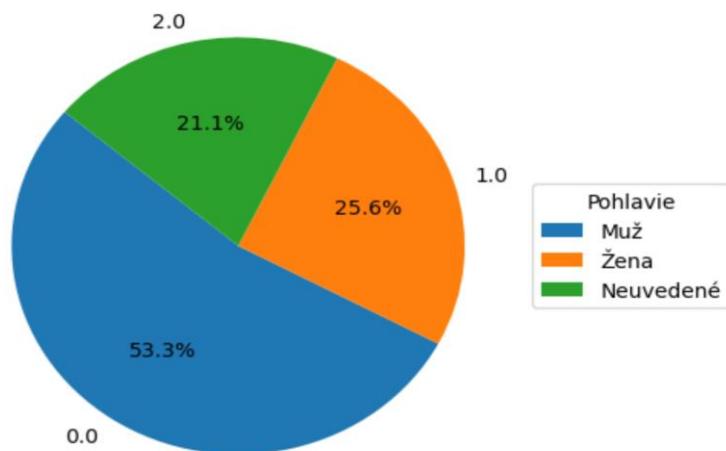
Na Obrázok 8 je zobrazené rozdelenie žiadostí podľa toho, čí sú zaradené medzi legitímne alebo podvodné žiadosti. Vidíme jasný rozdiel vo frekvencií medzi legitímnymi a podvodnými, dataset je nevyvážený.



Obrázok 8 – Rozdelenie žiadosti podľa stavu žiadosti v datasete

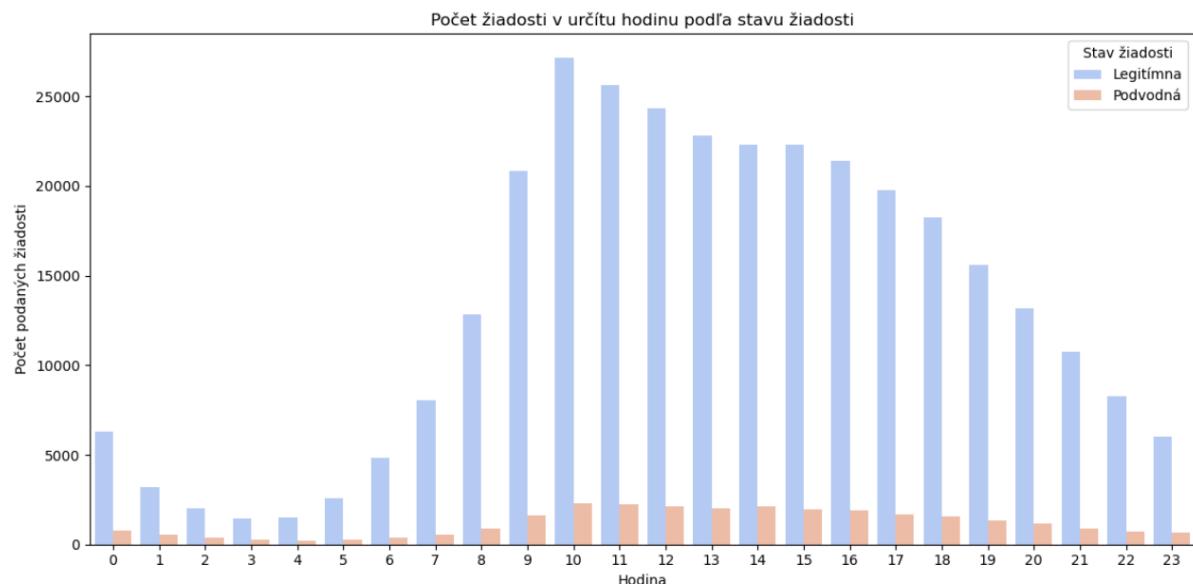
Obrázok 9 ukazuje, že v prípade mužov je viac ako dvojnásobná šanca, že ich úverová žiadosť bude podvodná ako u žien.

Podvodné úverové žiadosti podľa pohlavia žiadateľa



Obrázok 9 – Podvodné úverové žiadosti podľa pohlavia žiadateľa

Podľa grafu Obrázok 10 je možno vidieť zvýšený počet podvodných žiadostí vo večerných hodinách najmä medzi 1:00 a 4:00 hodinou v noci. Toto zistenie potvrdzuje aj Tabuľka 2.

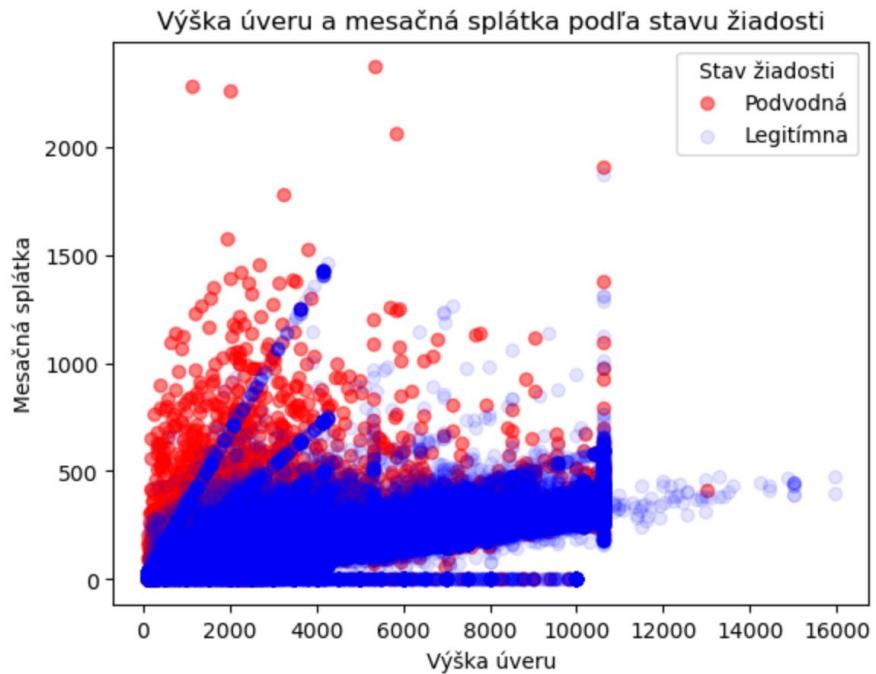


Obrázok 10 – Počet žiadostí v určitú hodinu podľa stavu žiadosti

Tabuľka 2 – Pomer podvodných úverových žiadostí ku legitímnym

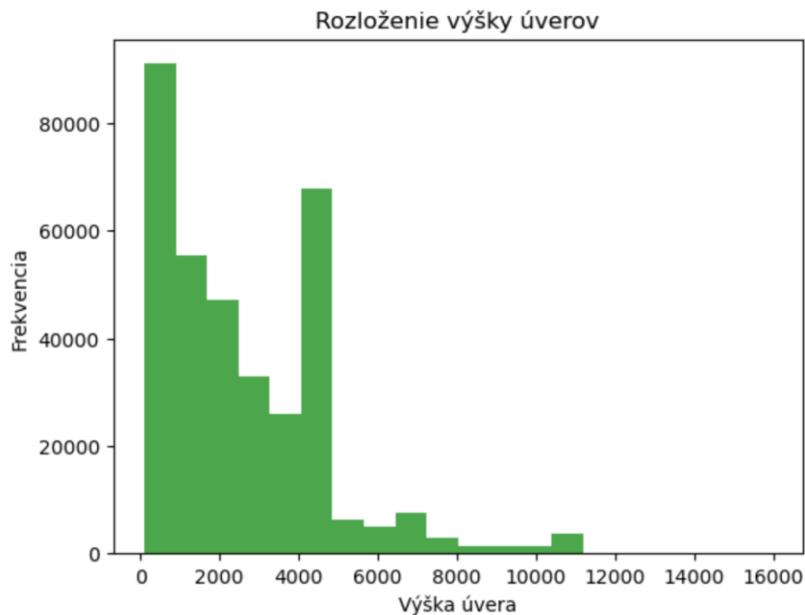
Hodina	Pomer	Hodina	Pomer
0	12.41%	12	8.85%
1	16.83%	13	8.74%
2	18.17%	14	9.50%
3	17.06%	15	8.83%
4	14.26%	16	8.88%
5	9.68%	17	8.52%
6	7.99%	18	8.52%
7	7.01%	19	8.47%
8	6.99%	20	8.73%
9	7.67%	21	8.45%
10	8.48%	22	8.68%
11	8.64%	23	10.91%

Scatter plot na Obrázok 11 hovorí, že nižšia výška úveru a pridelená vysoká mesačná splátka zvyšuje riziko, že úverová žiadosť bude podvodná.



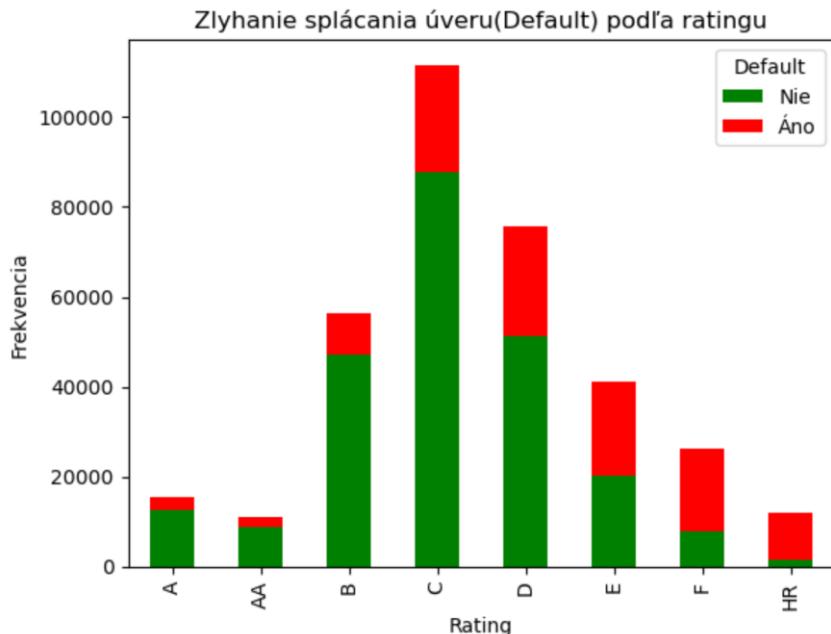
Obrázok 11 – Výška úveru a mesačná splátka podľa stavu žiadosti

Žiadatelia o úver, ktorí si podali žiadosť na P2P platforme Bondora majú tendenciu podľa Obrázok 12 žiadať o úvery vo finančnej hodnote **do 5000€**.



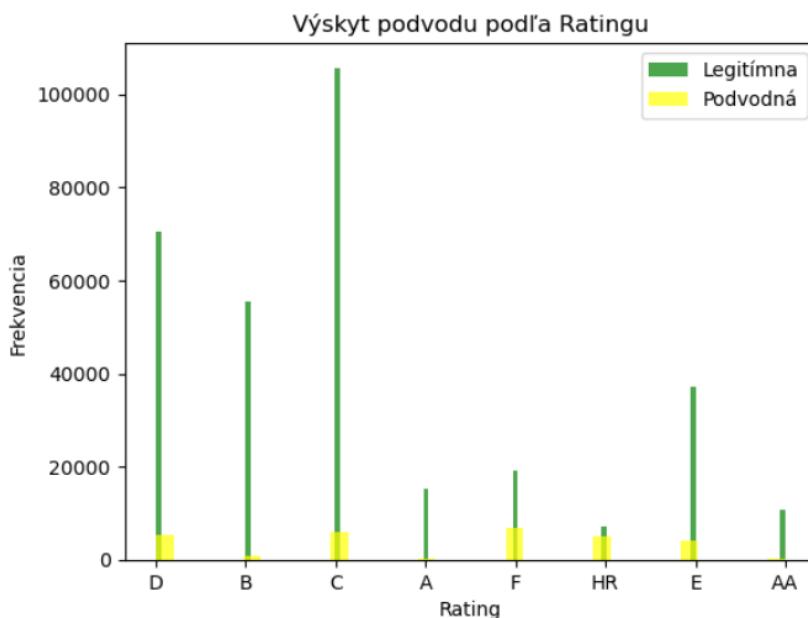
Obrázok 12 – Rozloženie výšky úverov

Graf na Obrázok 13 jasne vyjadruje, že úverové žiadosti s nižším ratingom majú veľký predpoklad dostať sa do stavu zlyhania splácania úveru. Hlavne pri kategórií F a HR je počet úverov v stave zlyhania splácania úveru zreteľne vyšší ako počet úverov ktoré v tomto stave nie sú.



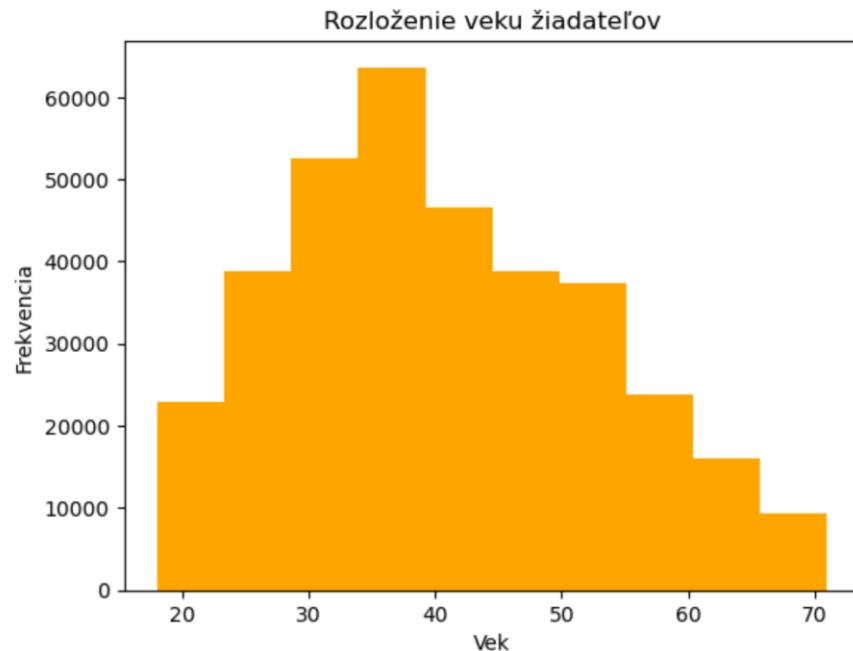
Obrázok 13 – Zlyhanie splácania úveru podľa ratingu

Fakt, že Rating má nezanedbateľný vplyv na odhad toho, aký stav dosiahnu úverové žiadosti potvrzuje aj Obrázok 14, ktorý zobrazuje rozdelenie výskytu podvodnej žiadosti podľa Ratingu.



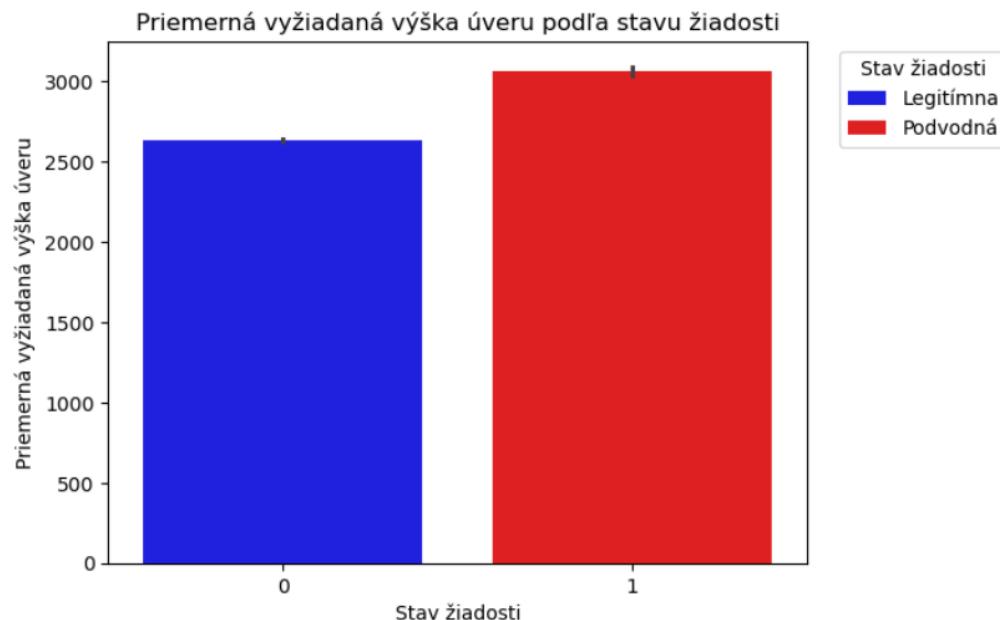
Obrázok 14 – Výskyt podvodu podľa Ratingu

Najväčšou skupinou ľudí, ktorý si požičiavajú na P2P platforme Bondora sú ľudia vo veku medzi 30 a 40 rokov života. Požičiavajú si aj ľudia s vyšším vekom, až približne do 70 rokov, najstarší žiadateľ má 71 rokov. Rozloženie vekových skupín je na Obrázok 15.



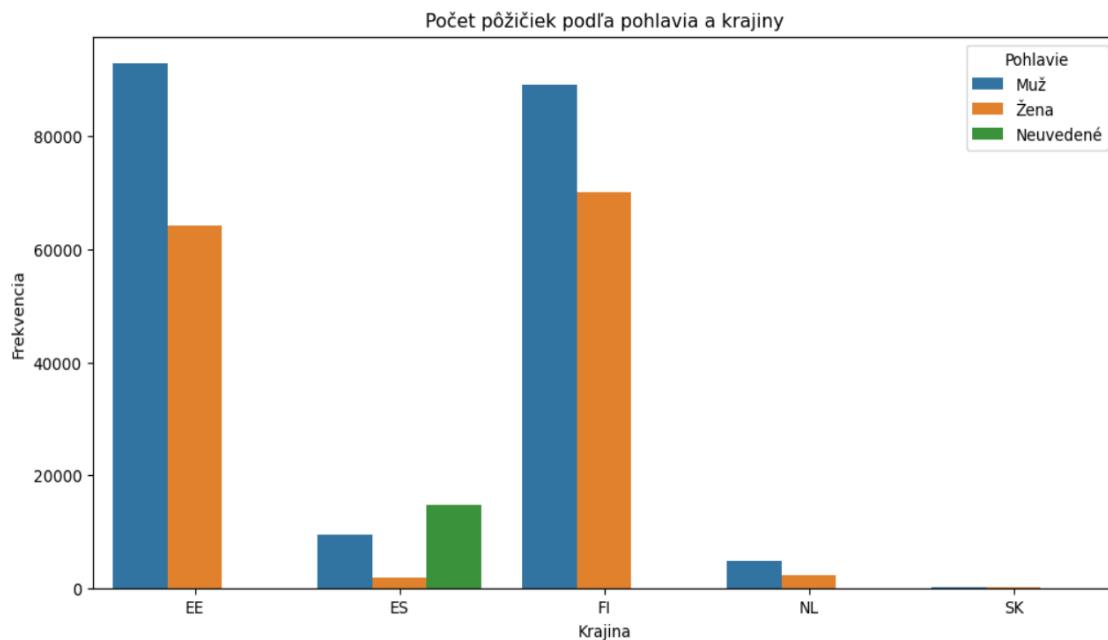
Obrázok 15 – Rozdelenie žiadateľov podľa veku

Obrázok 16 ukazuje, že priemerná výška úveru u podvodnej žiadosti je približne o 400 eur vyššia ako tá pri legitímej žiadosti.



Obrázok 16 – Priemerná výška úveru podľa stavu žiadosti

Najviac si podávajú žiadosť o úver muži z Estónka a Fínska, pričom viac žien má záujem o úver na Bondore z Fínska. Slovensko a Holandsko sú novo pridané krajiny, ktoré môžu žiadať o úver, zastúpenie úverových žiadostí je v porovnaní z ostatnými krajinami zanedbateľné. V Španielsku si najčastejšie svoje pohlavie neuvádzajú pri vyplňovaní úverovej žiadosti. Viac pozri Obrázok 17.



Obrázok 17 – Počet pôžičiek podľa pohlavia a krajiny

4.6 Škálovanie numerických hodnôt

Naše nezávislé numerické hodnoty majú rozdielne rozsahy hodnôt, ktoré môžu nadobúdať. Je preto vhodné ich zmeniť pred využitím na modelovanie predikčného modelu na jednotnú škálu. Škálovanie numerických hodnôt je obzvlášť potrebné pri modeloch založených na lineárnej alebo logistickej regresií z takého dôvodu, že model môže **prisudzovať väčšiu váhu** vyšším hodnotám, aj keď vyššie hodnoty nemajú väčší vplyv na správnu predikciu ako hodnoty nižšie. Naopak, algoritmy založené na báze rozhodovacích stromov nemajú problém z hodnotami rôznymi rozsahmi. Avšak rozhodovacím stromov proces škálovania nemôže uškodiť. **Nevýhodou** škálovania je často znížená vysvetliteľnosť, ľažká interpretácia hodnôt po procese škálovania. Knižnica Scikit-learn obsahuje viacero spôsobov škálovania, ako je proces **štandardizácie** pomocou StandardScaler [23] alebo **normalizácie** pomocou MinMaxScaler[23]. Proces transformácie týmto metódami je na Obrázok 18. V našej práci sme použili metódu **PowerTransformer** [23], ktorý dáta škáluje aby boli viac podobné **Gaussovému rozdeleniu**. S touto technikou sme dosiahli suverénne najlepšie výsledky.

StandardScaler

$$z = \frac{x - \mu}{\sigma}$$

MinMaxScaler

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Obrázok 18 – Transformácie hodnôt pomocou StandardScaler a MinMaxScaler [23]

4.7 Kódovanie (encoding) kategorických premenných

Všetky **kategorické premenné** musia byť **zakódované** predtým, ako ich môže model strojového učenia použiť. Je vhodné zakódovať aj kategorické premenné, ktoré sú už identifikované numerickou hodnotou aby bolo zachované rovnaké spracovanie pre všetky kategorické premenné daného typu. **Rozdielne kódovanie** musí byť keď ide o **ordinálnu** alebo **nominálnu** hodnotu. Ordinálne hodnoty majú určenú postupnosť podľa ktorej sú zoradené. V takýchto prípadoch má napríklad kategória

zakódovaná pod číslom 3 väčšiu váhu ako kategória zakódovaná pod číslom 1. Na takéto kódovanie využíva Scikit-learn **OrdinalEncoder** [23]. Tento encoder zakóduje hodnoty podľa ich určeného poradia od 0 až n-1 kategórií – hodnôt. Pre **závislé kategorické premenné** sa používa **LabelEncoder**, ktorý taktiež zakóduje každú hodnotu patriacu pod závislú premennú na hodnotu 0 až n-1 hodnotu [24]. Pre **nominálne** hodnoty, ktoré nemajú určenú postupnosť je populárny **OneHotEncoder** [23], ktorý zakóduje každú hodnotu do binárnej maticovej reprezentácie. Určí v stĺpci jednotku tam, kde je zakódovaná daná hodnota a zvyšok budú nuly. Pozri Tabuľka 3.

Tabuľka 3 – Kódovanie pomocou OneHotEncoding

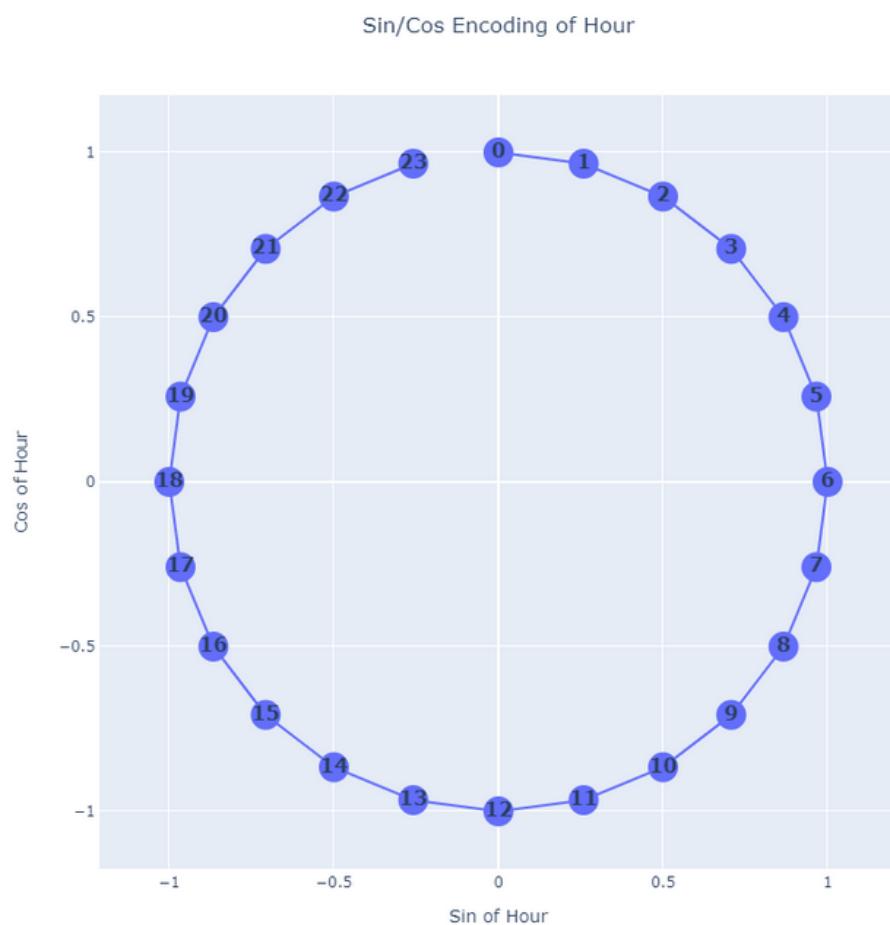
Pôvodný stĺpec	Transformované stĺpce		
Farba	Červená	Modrá	Zelená
Červená	1	0	0
Modrá	0	1	0
Zelená	0	0	1

Ako vidíme, na zakódovanie 3 kategórií jednej nezávislej premennej potrebujeme 3 nové stĺpce. Náš dataset obsahuje 13 ordinálnych premenných v rozsahu 2 až 11 kategórií. V priemere je to takmer 85 nových stĺpcov. Aj po odstránení pôvodných, nezakódovaných hodnôt je to obrovské číslo v už tak rozsiahлом datasete. V bakalárskej práci sme preto využili **BaseN encoder** [25], ktorý kategórie **zakóduje do jednotkovej sústavy**, ktorú mu určíme. BaseN encoder s bázou 1 je totožný s OneHotEncoderom. S bázou dva je ekvivalentný binárnej sústave. Čiže na pokrytie premennej so 7 kategóriami potrebuje vytvoriť na zakódovanie každej z nich dokopy 3 stĺpce. My sme si zvolili **BaseN encoder s bázou 4**, aby sme čo najmenej expandovali náš dataset, ale zároveň aby nenastala implicitná postupnosť medzi zakódovanými kategóriami. V takomto prípade vzniknú 2 nové stĺpce (po odstránení pôvodného nezakódovaného) pri zakódovaní premennej s 11 kategóriami. BaseN encoder je stredná cesta medzi OrdinalEncoderom a OneHotEncoderom. Má väčšiu mieru postupnosti medzi kategóriami oproti OneHotEncoder-u ale na rozdiel od neho viac šetrí miesto, nie však na takej úrovni ako OrdinalEncoder. Ďalšími spôsobmi zakódovania premenných môže byť **hashovanie** kategórií. Tieto hashovacie encodery [26] zahashujú kategóriu do čísla vopred určenej dĺžky, kde jedno číslo sa rovná jeden stĺpec t.j. ak je dĺžka hashu 5, tak je vytvorených 5 stĺpcov určených na zakódovanie každej kategórie danej premennej. Takto môže vzniknúť ľubovoľný počet stĺpcov, závisí od prvotného nastavenia. Nevýhodou tohto spôsobu sú kolízie, ktoré nastávajú

pri nízkej dĺžke hashu a tým vznikajúca strata interpretácie, lebo dochádza k tomu, že algoritmus priradí rovnaké číslo rôznym kategóriám. Výhodou je redukcia dimenzionality výslednej matice kódovania, pri vhodne zvolenom počte dĺžky hashu oproti metódam ako je OneHotEncoding alebo BaseN encoding. Populárne sú aj **Target encodery** [23], ktoré zakódujú kategóriu tak, že pri tvorbe kódovania zohľadňujú hodnotu závislej premennej. Nevýhodou je únik dát (data leakage), kde zakódovaná premenná má prístup k informáciám o závislej premennej. Výhodou je dimenzionalita, ktorá je na úrovni OrdinalEncoder-u.

Poslednou kategóriou sú premenné ktoré sú **cyklickej natury** ako sú hodiny, dni v mesiaci, dni v týždni [27]. Tieto hodnoty sa nedajú úplne vyjadriť ako ordinálne ani ako nominálne hodnoty. Na ich zakódovanie sme použili kódovanie pomocou sínusu a kosínusu, ktoré reprezentujú cyklickosť týchto časových hodnôt (zobrazenie na Obrázok 19). Vznikne nám jeden nový stĺpec pre každú premennú.

Stav po všetkých transformáciách je **349813 záznamov a 74 premenných**.



Obrázok 19 – Zakódovanie hodín pomocou sínusu a kosínusu[27]

5 IMPLEMENTÁCIA MODELOV

Hlavné rozdelenie algoritmov, ktoré využívajú učenie s učiteľom, ktoré aplikujeme v našej bakalárskej práci je na klasifikačné a regresné problémy. Naša práca je **klasifikačný problém** zameraný na detegovanie podvodných úverových žiadostí, kde sa snaží algoritmus rozdeliť dataset do určitých **diskrétnych** tried, v našom prípade na Fraud(1 – podvod) / Not Fraud (0 – nie je podvod / legitímna žiadosť). Pri regresných problémov sa predikuje spojité premenné na rozdiel od presne zadefinovaných tried ako pri klasifikačných problémoch. Pri regresných problémov sa využívajú napríklad modely lineárnej regresie alebo aj LASSO regresia(Least Absolute Shrinkage and Selection Operator). Pri klasifikačných problémoch sú to modely **logistickej regresie, rozhodovacích stromov, náhodných lesov** a iné. Vymenované modely klasifikačných problémov sme využili aj v našej práci [28].

5.1 Rozdelenie vstupných dát, Overfitting a Underfitting

Predtým, než sa model začne učiť zo vstupných dát je vhodné ich rozdeliť na trénovaciu, validačnú a testovaciu množinu.

Na **trénovacej množine** sa model naučí a vyhodnotí skryté vzťahy medzi premennými a potom na ich základe robí predpoveď a rozhodnutia.

Validačná množina sa využíva na dolaďovanie vstupných **hyperparametrov** modelu, ktorých cieľom je zlepšiť predikčné schopnosti modelu na dátach, na ktorých neboli natrénované, aby sa priblížilo nastavenie modelu na formu vhodnú na aplikovanie predikcie na reálnych, nových dátach.

Testovacia množina predstavuje reálne dátá, kde schopnosť predikcie modelu reflekтуje skutočné predikčné schopnosti vytvoreného modelu na detekciu podvodných úverových žiadostí.

V praxi sa bežne používa rozdelenie v pomere 70/15/15, 60/20/20 a podobne pre trénovaciu/validačnú/testovaciu množinu.

Overfitting alebo pretrénovanie [29] je situácia, ktorá vzniká, ak sa model natrénuje na trénovacích dátach do takej miery, že nedokáže vhodne predikovať nové, nevidené dátá. Znamená to, že model stratil generalizačnú schopnosť predikcie. Môže vzniknúť vtedy, ak vstupné dátá sú príliš komplexné a obsahujú veľa dát, ktoré sú pre predikciu daného problému nepodstatné. Overfitting je reprezentovaná dobrými predikčnými schopnosťami na trénovacej množine ale zlými výsledkami predikcie na testovacej množine. Predchádzať danej situácií môžeme odstránením nepodstatných nezávislých premenných alebo použitím krízovej validácie. **Krízová validácia** je

popísaná v podkapitole 5.3. Taktiež je vhodné pri škálovaní numerických hodnôt a kódovaní kategorických premenných natrénovať transformačnú metódu iba na trénovacej množine a aplikovať až vyhodnotené poznatky metódy na obe množiny (transformovať ich podľa naučených poznatkov). Týmto spôsobom sa zamedzí úniku dát (data leakage), ktorý by vznikol, ak by sa trénovali tieto metódy aj na testovacích dátach. Ak by nebol tento postup dodržaný, môže dôjsť k overfittingu a predikcie modelu budú nereprezentatívne jeho skutočným schopnostiam predikcie z dôvodu, že mal prístup k testovacej množine, ktorá by mala byť fungovať ako úplne neznáma množina dát.

Underfitting alebo nedostatočné natrénovanie [29] je opačná situácia, ku ktorej môže dôjsť, ak vzorka dát je príliš malá a nedostatočne reprezentuje reálnu situáciu, na ktorú bol prediktívny model aplikovaný. Ak je model príliš jednoduchý a bol natrénovaný na malom počte vstupných dát, tak má model problém vytvoriť si generalizačné prediktívne správanie, ktoré by mal aplikovať na predikovanie nových dát. Túto situáciu môžeme riešiť zvýšením komplexnosti modelu, napríklad vytvorením nových relevantných nezávislých premenných, ktoré pomôžu zlepšiť predikčné schopnosti, ktorých účelom je správna predikcia závislej premennej. Taktiež môžeme zvýšiť počet získaných vzoriek alebo sa zamerať na zber dát o ďalších iných, relevantných vlastnostiach skúmaného systému.

5.2 Hyperparametre modelu

Každý model strojového učenia má určité hyperparametre, ktoré nám dovoľujú pozmeniť nastavenia modelu, ktoré sa používajú počas procesu učenia. **Hyperparameter** je napríklad počet stromov v náhodnom lese, vnútorný algoritmus použitý pri logistickej regresii alebo maximálna hĺbka, do ktorej sa bude rozhodovací strom rozdeľovať. Vybranie správnych hyperparametrov môže zlepšiť výkon prediktívneho modelu, zároveň však nesprávne nastavené hyperparametre môžu brániť modelu dosiahnutie jeho skutočných prediktívnych schopností.

Výber vhodných hyperparametrov je veľmi nákladná činnosť, kde je veľkým vplyvom kvalita výpočtovej techniky, ktorá aplikuje metódy na vyhľadanie najlepšej kombinácie hyperparametrov. Výpočtovou technikou môže byť napríklad procesor alebo grafická karta ,ich výkon môžu ovplyvniť aj pamäťové média ako je pevný disk, SSD disk, pamäť ram a iné. Časová náročnosť je ovplyvnená touto výpočtovou technikou, spôsobom výberu hyperparametrov metódami vyhľadávania najlepšieho hyperparametra a počtom skúmaných hyperparametrov.

Scikit-learn poskytuje implementáciu GridSearch a RandomizedSearch ako metódy určené na vhodný výber hyperparametrov [30].

GridSearch preskúma každú kombináciu hyperparametrov zo zoznamu hyperparametrov, ktoré mu boli zadané na prehľadávanie. Poskytne exaktné, najlepšie riešenie tvorené kombináciou vstupných hyperparametrov, ktorému boli zadané na prehľadávanie. Tento prístup je veľmi časovo náročný, hlavne keď GridSearch má prehľadávať veľký počet kombinácií hyperparametrov. GridSearch pri veľkom počte kombinácií, môže trvať hodiny až dni a mesiace.

RandomizedSearch na rozdiel od GridSearch neskúša každú kombináciu hyperparametrov, ale vyberá z nich náhodne, tak aby vytvoril funkčnú kombináciu. Kedže postupuje náhodne, môže dosiahnuť neoptimálnu kombináciu hyperparametrov. Avšak preto, že neskúša každú kombináciu hyperparametrov, tak je oveľa rýchlejší ako GridSearch. Je vhodný najme pri veľkom počte kombinácií hyperparametrov, kde kompletné prehľadávanie pomocou GridSearch bolo veľmi časovo náročné. Metódy GridSearch a RandomizedSearch je vhodné použiť spoločne, najprv nájsť sľubné kombinácie hyperparametrov náhodným výberom pomocou RandomizedSearch a potom do oblasti, kde sa nachádzajú najsľubnejšie parametre sústrediť GridSearch, ktorý prehľadá celý tento sľubný priestor.

GridSearch a RandomizedSearch by mali byť aplikované na validačnú množinu, aby nedošlo k pretrénovaniu (overfitting) na trénovacích dátach a zároveň aby nedošlo k úniku dát, ku ktorému by mohlo dôjsť, ak by sme aplikovali GridSearch a RandomizedSearch na testovaciu množinu.

5.3 K-násobná krízová validácia (K-fold cross validation)

Krízová validácia [31] je metóda, pri ktorej sa dáta rozdelia na určitý počet **k-násobných (k-fold)** celkov, kde sa model následne trénuje na k-1 celkoch a testovanie modelu sa vykonáva na zvyšnom celku. Napríklad v prípade 10-násobnej krízovej validácie sa rozdelia dáta na 9 celkov, na ktorých sa bude model trénovať a 1 celok, ktorý poslúži ako testovací celok. Následne sa testovací celok zmení na iný a toto prebieha celkovo 10 krát, až kým sa každý celok nevystrieda ako testovací celok (pozri Obrázok 20, zelenou je testovací celok). **Výsledné metriky**, ktoré boli aplikované na testovaciu množinu v každej iterácii k-násobnej validácie **sú následne spriemerované**, aby poskytli všeobecný odhad efektivity modelu na predikovanie podvodných úverových žiadostí. Vďaka tejto metóde sa predchádza pretrénovaniu modelu a zároveň sa odstraňuje náhodnosť, pri ktorej môže dôjsť k šťastnému výberu testovacej množiny, ktorá môže vzniknúť pri klasickom postupe, kde sa vyberie iba

jeden celok, ktorý bude predstavovať testovaciu množinu. Keďže pri k-násobnej validácii sa stáva testovacou množinou každý s k celkov, nemôže dôjsť k šťastnému výberu, ktorý by skresľoval skutočné predikčné schopnosti modelu. Táto metóda má aj však svoje nevýhody a to je k-násobne zväčšená časová náročnosť, keďže proces trénovalia a následného testovania prebieha k-krát.

Zároveň sa dá K-násobná krízová validácia uplatniť aj pri vyhľadávaní najlepších hyperparametrov, kde vďaka k iteráciám testovacej množiny, ktorou sa stáva každý celok z k celkov, odpadá potreba validačnej množiny. Týmto spôsobom nemôže dôjsť k pretrénovaniu a zároveň odpadá starosť o únik dát, pretože výsledne metriky sú spriemerované a najlepšie nastavenie hyperparametrov je vybrané na základe týchto spriemerovaných metrík.

Z dôvodu **nerovnováhy** nášho datasetu, kde podvodné žiadosti predstavujú **8.14%** z celkového počtu žiadostí, musíme použiť **stratifikovanú** K-násobnú krízovú validáciu, ktorá bude zohľadňovať túto skutočnosť a rovnomerne rozdelí naše podvodné žiadosti do každého celku.

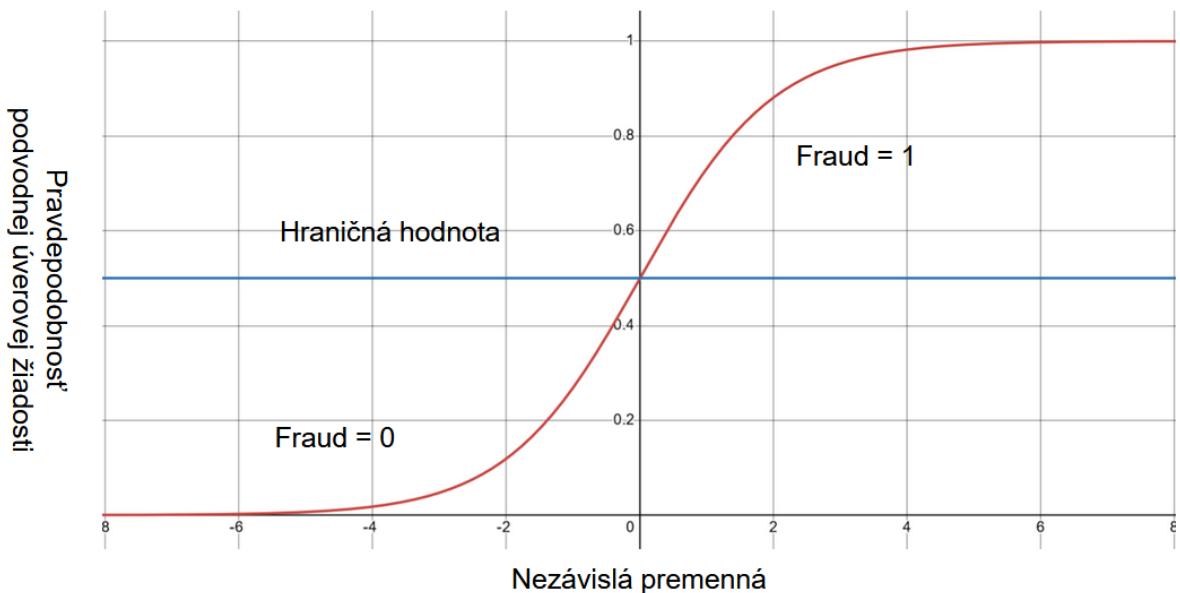


Obrázok 20 – Rozdelenie dát podľa 10-násobnej krízovej validácie

5.4 Aplikované algoritmy na tvorbu modelu strojového učenia

5.4.1 Logisticá regresia

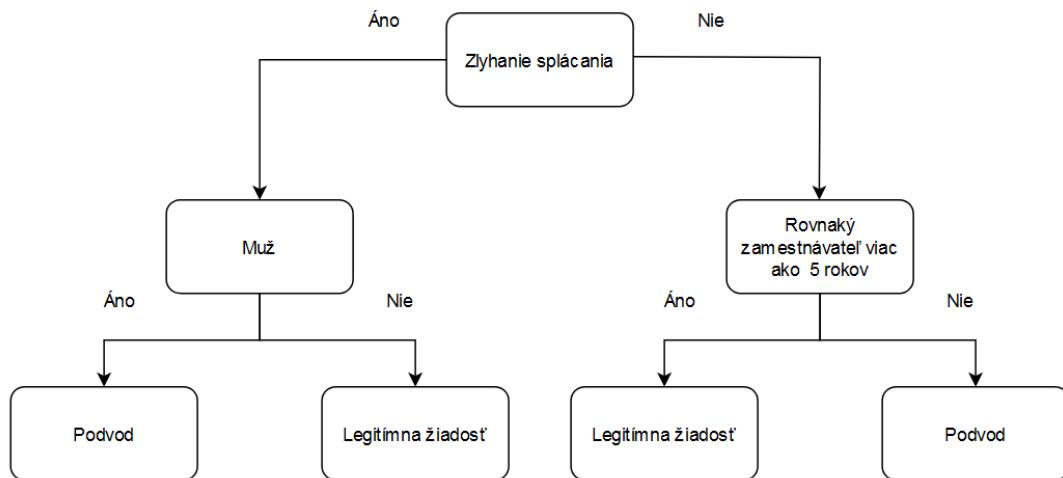
V doméne strojového učenia je **logisticá regresia** [32] algoritmus strojového učenia, ktorý sa snaží predikovať hodnotu závislej premennej na základe nezávislých premenných. Dokáže predikovať iba diskrétnu hodnotu z dvoch kategórií (alebo viac v prípade multinomickej logistickej regresie), môže byť napríklad 0 alebo 1, legitímna žiadosť alebo podvod. Používa sa na klasifikačné problémy. Model vytvára rovnicu, kde každej nezávislej premennej priradí koeficient a následne aplikuje **logisticú funkciu** (pozri Obrázok 21), ktorej výsledkom je hodnota pravdepodobnosti medzi 0 a 1. Model má nastavenú hraničnú hodnotu, zvyčajne 0.5. Ak výsledná pravdepodobnosť je väčšia ako hraničná hodnota, tak je priradená do jednej kategórie(1) inak, ak je menšia alebo rovná tejto hodnote tak do druhej(0). Model, ktorý využíva logisticú regresiu sa snaží nájsť najlepšie hodnoty koeficientov pre nezávisle premenné tak, aby sa maximálne zvýšila pravdepodobnosť správneho priradenia kategórie závislej premennej podľa skutočnosti. Hodnoty koeficientov každej nezávislej premennej sú vypočítane pomocou algoritmu **gradientového zostupu**, ktorý je aplikovaný na chybovú funkciu, ktorá hovorí ako moc (ne)vhodné sú aktuálne váhy priradené k nezávislým premenným. Gradient tejto chybovej funkcie ukazuje smer najvyššieho rastu chybovej funkcie, avšak model sa snaží minimalizovať túto funkciu, preto gradientový zostup používa opačný smer (antigradient). Koeficienty sa aktualizujú týmto spôsobom až pokiaľ model nedokáže viac zlepšiť presnosť modelu.



Obrázok 21 – Logisticá funkcia

5.4.2 Rozhodovací strom

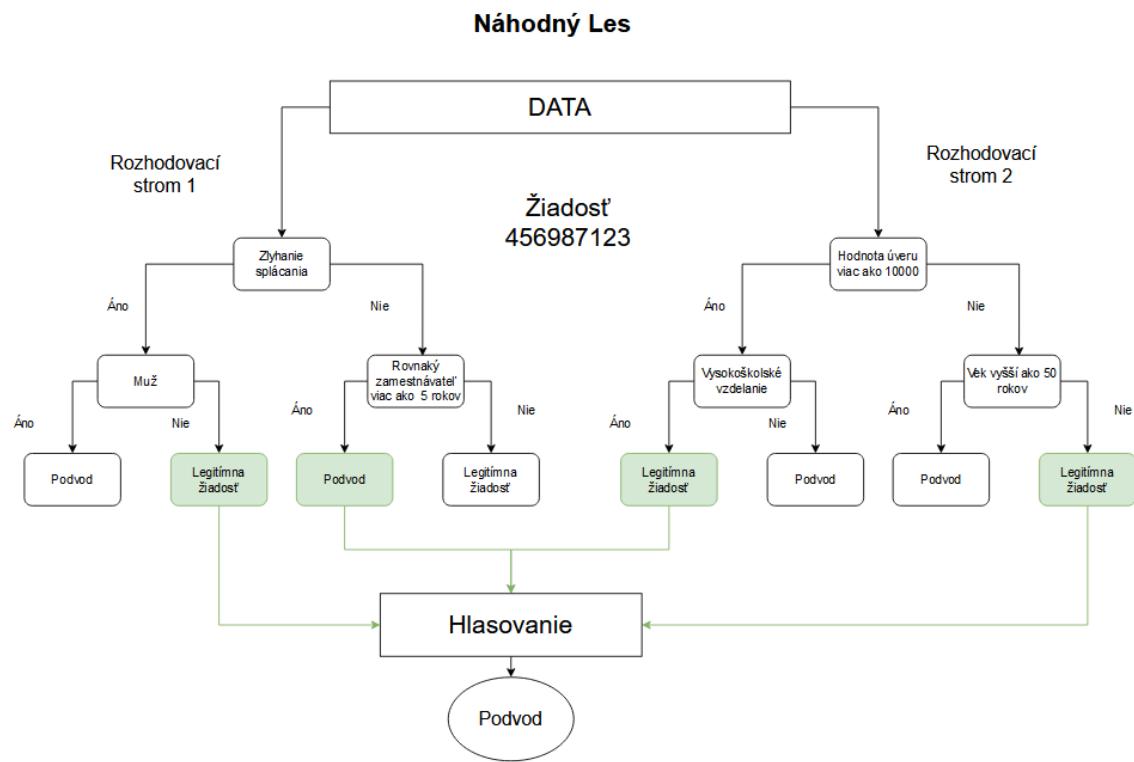
Rozhodovací strom [33] je algoritmus, ktorý sa rozhoduje podľa jednoduchej binárnej logiky. Rozhodovací strom začína koreňom, kde sa vyberie **nezávislá premenná** na ktorú je aplikovaná nejaká **podmienka**. Podľa odpovede sa strom ďalej rozdeľuje, vytvára vetvy podľa odpovedí a na týchto vetách sa nachádzajú uzly – miesta kde je ďalšia podmienka, podľa ktorej sa bude rozhodovací strom ďalej vetviť, až kým sa nepríde k listom stromu, ktoré predstavujú rozhodnutie na základe predošlých odpovedí. V našom prípade to je legitímna žiadosť alebo podvod. Celú hierarchiu rozhodovacieho stromu pozri Obrázok 22. Rozhodovací strom môže byť použitý na klasifikačné i regresné problémy na rozdiel od logistickej regresie.



Obrázok 22 – Hierarchia rozhodovacieho stromu

5.4.3 Náhodný les

Náhodný les [34] patrí pod takzvané **ansámblové metódy**. Tieto metódy používajú viaceré **klasifikátory**, aby dosiahli lepšiu presnosť výsledkov. V prípade náhodného lesa sú týmito klasifikátormi **rozhodovacie stromy**. Algoritmus náhodného lesa vytvorí určitý počet rozhodovacích stromov a každému stromu pridelí náhodnú časť dát a náhodnú množinu nezávislých premenných (vlastnosti úverovej žiadosti), ktorú bude rozhodovací strom spracovať svojim algoritmom. Následne po spracovaní dát stromami je zahájené **hlasovanie**, kde výsledné označenie žiadosti o úver je určené podľa toho, koľko hlasov získalo dané rozhodnutie – legitímna žiadosť alebo podvod. Každý rozhodovací strom má **jeden hlas** a rozhodnutie závisí podľa **najčastejšej** predikcie. Ukážka náhodného lesa aplikovaného na náš dataset pozri Obrázok 23.



Obrázok 23 – Náhodný les

6 PREZENTÁCIA VÝSLEDKOV

Pri určovaní podvodných žiadostí musíme zvážiť závažnosť nesprávnej predikcie. Z dôvodu nevyváženého počtu podvodných a legitímnych žiadostí v našom datasete ale aj v realite, kde taktiež dochádza o niekoľkonásobne viac k legitímnych žiadostiam ako tým podvodným je nutné zohľadniť to, že väčšina algoritmov strojového učenia pracuje na báze predpokladu, že počet hodnôt, ktoré môže závislá premenná nadobúdať je vyrovnaný. V takomto prípade má tendenciu model, ktorý obsahuje nevyvážený počet hodnôt závislej premennej dosahovať vysokú hodnotu presnosti, čo môže mylne nadobudnúť dojem vysokej prediktívnej schopnosti modelu. Treba si však uvedomiť, ak náš model obsahuje 100 000 žiadostí a z toho 90% je legitímnych a náš model bude predikovať všetky žiadosti ako legitímne, tak presnosť, ktorá je obvykle pokladaná za najhlavnejšiu metriku takéhoto modelu bude vysokých 90%. Preto je potrebné vyhodnocovať prediktívne schopnosti modelu inými metrikami ako sú **confusion matrix, senzitivita, precíznosť** a **F1 skóre**.

Confusion Matrix (matica zámen) [35] je matica, ktorá dokáže vyhodnotiť úspešnosť modelu pri klasifikačných problémoch. Ak klasifikuje binárnu závislú premennú, tak sa skladá z dvoch riadkov a dvoch stĺpcov. V prvom riadku a stĺpci sa nachádza počet **True negative (TN)** predikcií. Toto sú také predikcie, ktoré predikuje model ako legitímne úverové žiadosti a sú legitímne aj podľa skutočnosti. V prvom riadku a druhom stĺpci je počet **False positive (FP)** predikcií. Tieto predikcie sú také, ktoré model predikuje ako podvodné ale v skutočnosti sú legitímne. **False negative (FN)** predikcie sa nachádzajú v druhom riadku, prvý stĺpec. Tieto hodnoty sú také, ktoré model vyhodnotil ako legitímne ale v skutočnosti sú to podvodné žiadosti. Počet **True positive (TP)** predikcií sa nachádza v poslednom riadku a stĺpci. Táto hodnota hovorí koľko úverových žiadostí bolo predikovaných ako podvodných a aj v skutočnosti to sú podvodné úverové žiadosti. Confusion matrix je obzvlášť dôležitá pri nevyváženom počte závislej premennej, pretože prehľadne zobrazuje ako model predikoval v každom prípade.

Presnosť (Accuracy) [35] – najzákladnejšia metrika v prípade, že sa nejedná o nevyvážený dataset. Táto metrika hovorí, o pomere správnej predikcie podvodných a legitímnych úverových žiadostí.

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

Precíznosť (Precision) [35] – táto metrika reprezentuje pomer medzi správne predikovanými podvodnými žiadostami proti všetkým žiadostiam, ktoré model označil ako podvodné. Hovorí o tom, ako dobre náš model dokáže zredukovať počet zamietnutých úverových žiadostí kvôli nesprávnemu vyhodnoteniu žiadosti ako podvodnej.

$$Precíznosť = \frac{TP}{TP + FP}$$

Senzitivita(Recall alebo Sensibility) [35] – metrika, ktorá je pre nás veľmi dôležitá, keďže cena za neodhalenie podvodu vo finančnom sektore je vysoká, lebo je ohrozená reputácia spoločnosti aj financie veriteľov. Snažíme sa dosiahnuť čo najlepší pomer. Táto metrika reprezentuje pomer správne predikovaných podvodných žiadostí voči všetkým skutočným žiadostiam, ktoré sú podvodné.

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

F1 skóre [35] – ďalšia metrika, ktorá je veľmi dôležitá pri datasetoch kde nie je rovnováha vo výskytu závislej premennej. Táto metrika je dôležitá lebo spája informácie získane vďaka precíznosti a senzitívity – je to ich harmonický priemer, ktorý ma vlastnosť, že čím viac sú od seba precíznosť a senzitivita od seba odlišné, tým horšie je F1 skóre. V prípadoch, keď je dataset nevyvážený najlepšie reflektuje skutočné predikčné vlastnosti modelu.

$$F1 = \frac{2 * precíznosť * senzitivita}{precíznosť + senzitivita} = \frac{2 * TP}{2 * TP + FP + FN}$$

Špecificka(Specificity) [35] – Táto metrika reprezentuje pomer správne predikovaných legitímnych žiadostí voči všetkým skutočným žiadostiam, ktoré sú legitímne.

$$\check{S}pecificka = \frac{TN}{TN + FP}$$

Vyvážená presnosť(Balanced Accuracy) [36] – je to ďalšia metrika, ktorá je vhodná pre klasifikačné problémy s nevyváženou závislou premennou v datasete, ako je náš dataset LoanData.csv. Je to vlastne aritmetický priemer senzitivity pre všetky triedy závislej premennej. Ak sa jedná o binárnu závislú premennú, tak sa dá vypočítať aj ako aritmetický priemer senzitivity a špecificity.

$$\begin{aligned} \text{Vyvážená Presnosť} &= \frac{\sum_{i=1}^n \text{Senzitivita}_i}{n} \\ &= \frac{\text{Senzitivita} + \text{Špecificita}}{2} \end{aligned}$$

Makro priemer F1 skóre [37] – táto metrika je vhodná pri nevyvážených datasetoch a pri nebinárnych klasifikačných problémoch (predikujeme viac ako dve hodnoty - triedy závislej premennej), lebo sčíta F1 skóre pre každú triedu predikovanej závislej premennej a túto sumu podelí počtom tried. Makro priemer zabezpečuje, že menej početná trieda rovnakým podielom participuje na výslednom hodnotení predikčných schopností modelu, ako trieda s vyššou frekvenciou výskytu.

$$\begin{aligned} F1_{makroPriemer} &= \frac{\sum_{i=1}^n F1_i}{n} \\ &= \frac{F1_{legitímnažiadost} + F1_{podvodnážiadost}}{2} \end{aligned}$$

Vážený priemer F1 skóre [37] – Využiteľnosť tejto metriky je v modeloch, ktoré by mali byť veľmi výkonné pri správnom predikovaní tried (hodnôt, ktoré môže závislá premenná nadobúdať), ktoré sa vyskytujú často a ich predikčné schopnosti môžu byť menej úspešne pri triedach, ktoré sa vyskytujú menej frekventované. Je to tým, že táto metrika vynásobí každé F1 skóre danej triedy jej výskytom (w_i) v datasete. Vďaka tomu, dokáže byť vážený priemer F1 skóre vyššie ako klasické F1, ak správne predikuje frekventovanejšie triedy. Pri vyváženom počte hodnôt nezávislej premennej je zas veľmi blízko hodnoty klasického F1 skóre. Táto metrika nie je pre náš prípad príliš smerodajná, vďaka obrovskej prevahy legitímných žiadostí oproti podvodným, čo udáva príliš veľkú váhu legitímnym žiadostiam.

$$F1_{váženýPriemer} = \sum_{i=1}^n (w_i) * F1_i \quad w_i = \frac{n_i}{n}$$

Beta F1 skóre [37] – táto metrika je modifikáciou F1 skóre, kde je pokladaný väčší dôraz buď na senzitivitu alebo precíznosť. Tento dôraz sa určuje podľa parametra β . Ak je β väčšia ako 1, senzitivita má väčší dôraz na výsledok. Napríklad, ak $\beta=3$, tak na senzitivitu je trojnásobne väčší dôraz ako na precíznosť. Hodnota β menšia ako 1 zvyšuje dôležitosť precíznosti voči senzitivite. Hodnota $\beta =1$ je rovná F1 skóre. Pre finančný sektor je dôležitejšie predísť podvodu, závažnosť označenia žiadosti ako podvodnej nie je až tak veľká ako vyhodnotiť žiadosť ako legitímnu. Preto v našom prípade dáva vyšší dôraz na senzitivitu väčší zmysel.

$$\begin{aligned}\beta F1 &= \frac{(1 + \beta^2) * \text{Precíznosť} * \text{Senzitivita}}{(\beta^2 * \text{Precíznosť}) + \text{Senzitivita}} \\ &= \frac{(1 + \beta^2) * TP}{(1 + \beta^2) * TP + FP + \beta^2 * FN}\end{aligned}$$

Matthewsov korelačný koeficient (MCC – Matthews correlation coefficient) [38] je ďalšia metrika, ktorá veľmi dobre zhodnocuje prediktívne schopnosti modelu v prípade, že dataset nie je vyvážený. Je využívaný v binárnych kategorických problémoch. Jeho výhoda oproti F1 skóre a jeho variantom je tá, že sa nezameriava iba na pozitívnu triedu, ale aj na negatívnu a tým poskytuje komplexný pohľad na prediktívne schopnosti modelu. Dosahuje vysoké hodnoty iba v prípade ak sú prediktívne schopnosti modelu nadstandardne dobré vo všetkých smeroch - zohľadňuje vplyv FN, FP, TP, TN. Táto metrika je vhodná ako ukazovateľ, ktorý vyjadruje ako je nás model vhodný pri identifikovaní legitímnych žiadostí.

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

ROC krivka a AUC

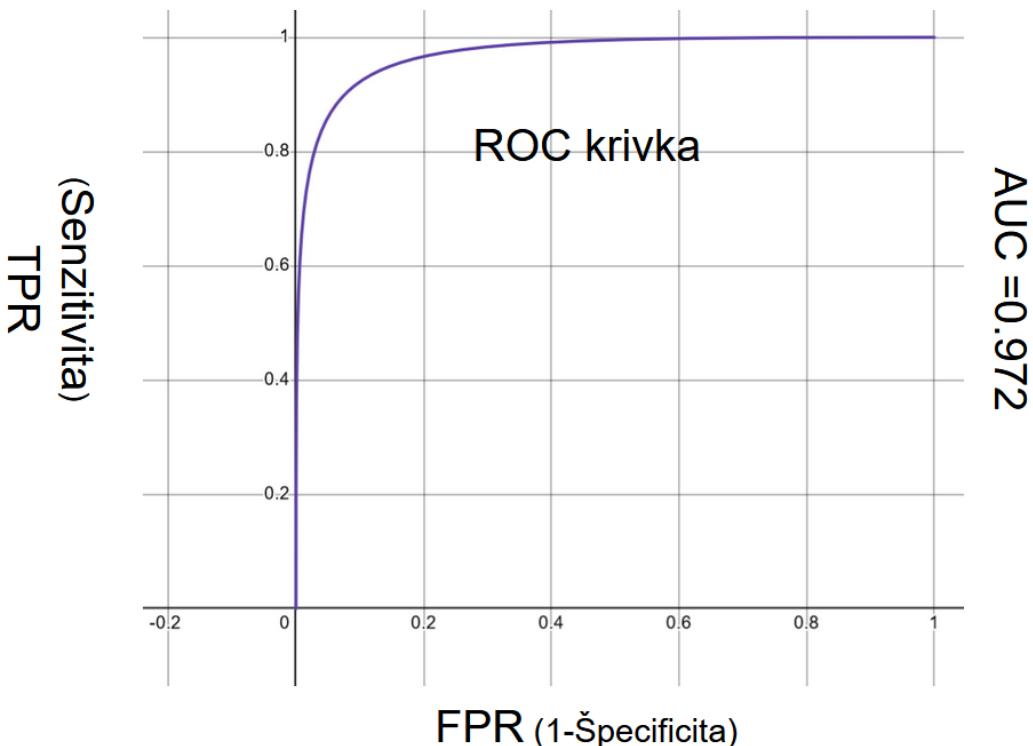
ROC [39] je grafické znázornenie, ktoré pomáha vyhodnotiť predikčné schopnosti klasifikačného binárneho modelu. Graf ROC krivky má na osi x hodnoty, ktoré predstavujú pomer falošne pozitívnych predikcií (FPR), ktoré dokážeme získať ako $1 - \text{špecificita} = \frac{FP}{TN+FP}$. Na osi y sa nachádza pomer pravdivých pozitívnych predikcií (TPR) alebo inak Senzitivita. ROC krivka znázorňuje predikčné schopnosti modelu v určitej hraničnej hodnote. Hraničná hodnota postupne dosahuje hodnoty v rozmedzí 0 a 1. Napríklad, pri menších hodnotách hraničnej hodnoty model frekventovane hodnotí úverové žiadosti ako podvodné, čo môže mať za následok, že veľa legitímnych žiadostí

bude hodnotených ako podvodné – falošne pozitívne. Naopak, pri väčšej hraničnej hodnote, model frekventovane vyhodnocuje úverové žiadosti ako legitímne, čo má za následok opačný efekt a to možný zvýšený počet úverov, ktoré boli označené ako legitímne – falošne negatívne.

Grafické znázornenie ROC krivky [39] je v Príloha H.

V pravom hornom rohu je hraničná hodnota najnižšia – je vždy nižšia ako všetky pravdepodobnosti predikcií, ktoré predpovedal model , čo znamená, že model všetko hodnotí ako podvodnú žiadosť, FPR a TPR sú vysoké. Následne sa táto hodnota zväčšuje a tým by sa mal zvyšovať pomer TPR a klesať pomer FPR (graf 3 v Príloha H). Ak však klesajú súčasne a tvoria priamku z pravého horného rohu do dolného ľavého rohu(diagonála), znamená to, že model nedokáže správne rozlíšiť podvodnú žiadosť od legitímnej žiadosti, robí náhodné predikcie (graf 2). Dokonalý prediktívny model by mal vyzeráť tak, že ROC krivka postupuje z pravého horného rohu do ľavého a potom strmo do ľavého dolného rohu(graf 1). Ak ROC krivka postupuje z pravého horného rohu strmo do pravého dolného rohu a následne priamo do ľavého dolného rohu znamená to, že model predikuje binárne kategórie presne opačne(predikuje 0 ako 1 a 1 ako 0, graf 4).

AUC [39] je metrika ktorá hovorí o ploche pod ROC krivkou, čím vyššia hodnota tejto metriky, tým lepšie. Hodnota 1 znamená, že model má dokonale presné predikcie, 0.5 značí, že predikcie sú náhodne, menšia hodnota ako 0.5 značí, že model má vymenenú klasifikáciu hodnôt závislej premennej. Obrázok 24 je ROC krivka s hodnotou $AUC = 0.972$. Za dobré klasifikačné schopnosti modelu sa pokladá model, ktorý dosiahol AUC s hodnotou 0.7 a vyššie.



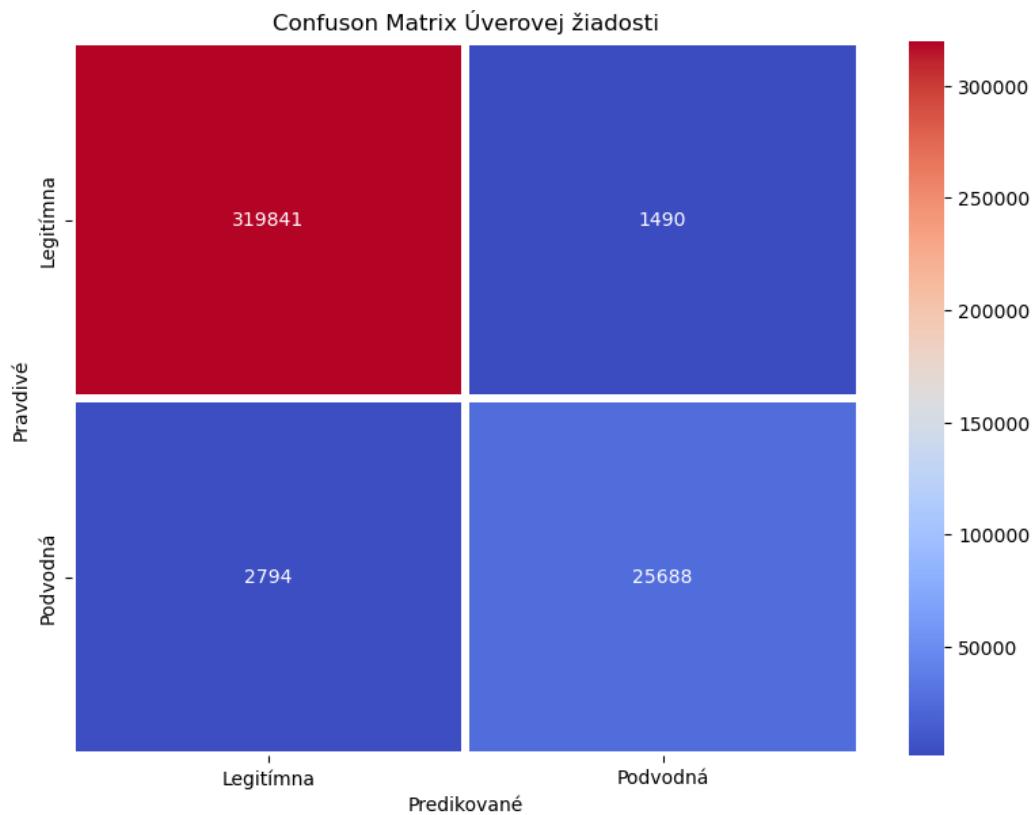
Obrázok 24 – ROC krivka s AUC v hodnote 0.972

6.1 Vyhodnotenie výsledkov modelov

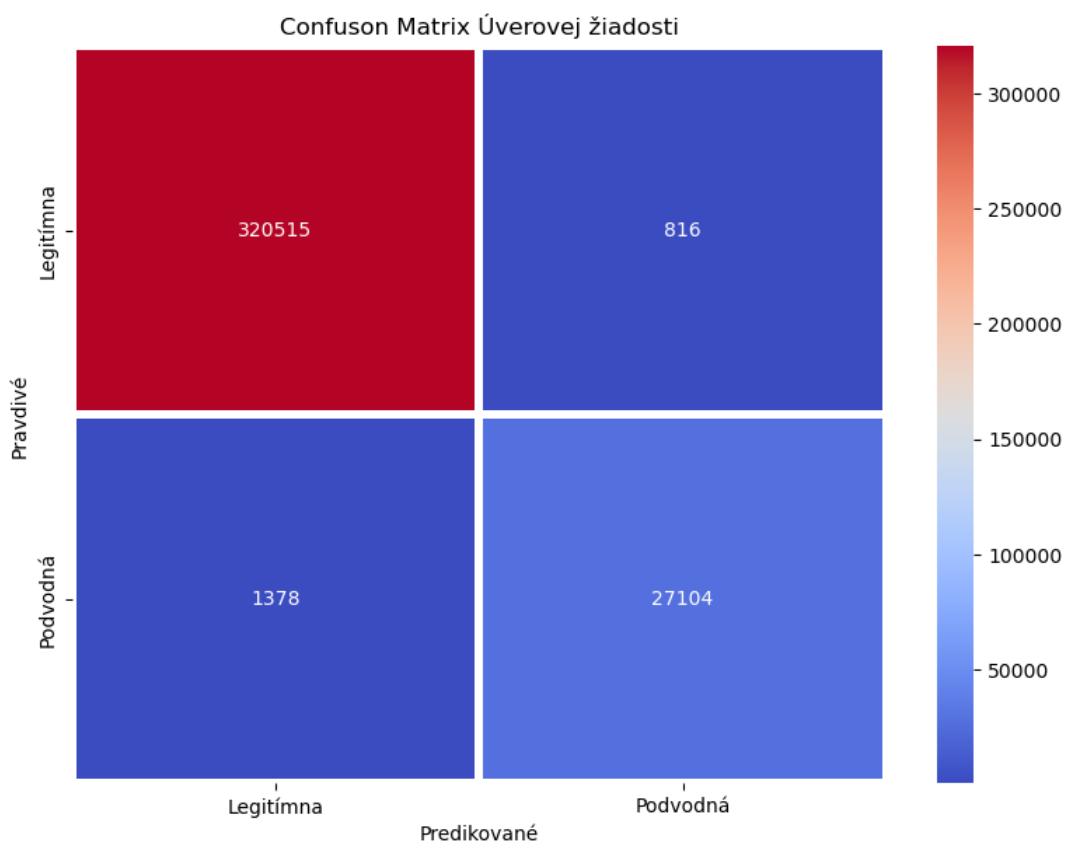
Na každú z modelov (Logistická Regresia, Rozhodovací strom, Náhodný les) bola aplikovaná 10-násobná krížová stratifikovaná validácia (pozri 0). V Príloha I sa nachádza nastavenie modelu pre logistickú regresiu, v Príloha J sa nachádza nastavenie modelu pre rozhodovací strom a v Príloha K sa nachádza nastavenie modelu pre náhodný les. Najvhodnejšie nastavenia modelu boli natrénované pomocou 5-násobnej krížovej validácie s prioritou na najlepšie beta F1 skóre. Pre každú model bola vytvorená confusion matrix, klasifikačný report, ROC krivka a vyhodnotenie najdôležitejších parametrov pre daný model.

6.1.1 Confusion Matrix (matica zámen)

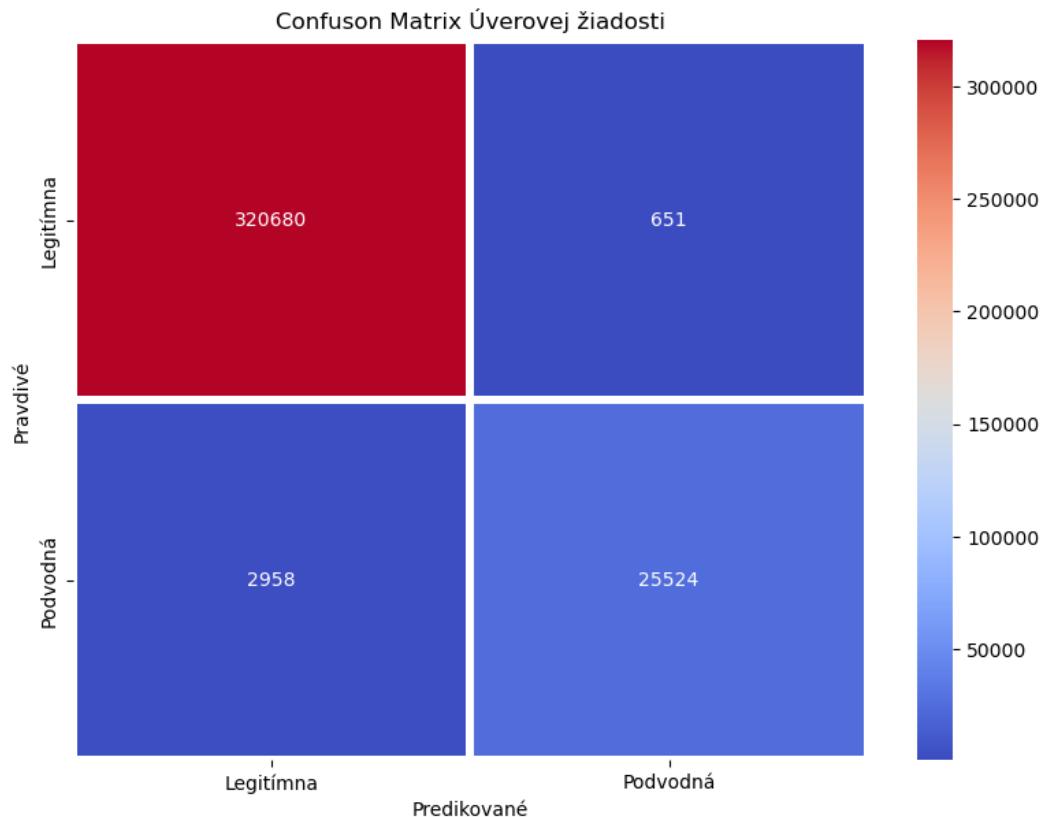
Na Obrázok 25 Obrázok 26 Obrázok 27 je zobrazená confusion matrix pre Logistickú regresiu. V ľavom hornom rohu sa nachádza počet TN predikcií, v pravom hornom rohu počet FP, v ľavom dolnom rohu FN a v pravom dolnom rohu TP. Modely dosahuje veľmi dobré výsledky pri určovaní **TN**, pričom najúspešnejší bol model **náhodného lesu**. Pri **FP** bol taktiež najviac úspešný **náhodný les**, ktorý mal iba 651 **FP** predikcií. Najmenej **FN** dosial **rozhodovací strom**. Pri určovaní **TP** sa žiadnen model nevyrovnal **rozhodovaciemu stromu**.



Obrázok 25 - Confusion Matrix pre Logistickú regresiu



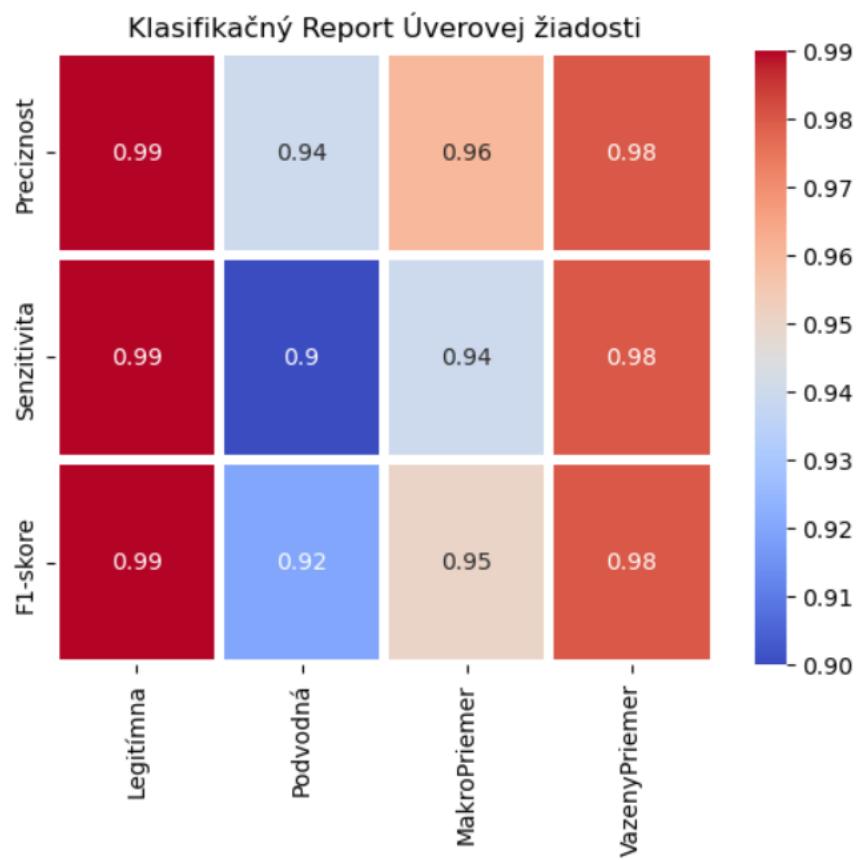
Obrázok 26 – Confusion Matrix pre Rozhodovací strom



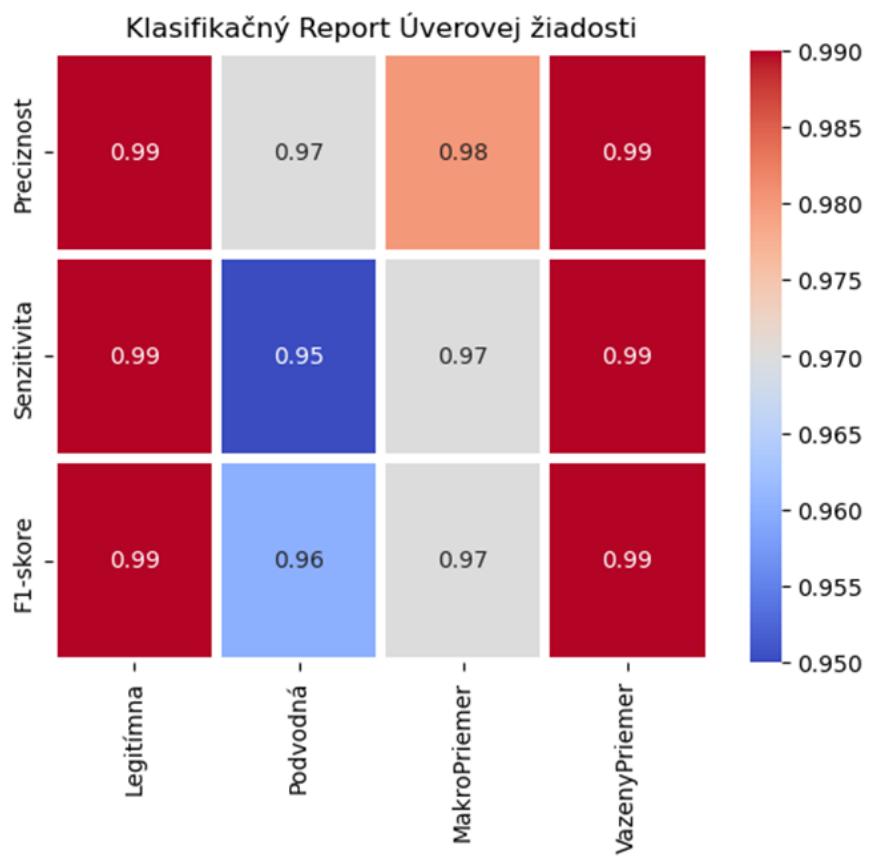
Obrázok 27 – Confusion Matrix pre Náhodný les

6.1.2 Klasifikačný report

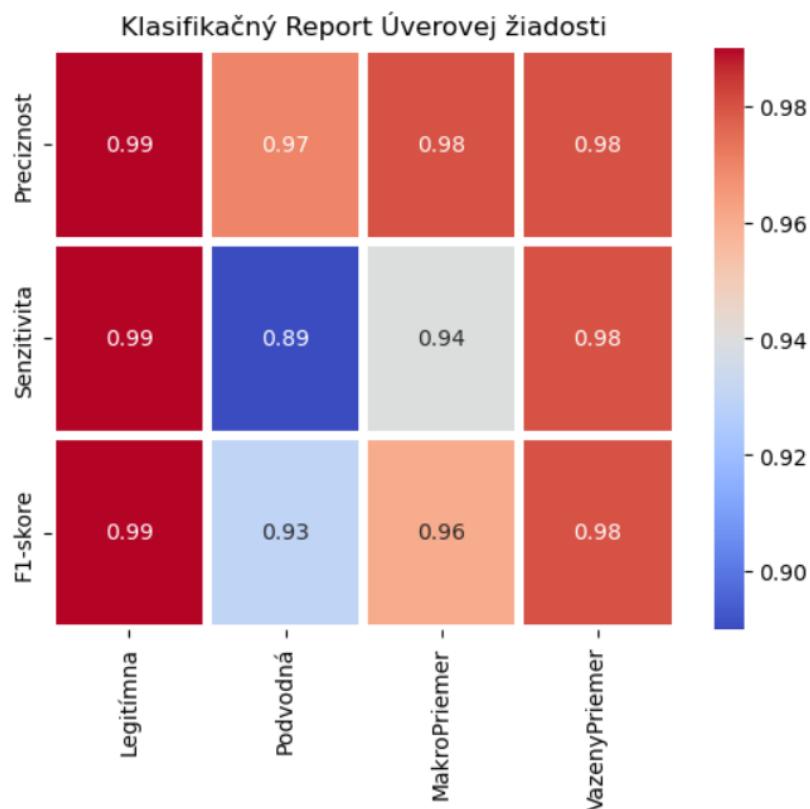
Klasifikačný report je Scikit-om poskytnuté prehľadne zobrazenie výsledných hodnôt metrík ako je precíznosť, senzitivita, F1 skóre, makro priemer (aritmetický) a vážený priemer. Klasifikačný report poskytuje tieto metriky pre obe kategórie (0 aj 1 respektívne legitímna žiadosť a podvodná žiadosť). Nás hlavne zaujímajú metriky aplikované na podvodné žiadosti. Klasifikačné reporty sa nachádzajú na obrázkoch Obrázok 28 Obrázok 29 Obrázok 30.



Obrázok 28 – Klasifikačný report pre Logistickej regresiu



Obrázok 29 – Klasifikačný report pre Rozhodovací strom

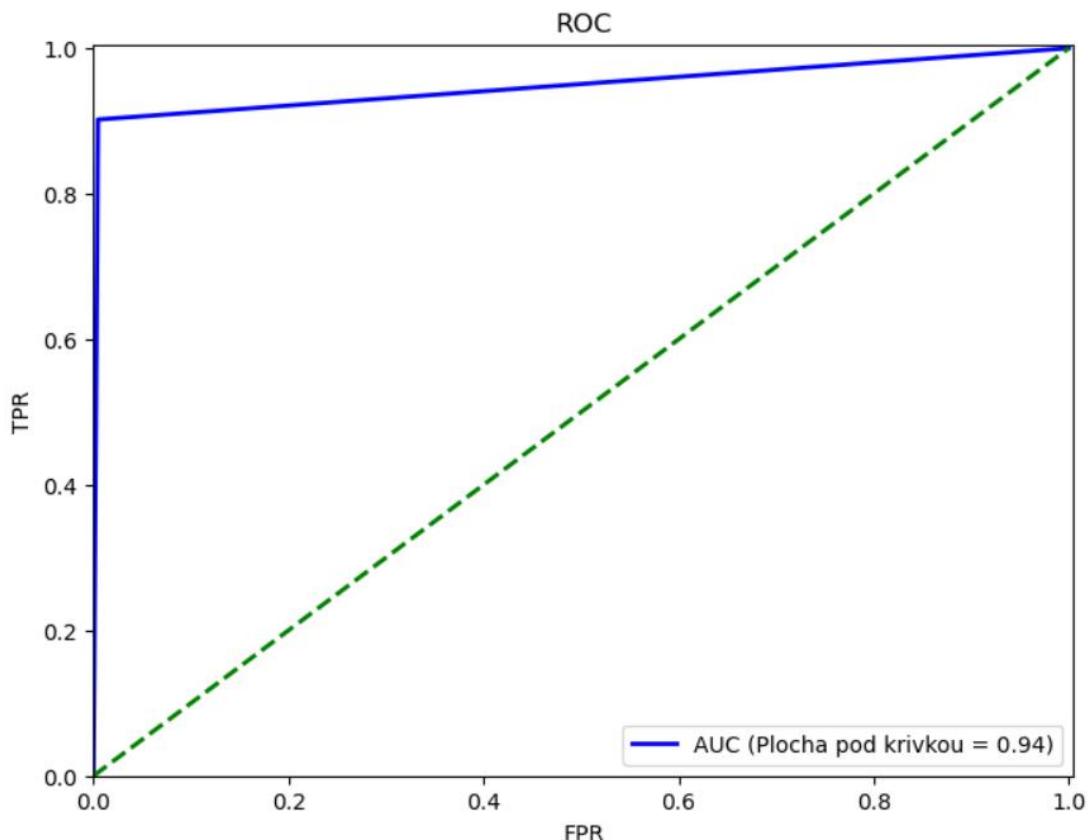


Obrázok 30 – Klasifikačný report pre Náhodný les

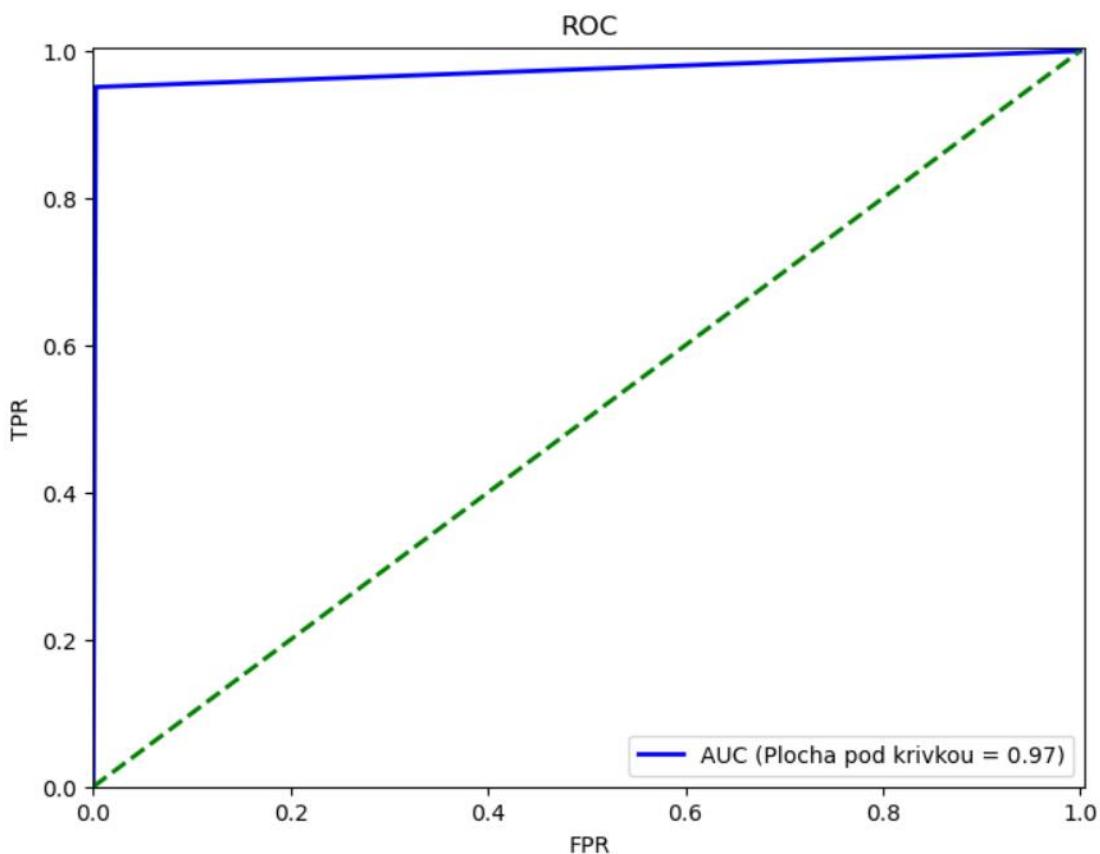
6.1.3 ROC krivka

Príslušné ROC krivky s hodnotou AUC sa nachádzajú:

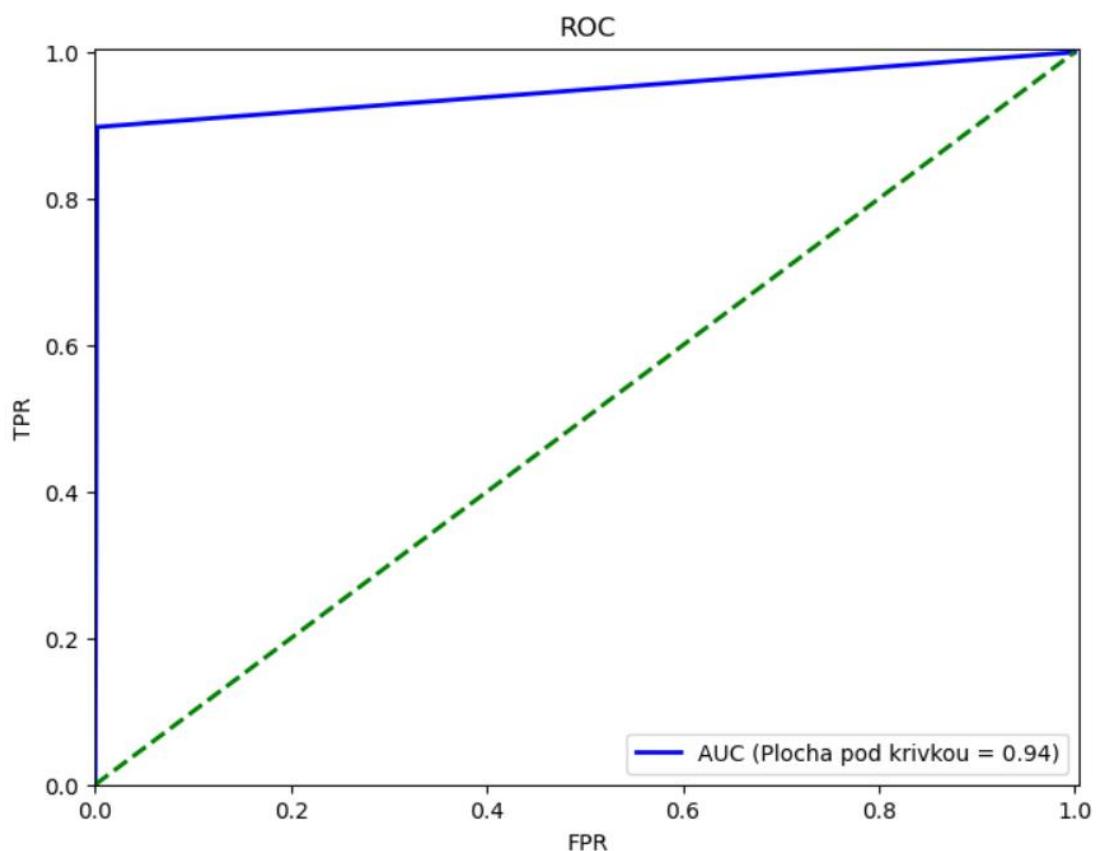
Na Obrázok 31 je Logistická regresia. Pre Rozhodovací strom, ktorý dosiahol najlepšiu hodnotu AUC bol priradený Obrázok 32. Obrázok 33 je pre náhodný les, ktorý dosiahol najmenšiu AUC hodnotu zo všetkých troch modelov.



Obrázok 31 – ROC a AUC pre Logistickú regresiu



Obrázok 32 - ROC a AUC pre Rozhodovací strom



Obrázok 33 - ROC a AUC pre Náhodný les

6.1.4 Výsledná tabuľka

Zvyšné metriky sme dopočítali bez grafického zobrazenia, výsledky možno vidieť v Tabuľke 4. Pre **Beta F1 skóre** sme použili hodnotu **parametra $\beta=2$** , aby sme zvýšili vážnosť senzitívity voči precíznosti pri výpočte F1 skóre. Všetky hodnoty boli zaokruhlene **na dve desatinné čísla nadol**.

Tabuľka 4 – Výsledné metriky hodnotenia modelov

Metriky	Logistická regresia	Rozhodovací strom	Náhodný les
Presnosť'	0.98	0.99	0.98
Precíznosť'	0.94	0.97	0.97
Senzitivita	0.9	0.95	0.89
Špecifickita	0.99	0.99	0.99
F1 skóre	0.92	0.96	0.93
Vyvážená presnosť'	0.94	0.97	0.94
Makro priemer F1 skóre	0.95	0.97	0.96
Vážený priemer F1 skóre	0.98	0.99	0.98
Beta F1 skóre	0.91	0.95	0.91
MCC	0.91	0.95	0.92
AUC	0.94	0.97	0.94

Z dôvodu vysokých negatívnych dôsledkov, ako sú strata financií, poškodenie reputácie, zdĺhavý a náročný proces vymáhania, ktoré dokážu spôsobiť neodhalené podvodné žiadosti ale zároveň aj s podstaty fungovania P2P platforem, kde žiadajú o úver aj osoby s horšou finančnou históriaou a po zohľadnený biznis filozofie spoločnosti Bondora, ktorá vyberá poplatky od úspešných žiadateľov o úver sme zvolili ako najdôležitejšiu hodnotiacu metriku **Beta F1 skóre**, kde sme určili aby senzitivita mala dvojnásobne väčší vplyv ako precíznosť na výsledok. Týmto rozhodnutím sme **zabezpečili väčšiu vyváženosť** medzi schválenými úverovými žiadostami a žiadostami ktoré by mali byť zamietnuté z dôvodu podozrenia o podvodnú žiadosť, ako keby sme použili za najdôležitejšiu metriku senzitívitu, ktorá by bola preferovaná

aby hlavnou pointou bolo určiť podvodnú žiadosť aj za cenu viacerých falošných pozitívnych predikcií. Najlepší model podľa našej prioritnej metriky je **Rozhodovací strom**. Tento model dosiahol najvyššie hodnoty vo všetkých metrikách.

6.2 Najdôležitejšie nezávislé premenné

Pre každý model sme vypísali a graficky vykreslili hodnoty desiatich najdôležitejších premenných. Premenné s podčiarkovníkom sú nominálne nezávislé premenné, ktoré boli zakódované pomocou BaseN encoder-u a ich interpretácia je zložitejšia ako u iných premenných.

1. Logistická regresia

Pri logistickej regresii, čím je hodnota koeficienta väčšia a kladná, tým viac prispieva k tomu, aby výsledná predikcia označila úverovú žiadosť ako podvodnú – 1. Naopak, čím je hodnota viac záporná, tým viac prispieva k tomu, aby bola úverová žiadosť predikovaná ako legitímná – 1.

Kladné koeficienty boli v poradí od najväčšieho: Status_0, Default_0, Status_1, MonthlyPayment, AppliedAmount, Amount.

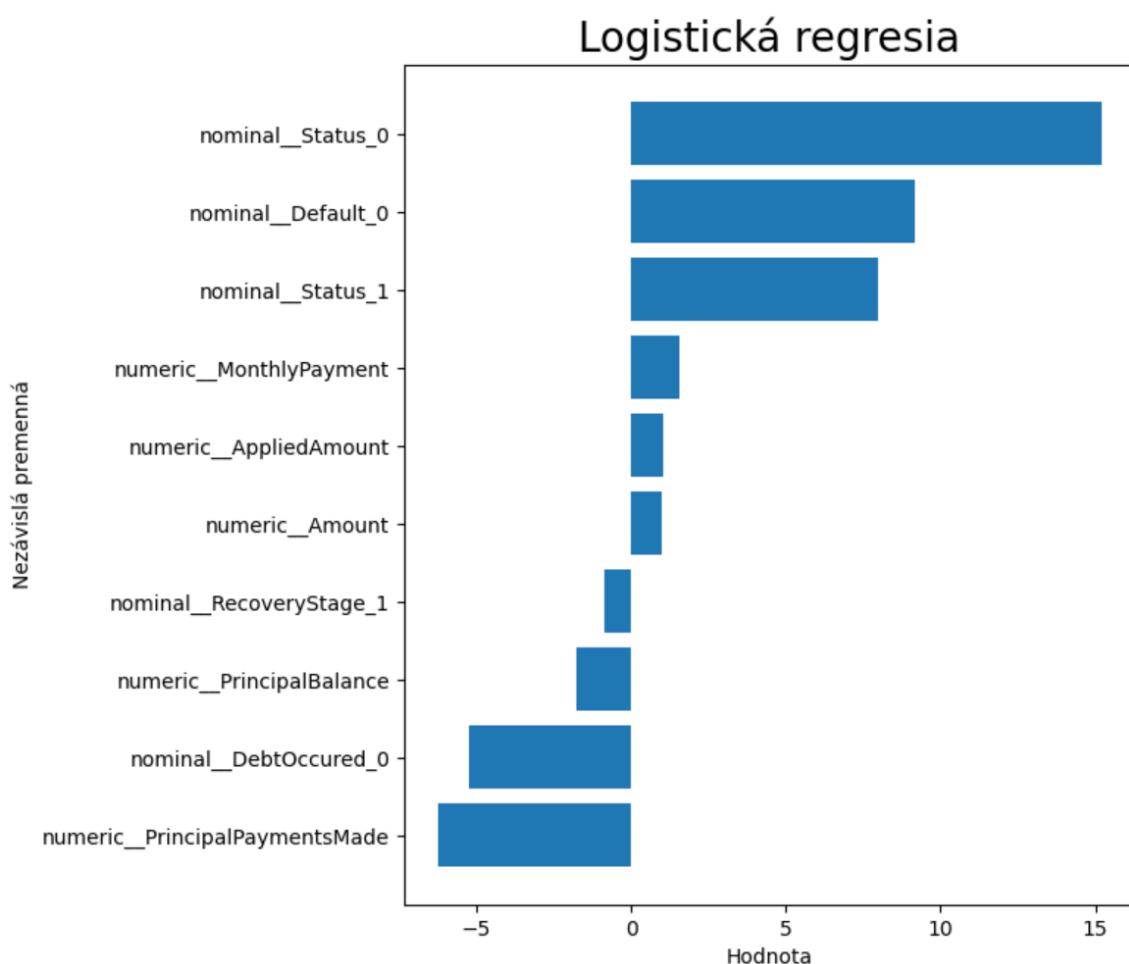
Záporné koeficienty boli v poradí od najväčšieho: PrincipalPaymentsMade, DebtOccured_0, PrincipalBalance, RecoveryStage_1. Výsledný graf je na Obrázok 34.

2. Rozhodovací strom

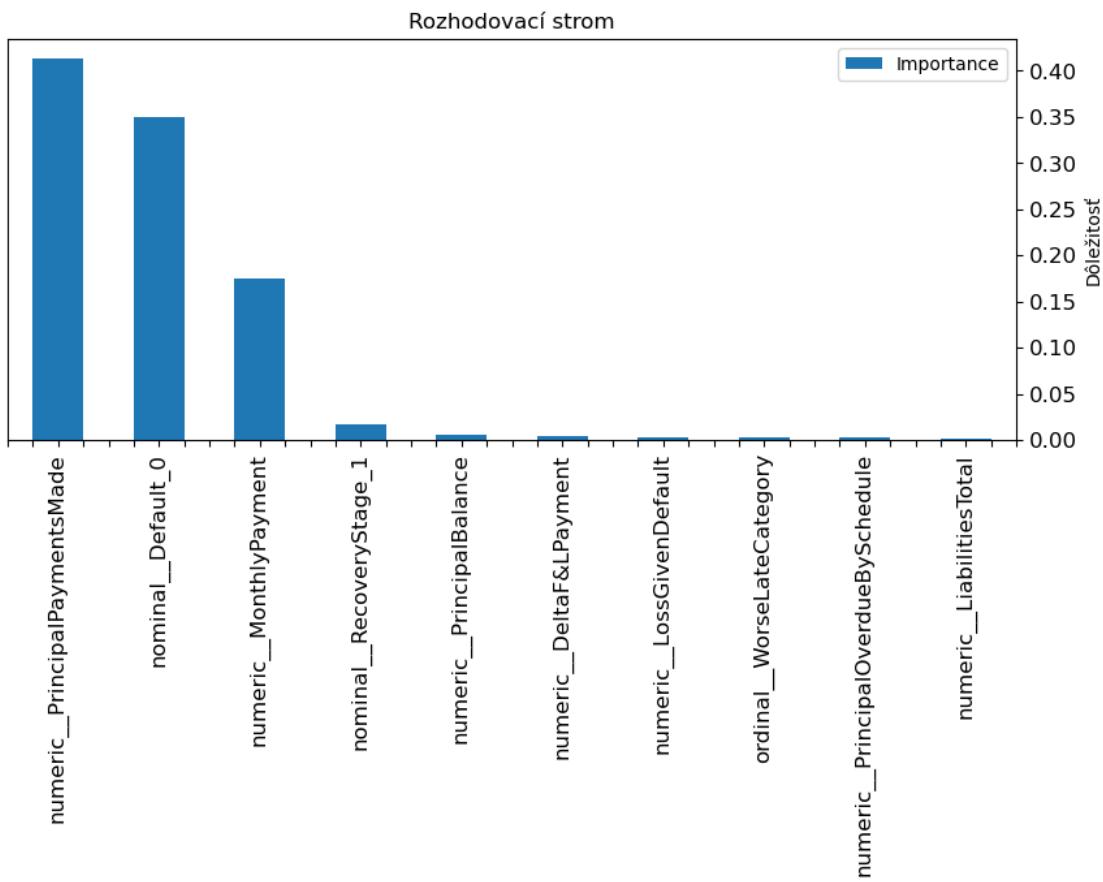
Vyššie hodnoty priradené nezávislým premenným hovoria o väčšej dôležitosti pri rozdeľovaní uzlov stromu a tak prispievajú k zlepšeniu kvality predikcií modelu. V poradí od najväčšej hodnoty priradenej dôležitosti nezávislej premennej: PrincipalPaymentsMade, Default_0, MonthlyPayment, RecoveryStage_1, PrincipalBalance, DeltaF&Payment, LossGivenDefault, WorseLateCategory, PrincipalOverdueBySchedule, InterestAndPenaltyPaymentsMade. Výsledný graf pozri Obrázok 35.

3. Náhodný les

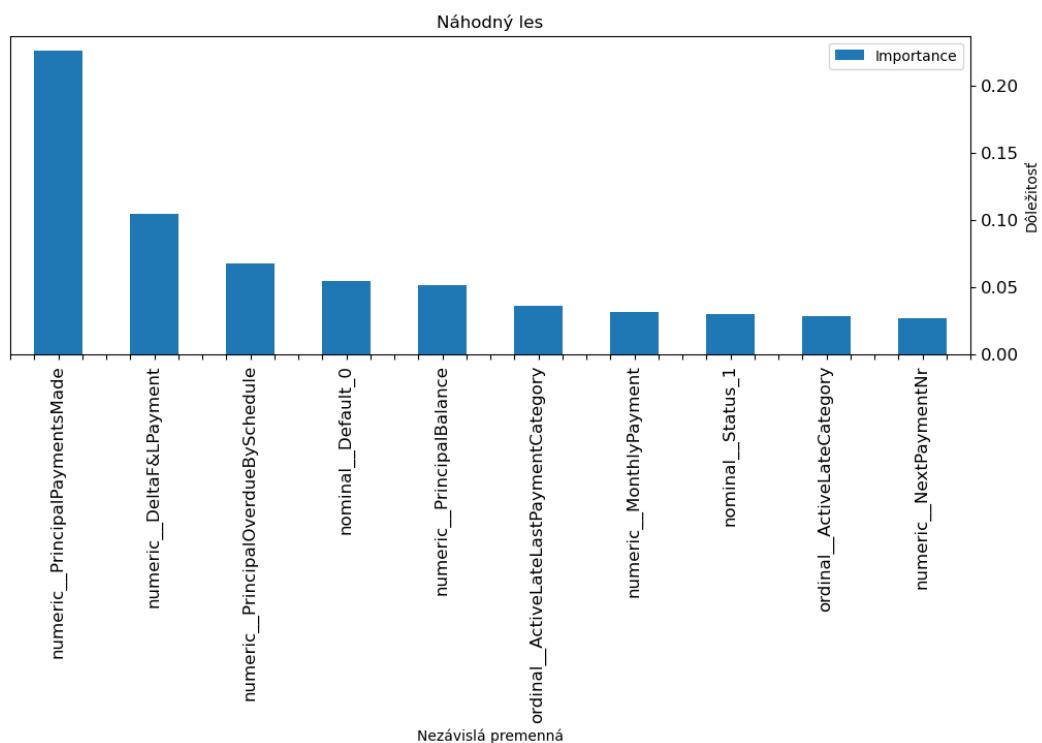
Rovnaký princíp ako pri rozhodovacom strome, avšak namiesto priamej hodnoty dôležitosti nezávislých premenných sú hodnoty každej nezávislej premennej z každého stromu spriemerované. Boli to tieto nezávislé premenné, zoradené od najväčšej hodnoty: PrincipalPaymentsMade, DeltaF&Payment, PrincipalOverdueBySchedule, Default_0, PrincipalBalance, ActiveLateLastPaymentCategory, MonthlyPayment, Status_1, ActiveLateCategory a NextPaymentNr. Výsledné ukazovatele sú na Obrázok 36.



Obrázok 34 – Hodnoty koeficientov Logistickej regresie



Obrázok 35 – Najvplyvnejšie nezávislé premenné Rozhodovacieho stromu



Obrázok 36 – Najvplyvnejšie nezávislé premenné Náhodného lesa

6.3 Hľadanie najlepšieho modelu na minimálnych vstupných dátach

Naše modely sme skúsili aplikovať na minimálnych vstupných dátach, ktoré sme určili, že sú prístupné pri prvnej žiadosti o úver. Ako vidieť v prvej polovici Tabuľka 5, všetky tri modely dosahujú značne nízke hodnoty hodnotiacich metrík. Táto skutočnosť je zapríčinená nielen nedostatočnými informačnými schopnosťami vstupných dát z ktorých by model vedel predikovať podvodnú žiadosť, ale aj veľkou nevyváženosťou datasetu. Tento efekt je viditeľný hlavne z výsledkov metrík ako je presnosť alebo vážený priemer F1 skóre, ktoré naproti ostatným metrikám dosahuje veľmi vysoké hodnoty.

Rozhodli sme sa preto zmeniť základnú hodnotu hraničnej hodnoty, ktorá je 0.5 a najšť najideálnejšiu hraničnú hodnotu modelu, ktorý dosahuje najlepšiu hodnotu F1 skóre. F1 skóre sme vybrali ako KPI preto, aby sme sa pokúsili zachovať rovnováhu medzi FN a FP, čo je problematické, ak dáta sú nevyvážené a nie sú dostatočne informatívne [40]. Ich veľká nerovnováha by mohla poškodiť financie biznisu z dôvodu pričastého odmietania úverovej žiadosti pri nadmernom množstve FP alebo naopak poškodenie reputácie a strate investícií veriteľov pri nadmernom množstve FN.

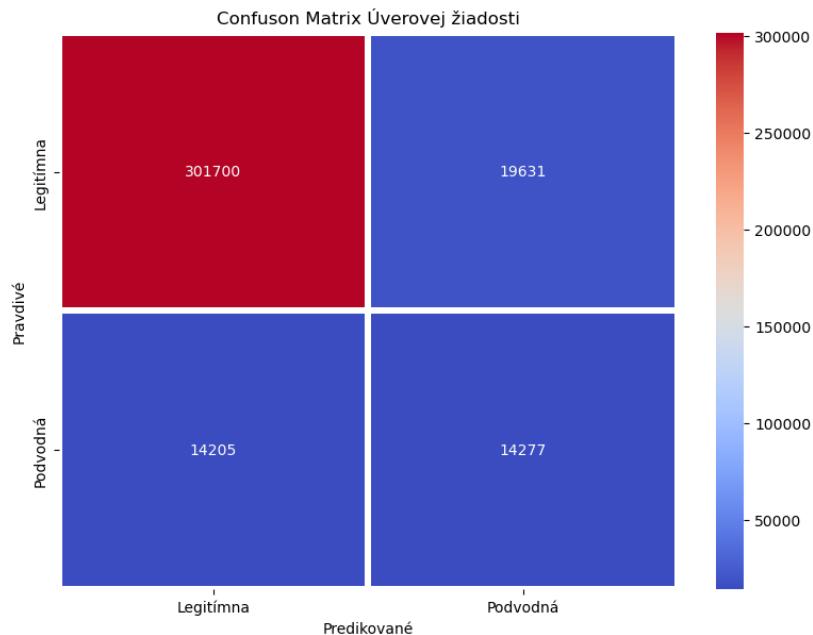
Vyhľadali sme preto takú kritickú hodnotu, kde daný model dosahoval najlepšie F2 skóre. **Najlepšie** hodnoty **F2 skóre** zo všetkých modelov sme dosiahli pri **náhodnom lese s kritickou hodnotou 0.221**. Táto kritická hodnota hovorí, že ak si je model aspoň na 22.1% istý, že úverová žiadosť je podvodná, označí ju ako podvodnú.

Výslednú confusion matrix daného náhodného lesa možno vidieť na Obrázok 37, klasifikačný report Obrázok 38, pre ROC krivku a AUC tu je Obrázok 39 a najvplyvnejšie nezávisle premenné sú na Obrázok 40. Ako vidíme, model stále dosahuje slabé výsledky v predikovaní podvodného stavu úverovej žiadosti. Preto vyhodnocujeme, že nás model nie je príliš vhodný ako predikčný model pred tým ako je úverová žiadosť schválená a je možné o nej získať viac informácií.

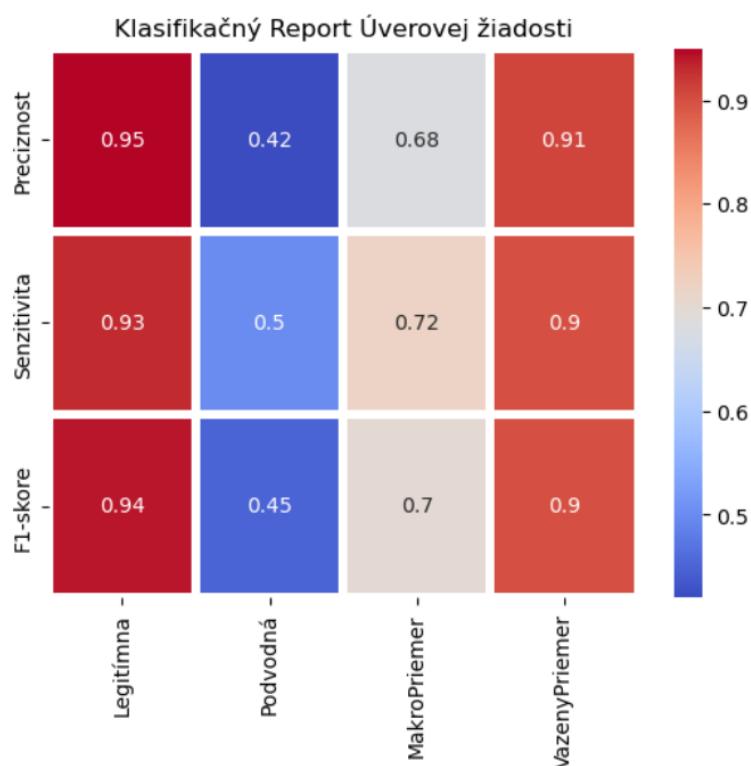
Vstupné nezávislé premenné, ktoré sme ponechali boli:

NewCreditCustomer, VerificationType, LanguageCode, Age, Gender, Country, AppliedAmount, Amount, Interest, LoanDuration, MonthlyPayment, Education, EmploymentStatus, EmploymentDurationCurrentEmployer, HomeOwnershipType , IncomeFromPrincipalEmployer , IncomeFromPension, IncomeFromFamilyAllowance, IncomeFromSocialWelfare, IncomeFromLeavePay, IncomeFromChildSupport, IncomeOther, IncomeTotal, ExistingLiabilities, LiabilitiesTotal, RefinanceLiabilities, DebtToIncome, FreeCash, , ExpectedLoss, LossGivenDefault, ExpectedReturn,

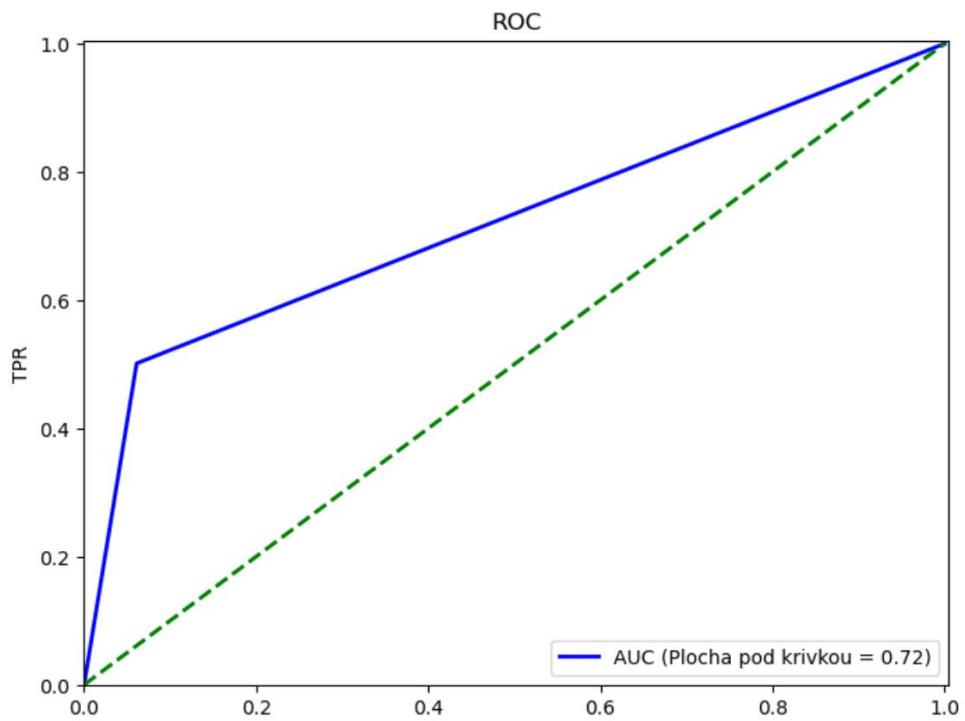
ProbabilityOfDefault, ModelVersion, Rating, NoOfPreviousLoansBeforeLoan, AmountOfPreviousLoansBeforeLoan, PreviousEarlyRepaymentsCountBeforeLoan, NrOfScheduledPayments, CreditScoreUnified, ApplicationSignedHour, ApplicationSignedWeekday, MonthlyPaymentDay.



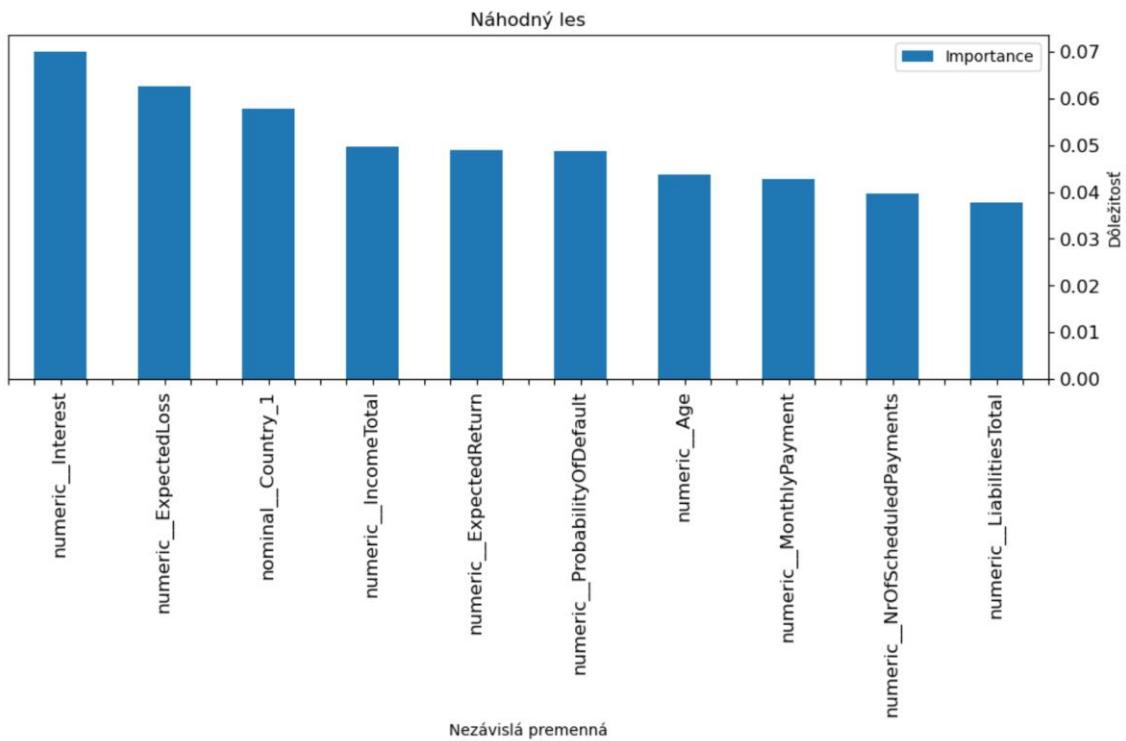
Obrázok 37 – Confusion matrix pre Náhodný les s minimálnymi vstupmi



Obrázok 38 – Klasifikačný report pre Náhodný les s minimálnymi vstupmi



Obrázok 39 – ROC a AUC pre Náhodný les s minimálnymi vstupmi



Obrázok 40 – Najvplyvnejšie nezávislé premenné Náhodného lesa s minimálnymi vstupmi

Tabuľka 5 – Výsledné metriky hodnotenia modelov pri rozličnej hraničnej hodnote

Metriky	Logistická regresia	Rozhodovací strom	Náhodný les
Hraničná hodnota (Threshold) = 0.5			
Presnosť'	0.92	0.9	0.92
Precíznosť'	0.6	0.4	0.63
Senzitivita	0.14	0.26	0.19
Špecificka	0.92	0.93	0.93
F1 skóre	0.23	0.32	0.29
Vyvážená presnosť'	0.56	0.61	0.59
Makro priemer F1 skóre	0.59	0.63	0.62
Vážený priemer F1 skóre	0.9	0.89	0.9
Beta F1 skóre	0.17	0.28	0.22
MCC	0.27	0.28	0.32
AUC	0.56	0.61	0.59
Hraničná hodnota (Threshold) = 0.221			
Presnosť'	0.9	0.85	0.9
Precíznosť'	0.41	0.26	0.42
Senzitivita	0.42	0.47	0.5
Špecificka	0.94	0.95	0.95
F1 skóre	0.41	0.34	0.45
Vyvážená presnosť'	0.68	0.68	0.72
Makro priemer F1 skóre	0.68	0.62	0.7
Vážený priemer F1 skóre	0.9	0.86	0.9
Beta F1 skóre	0.41	0.41	0.48
MCC	0.36	0.28	0.4
AUC	0.68	0.68	0.72

ZÁVER

Práca je limitovaná najmä potrebnými dátami k správnej predikcii, ktoré sú prístupne až po tom, ako je úverová žiadosť schválená. Zároveň je náš prediktívny model naviazaný na dáta od spoločnosti Bondora, ktorá poskytuje rôzne ohodnotenia tretími stranami ale aj ich vlastným ratingovým modelom. Limitujúca môže byť aj naša metodika na určenie podvodu a neprístupnosť k dátam, ktoré si finančné inštitúcie strážia a nechcú zverejňovať, či už kvôli ochrane GDPR alebo zachovaniu interného know-how.

Možné rozšírenia práce by mali byť smerované na zvýšenie úspešnosti modelu v oblasti vstupného schválenia úverových žiadostí. Profilovanie nezávislých premenných, zmena dimenzionality spracovaných dát, vyváženie zastúpenia hodnôt závislej premennej v datasete syntetickými metódami, zapracovanie predikčného modelu do webovej aplikácie, skúmanie stability datasetu a využitie iných modelov strojového učenia sú rozšírenia, ktoré majú potenciál ďalej posunúť našu prácu smerom k úspechu.

Cieľom našej práce bolo vytvoriť vhodný prediktívny model na detekciu podvodných žiadostí. Vhodný prediktívny model sme určili na základne úspešnosti modelu podľa najlepších KPI vyhodnotených pomocou hodnotiacich metrík. Sú to hlavne metriky beta F1 skóre a senzitivita, ktoré uprednostňujú správne predikovanie podvodnej žiadost aj za cenu vyššieho počtu omylných predikcii. Zároveň sme však zohľadnili aj ostatné metriky ako sú klasické, vážené a makro F1 skóre, precíznosť, AUC, MCC, vyvážená presnosť a iné. Podľa týchto metrík bol **najúspešnejším modelom Rozhodovací strom**.

V našej práci sme využili verejný dataset úverových žiadostí od spoločnosti Bondora, ktorá zverejňuje na pravidelnej báze stav úverových žiadostí. Využili sme väčšinu vstupných dát, ktoré Bondora poskytuje. Náš model preto poskytuje najlepšie predikcie vtedy, keď mu dokážeme **poskytnúť čo najviac relevantných dát** o úverovej žiadosti. Vzhľadom na tento fakt náš model nie je robustný natoľko, aby dokázal odhaliť uspokojivo úverovú žiadost pri jej vzniku, ale po zapracovaní nášho modelu môžu získať finančné inštitúcie nástroj, ktorý dokáže predikovať podvodnú úverovú žiadost po jej vzniku a tak docieliť promptné právne kroky. Zároveň by náš model po zapracovaní dokázal s určitou presnosťou určiť, či veriteľ alebo finančná inštitúcia získa návratnosť investície v prípadoch ak sa úverová žiadost dostane do stavov ako dlh, zlyhanie splácania úveru, ak mu bola poskytnutá úľava v rámci dlhu a úrokov , alebo ak dostala úverová žiadost nový splátkový kalendár.

Zistili sme, že na detekciu podvodov s **minimálnymi vstupnými dátami**, ktoré získame ešte pred schválením úverovej žiadosti je vhodný **náhodný les**. Náhodný les dosahuje zo všetkých modelov najlepšie **F1 skóre**, ktoré sme si určili ako najdôležitejšie **KPI pri minimálnych vstupných dátach**, ktoré sú veľmi nevyvážené a neposkytujú dostatok informácií. Napriek tomu, že náhodný les je náš najlepší model pre minimálne vstupné dáta, tak jeho predikčné schopnosti sú značne nespoľahlivé a zaostávajú za predikčnými schopnosťami modelu rozhodovacieho stromu, ktorý používa čo najviac relevantných dát.

Zoznam použitej literatúry

- [1] **The Business Research Company.** 2023. *Lending and payments global market report 2023*. In ReportLinker [online]. [cit. 2024-03-18]. Dostupné na internete: https://www.reportlinker.com/p06277919/Lending-And-Payments-Global-Market-Report.html?utm_source=GNW
- [2] **UK Finance.** 2023. *ANNUAL FRAUD REPORT: The definitive overview of payment industry fraud in 2022*. In UK Finance [online]. [cit. 2024-03-17]. Dostupné na: [https://www.ukfinance.org.uk/system/files/2023-05/Annual Fraud Report 2023 .pdf](https://www.ukfinance.org.uk/system/files/2023-05/Annual%20Fraud%20Report%202023%20-%200.pdf)
- [3] **Australian Competition and Consumer Commission.** 2023. *Targeting scams: Report of the ACCC on scams activity 2022*. In Australian Competition and Consumer Commission [online]. [cit. 2024-03-18]. Dostupné na internete: <https://www.accc.gov.au/system/files/Targeting%20scams%202022.pdf>
- [4] **Federal Trade Commission.** 2023. *Fraud Reports*. In Federal Trade Commission [online]. [cit. 2024-03-18]. Dostupné na internete: https://public.tableau.com/shared/XD6ZP4WS4?:display_count=n&:origin=viz_share_link
- [5] **European Central Bank.** 2023. *Report on card fraud in 2020 and 2021*. In European Central Bank [online]. [cit. 2024-03-18]. Dostupné na internete: <https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202305~5d832d6515.en.html>
- [6] **Anderson, R.** 2007. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. New York: Oxford University Press, 2007. 731s. ISBN 978–0–19–922640–5.
- [7] **GoCardless.** 2021. *First, Second and Third Party Fraud*. In GoCardless [online]. [cit. 2024-03-17]. Dostupné na: <https://gocardless.com/guides/posts/first-second-and-third-party-fraud/>
- [8] **Miller, M. A.** 2023. *What does the rise in credit and Loan Application Fraud Mean for banks?*. In Prove [online]. [cit. 2024-03-18]. Dostupné na internete: <https://www.prove.com/blog/what-does-rise-in-credit-and-loan-application-fraud-mean-for-banks>
- [9] **Oracle.** 2024. *What is machine learning?*. In Oracle [online]. [cit. 2024-03-24]. Dostupné na internete: <https://www.oracle.com/artificial-intelligence/machine-learning/what-is-machine-learning/>

- [10] **ALI, A. et al.** 2022. *Financial fraud detection based on machine learning: A systematic literature review*. In Applied Sciences [online]. 2022, roč. 12, č. 19, s. 1-24. [cit. 2024-03-24]. ISSN 2076-3417. Dostupné na internete: <https://doi.org/10.3390/app12199637>
- [11] **SULAIMAN, R.B. - SCHETININ, V. - SANT, P.** 2022. *Review of Machine Learning Approach on credit card fraud*. In Human-Centric Intelligent Systems [online]. 2022, roč. 2, č. 1-2, s. 55-68. [cit. 2024-03-26]. ISSN 2667-1336. Dostupné na internete: <https://doi.org/10.1007/s44230-022-00004-0>
- [12] **ZHAO, S. et al.** 2023. *Loan fraud users detection in online lending leveraging multiple data views*. In Proceedings of the AAAI Conference on Artificial Intelligence [online]. 2023, roč. 37, č. 4, s. 5428-5436. [cit. 2024-03-30]. ISSN 2374-3468. Dostupné na internete: <https://ojs.aaai.org/index.php/AAAI/article/view/25675>
- [13] **XU, J.J. - LU, Y. - CHAU, M.** 2015. *P2P lending fraud detection: A big data approach*. In Lecture notes in computer science [online]. 2015. Dostupné na internete: https://link.springer.com/chapter/10.1007/978-3-319-18455-5_5
- [14] **WANG, H. - WANG, Z. - ZHANG, B. - ZHOU, J.** 2019. *Information collection for fraud detection in P2P financial market*. In MATEC Web of Conferences [online]. 2018. roč. 189, p. 06006. [cit. 2024-03-30]. Dostupné na internete: https://www.matec-conferences.org/articles/matecconf/pdf/2018/48/matecconf_meamt2018_06006.pdf
- [15] **LYÓCSA, Š. et al.** 2022. *Default or profit scoring credit systems? evidence from European and US peer-to-peer lending markets*. In Financial Innovation [online]. 2022, roč. 8, č. 32, s. 1-21. [cit. 2024-03-31]. Dostupné na internete: <https://doi.org/10.1186/s40854-022-00338-5>
- [16] **BACÁ, A.** 2020. *Žltý Melón - Spôsob ako Dosiahnuť Vysoký cashflow!* In YouTube [online].[cit. 2024-04-01]. Dostupné na internete: <https://youtu.be/n9jjogpgJaw?t=1055>
- [17] **BONDORA.** 2009. *What is the General Business Information of bondora?* In Bondora [online]. [cit. 2024-04-01]. Dostupné na internete: <https://help.bondora.com/hc/en-us/articles/14898362614161-What-is-the-general-business-information-of-Bondora>
- [18] **BONDORA.** 2009. *Public reports*. In Bondora.com [online]. [cit. 2024-04-04]. Dostupné na internete: <https://www.bondora.com/en/public-reports>

- [19] **BONDORA**. 2009. *What is Bondora's 3-step collection & recovery process?* . In Bondora.com [online]. [cit. 2024-04-06]. Dostupné na internete: <https://help.bondora.com/hc/en-us/articles/14816734538897-What-is-Bondora-s-3-step-Collection-Recovery-Process>
- [20] **BONDORA**. 2014. *Explaining Bondora Rating*. In Bondora Blog [online]. [cit. 2024-04-09]. Dostupné na internete: <https://www.bondora.com/blog/explaining-bondora-rating/>
- [21] **WAGAVKAR, S.** 2023. *Introduction to the correlation matrix*. In BuiltIn [online]. [cit. 2024-04-19]. Dostupné na internete: <https://builtin.com/data-science/correlation-matrix>
- [22] **ZYCHLINSKI, S.** 2018. The Search for Categorical Correlation. In Medium [online]. [cit. 2024-04-19]. Dostupné na internete: <https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9>
- [23] **SCIKIT-LEARN**. 2024. *Preprocessing data*. In scikit-learn [online]. [cit. 2024-04-14]. Dostupné na internete: <https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing>
- [24] **SCIKIT-LEARN**. 2024. *sklearn.preprocessing.LabelEncoder*. In scikit-learn [online]. [cit. 2024-04-14]. Dostupné na internete: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>
- [25] **SCIKIT-LEARN**. 2022. *BaseN - Category Encoders*. In BaseN - Category Encoders 2.6.3 documentation [online]. [cit. 2024-04-14]. Dostupné na internete: https://contrib.scikit-learn.org/category_encoders/basen.html
- [26] **SCIKIT-LEARN**. 2024. *sklearn.feature_extraction.FeatureHasher*. In scikit-learn [online]. [cit. 2024-04-14]. Dostupné na internete: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.FeatureHasher.html
- [27] **SAHINOGLU, O.** 2020. *How to encode cyclic time data for machine learning models*. In IdeaDrops [online]. [cit. 2024-04-15]. Dostupné na internete: <https://www.ideadrops.info/post/how-to-encode-cyclic-time>
- [28] **GEEKSFORGEEKS**. 2023. *Classification vs regression in machine learning*. In GeeksforGeeks [online]. [cit. 2024-04-15]. Dostupné na internete: <https://www.geeksforgeeks.org/ml-classification-vs-regression/>
- [29] **GEEKSFORGEEKS**. 2024. *ML: Underfitting and overfitting*. In GeeksforGeeks [online]. [cit. 2024-04-15]. Dostupné na internete: <https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/>

- [30] **SCIKIT-LEARN.** 2024. *Tuning the hyper-parameters of an estimator*. In scikit-learn [online]. [cit. 2024-04-15]. Dostupné na internete: https://scikit-learn.org/stable/modules/grid_search.html#grid-search-tips
- [31] **SCIKIT-LEARN.** 2024. *Cross-validation: evaluating estimator performance*. In scikit-learn [online]. [cit. 2024-04-15]. Dostupné na internete: https://scikit-learn.org/stable/modules/cross_validation.html#multimetric-cross-validation
- [32] **IBM.** 2024. *What is logistic regression?* In IBM [online]. [cit. 2024-04-16]. Dostupné na internete: <https://www.ibm.com/topics/logistic-regression>
- [33] **IBM.** 2024. *What is a decision tree?* In IBM [online]. [cit. 2024-04-16]. Dostupné na internete: <https://www.ibm.com/topics/decision-trees>
- [34] **GEEKSFORGEEKS.** 2024. *Random Forest algorithm in machine learning*. In GeeksforGeeks [online]. [cit. 2024-04-16]. Dostupné na internete: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>
- [35] **GEEKSFORGEEKS.** 2024. *Confusion matrix in machine learning*. In GeeksforGeeks [online].. [cit. 2024-04-16]. Dostupné na internete: <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>
- [36] **OLUGBENGA, M.** 2023. *Balanced accuracy: When should you use it?* In neptune.ai [online]. [cit. 2024-04-18]. Dostupné na internete: <https://neptune.ai/blog/balanced-accuracy>
- [37] **KUNDU, R.** 2022. *F1 score in Machine Learning: Intro & Calculation*. In v7labs [online]. [cit. 2024-04-17]. Dostupné na internete: <https://www.v7labs.com/blog/f1-score-guide>
- [38] **VOXCO.** 2021. *Matthews's correlation coefficient: Definition, formula and advantages*. In Voxco [online]. [cit. 2024-04-17]. Dostupné na internete: <https://www.voxco.com/blog/matthewss-correlation-coefficient-definition-formula-and-advantages/>
- [39] **DASH, S.** 2022. *Understanding the ROC and AUC intuitively*. In Medium [online]. [cit. 2024-04-17]. Dostupné na internete: <https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02>
- [40] **BROWNLEE, J.** 2021. *A gentle introduction to threshold-moving for Imbalanced Classification*. In MachineLearningMastery.com [online]. [cit. 2024-04-25]. Dostupné na internete: <https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/>

Prílohy

Zoznam príloh

Príloha A | Príklad úverovej žiadosti na platforme Žltý Melón

Príloha B | Hodnoty v stĺpci Comment v DebtEvent

Príloha C | Tabuľka odstránených nezávislých premenných

Príloha D | Prehľad nezávislých premenných typu kategória a typu dátum a čas

Príloha E | Prehľad vyplnenia jednotlivých premenných a ich kategórií pred spracovaním

Príloha F | Korelačná matica pre numerické premenné

Príloha G | Korelačná matica pre kategorické premenné

Príloha H | ROC krivka a AUC

Príloha I | Nastavenie modelu Logistickej regresie

Príloha J | Nastavenie modelu Rozhodovacieho stromu

Príloha K | Nastavenie modelu Náhodného lesa

Príloha L | Pamäťová karta

Príloha A | Príklad úverovej žiadosti na platforme Žltý Melón

STAV FINANCOVANIA

0 € 3 425 € / 68,50 % 5 000 €

VYUŽITIE: PÓŽIČKA NA ČOKOLVEK (DOVOLENKA, SVADBA)
(ID: 809339641)

Požadovaná výška pôžičky	5 000 €
Typ aukcie	Štandardná aukcia
Rating klienta	+
Krajina	Slovenská republika
Stav overovania	<input checked="" type="checkbox"/> Žiadateľ je zatiaľ čiastočne overený (viď tooltip nad ikonou)
Splatnosť	60 mesiacov
Odporučaná úroková sadzba	12,40 % p.a.
Priemerná úroková sadzba	16,37 % p.a.
Využitie prostriedkov	Pôžička na čokolvek (dovolenka, svadba)
Poistenie	Súbor poistenia B
Dodatočné zabezpečenie pôžičky	Rozhodcovská doživotné Poistenie schopnosti splácať Dohoda o zrážkach zo mzdy

Dátum a čas ukončenia aukcie 10.01.2020 15:03:40

Investujte Do konca aukcie zostáva 8 dní 17:36:51

Žiadateľ - Undertaker

Pohlavie	Muž
Vek	29 rokov
Vzdelanie	Stredoškolské bez maturity
Rodinný stav	Slobodný/Slobodná
Počet využívanych deli	0
Typ byvania	Bývanie u rodiny
Krajina	Slovenská republika
Kraj	Trenčiansky kraj
Typ zamestaneckého pomeru	Prijem na základe pracovnej zmluvy
Odvetvie	Doprava
Prečo si chcem požičať?	Renovacia auta

Príjmy a záväzky

Hlavný príjem	1 100,00 €
Iné príjmy	0,00 €
Náklady na domácnosť	100,00 €
Záväzky	Typ záväzku Pôvodná suma pôžičky Mesačná splátka Pôžička zo Žltého melóna
	Spotrebny úver 9 000 € 188 €
	Hypoteika 35 000 € 135 €
	Spotrebny úver 5 000 € 135 €

Ponuky investorov

Užívateľské meno	Vyhľadávajúca čiastka	Ponuka	Úroková sadzba	Stav pôžičky	Dátum ponuky
Investor99185	25 €	25 €	12,00 % p.a.	Vyhľadávajúca	31.12.2019 19:12:49
Investor57181	25 €	25 €	12,40 % p.a.	Vyhľadávajúca	27.12.2019 16:32:53
VeCo	25 €	25 €	12,40 % p.a.	Vyhľadávajúca	27.12.2019 16:33:26
Investor60695	50 €	50 €	12,40 % p.a.	Vyhľadávajúca	28.12.2019 20:05:54
Katarina	250 €	250 €	12,40 % p.a.	Vyhľadávajúca	30.12.2019 11:28:15
romelis	25 €	25 €	12,40 % p.a.	Vyhľadávajúca	31.12.2019 09:10:55
Investor28136	50 €	50 €	12,40 % p.a.	Vyhľadávajúca	31.12.2019 10:13:16
meriluk	25 €	25 €	12,40 % p.a.	Vyhľadávajúca	31.12.2019 19:12:02
finman	250 €	250 €	12,40 % p.a.	Vyhľadávajúca	01.01.2020 13:40:42
Mike5	25 €	25 €	13,00 % p.a.	Vyhľadávajúca	30.12.2019 12:32:17

← Previous 1 2 3 4 5 6 Next →

Ponuky investorov

Zadajte ponuku

Vofné prostriedky na mojom účte (dotácio účtu)	19,14 €
Odporučaná úroková sadzba ?	12,40 % p.a.
Priemerná úroková sadzba ?	16,37 % p.a.
Minimálna investícia	25 €
Maximálna investícia	250 €

Výška investicie: EUR
Úroková sadzba: % p.a.
Validačný kód č.12:

Investovať

Otázky a odpovede

Žiadne otázky

Položiť otázku:

Zadať otázku

Príloha B | Hodnoty v stĺpci Comment v DebtEvent

UniqueComments
Started
PaidOff
BSecureNewSchedule
LoanDefaulted
WaitingForPaymentOrder
PaymentOrderFiled
DecisionReceivedInOurFavor
WaitingForSendingTheApplication
ApplicationSent
DebtSale
Finished
DCA2NewSchedule
OutOfCourtAgreement
Objection
ObjectionAgreementInProcess
WaitingForCivilClaim
ClaimFiled
ObjectionNewSchedule
ObjectionDecisionRecievedInOurFavor
FullyPaid
DecisionReceivedNotInOurFavor
NewSchedule
FullyPaidAccordingToAgreement
NothingToDebit
ApplicationWaitingForResent
ApplicationResent
PaidOffAccordingToJudgement
CriminalCase_NotifyAboutCriminalProceedings
CriminalCase_PoliceRequestAnswered
CriminalCase_ClaimFiled
CriminalCase_CourtDecisionMade
CriminalCase_BailiffApplicationSent
Bankruptcy_NotifyAboutBankruptcy
Bankruptcy_ClaimFiled
Bankruptcy_DecisionDeclared
PaidOffAccordingToAgreement
Deceased_DeathCertificateReceived
HopelessCase
DebtRestructuring_DecisionDeclared
DebtRestructuring_NewSchedule
DebtRestructuring_NotifyAboutDebtRestructuring
HeavilyIndebted
CustomerLivesOutsideSupportedCountries
DebtRestructuring_ClaimFiled
Deceased_NotifyAboutDeathOfCustomer
Decision1InOurFavor
CompromiseAgreement
Deceased_ProcessStarted
Bankruptcy_AbatementOfBankruptcy
DeceasedSuccessionBankruptcy
Bankruptcy_NewSchedule
Bankruptcy_DistributionPlanProposed
Decision1NotInOurFavor
Deceased_HeritantTookOverLoan
LawsuitPresentedAgainstBondora
CriminalCase_PoliceInvestigationEndedWithoutResult
Decision2InOurFavor
HighInterestComplaint
WaitingForAssignment
PaidOffDCA
WaitingToBeResent

Príloha C | Tabuľka odstránených nezávislých premenných

Odstránené nezávislé premenné		
ReportAsOfEOD	ContractEndDate	LoanDate
BiddingStartedOn	BidsPortfolioManager	BidsApi
BidsManual	PartyId	DefaultDate
NextPaymentDate	DebtOccuredOnForSecondary	ListedOnUTC
MaturityDate_Last	MaturityDate_Original	GracePeriodEnd
GracePeriodStart	CurrentDebtDaysSecondary	LoanApplicationStartedDate
FirstPaymentDate	StageActiveSince	DebtOccuredOn
ReScheduledOn	LastPaymentOn	LoanId
EL_V0	EL_V1	"Rating_V0
Rating_V1	Rating_V2	LoanNumber

Príloha D | Prehľad nezávislých premenných typu kategória a typu dátum a čas

```
# stlpce ktoré sme vyhodnoti ako kategorické po preskumani excel suboru. Pandas automaticky vsetky hodnoty  
# premenia, ktoré nie su int/float premenia na object - ak vieme vsak ze su to kategorické premenne  
# a maju obmedzny pocet moznych hodnot, je lepsie ich premenit na category  
columns_to_category = ["Country", "NrOfDependants", "EmploymentDurationCurrentEmployer", "WorkExperience",  
    "ApplicationSignedHour", "ApplicationSignedWeekday", "VerificationType", "LanguageCode",  
    "Gender", "UseOfLoan", "Education", "MaritalStatus", "EmploymentStatus",  
    "OccupationArea", "HomeOwnershipType", "RecoveryStage", "ModelVersion", "Rating",  
    "Rating_V0", "Rating_V1", "Rating_V2", "Status", "Restructured", "ActiveLateCategory",  
    "WorseLateCategory", "CreditScoreEsMicrol", "CreditScoreEsEquifaxRisk",  
    "CreditScoreFiAsiakasTietoRiskGrade", "CreditScoreEeMini", "ActiveLateLastPaymentCategory",  
    "MonthlyPaymentDay"]  
  
#nacitanie csv suborov do Pandas DataFrame, engine=pyarrow boost v rychlosti nacitania  
dataLoan = pd.read_csv("LoanData/LoanData.csv", engine='pyarrow', dtype={column: "category" for column in columns_to_category})  
  
#transformacia datumov na datetime typ - iba tie, ktoré nebudu dropnute neskôr , ,errors='coerce' znamena aby datumy, ktoré nemozno konvertovat  
#boli vyhodnotene ako NaT - not a Time  
dates=["MaturityDate_Original","MaturityDate_Last","LastPaymentOn","DebtOccuredOn",  
    "GracePeriodStart","GracePeriodEnd","NextPaymentDate","ReScheduledOn","FirstPaymentDate"]  
for date in dates:  
    dataLoan[date] = pd.to_datetime(dataLoan[date], format='%d.%m.%Y', errors='coerce')  
    1  
    2  
    3  
    4
```

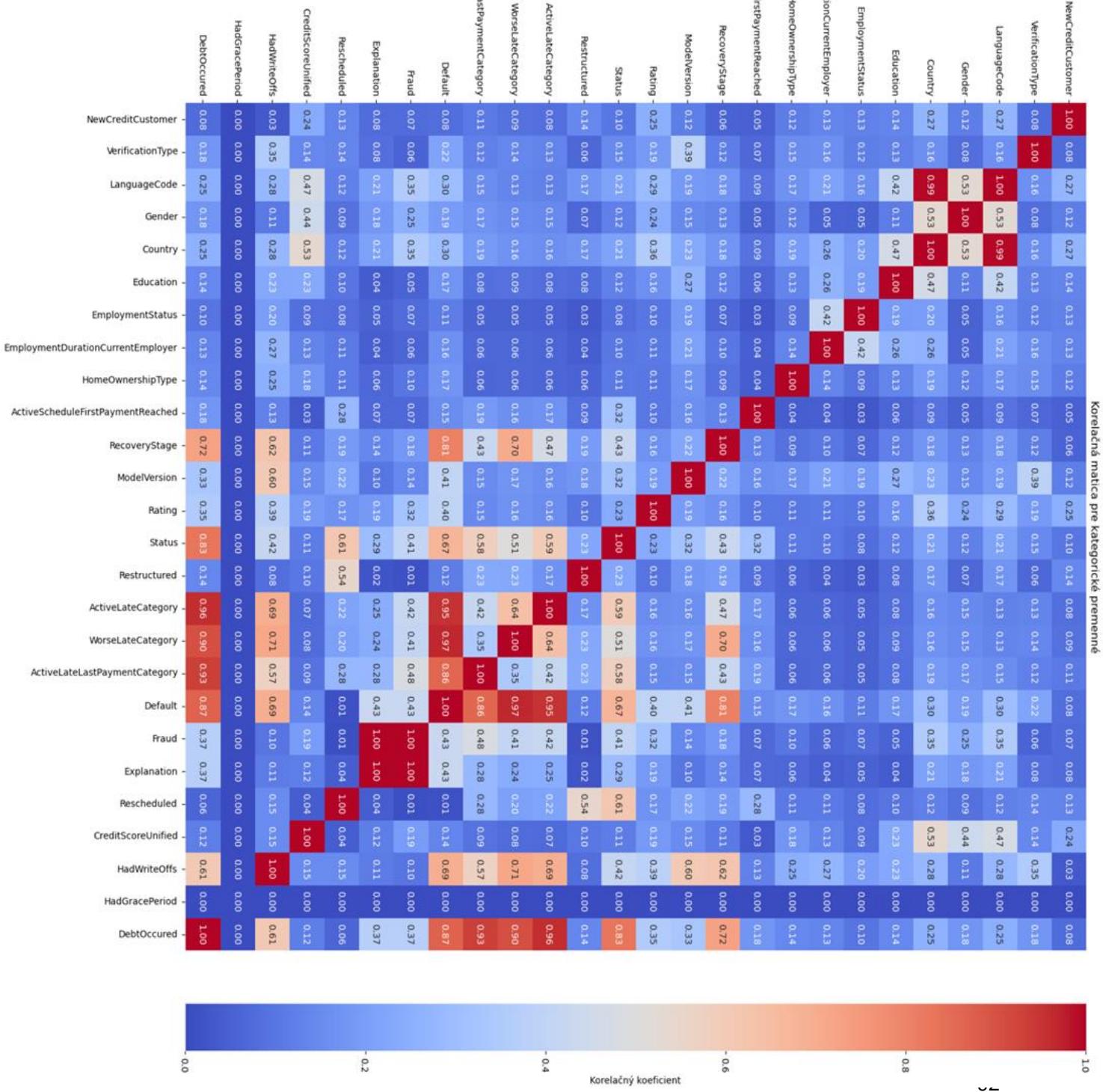
Príloha E | Prehľad vyplnenia jednotlivých premenných a ich kategórií pred spracovaním

Column	Non-Null	Dtype	Column	Non-Null	Dtype
ActiveLateCategory	135123	category	InterestAndPenaltyBalance	174041	float64
ActiveLateLastPaymentCategory	149350	category	InterestAndPenaltyDebtServicingCost	112009	float64
ActiveScheduleFirstPaymentReached	375479	bool	InterestAndPenaltyPaymentsMade	375479	float64
Age	375479	int64	InterestAndPenaltyWriteOffs	112009	float64
Amount	375479	float64	InterestRecovery	121249	float64
AmountOfPreviousLoansBeforeLoan	375466	float64	LanguageCode	375479	category
ApplicationSignedHour	375479	int64	LiabilitiesTotal	375479	float64
ApplicationSignedWeekday	375479	int64	LoanDuration	375479	int64
AppliedAmount	375479	float64	LossGivenDefault	372840	float64
Combined_EL	17493	float64	MaritalStatus	375429	category
Country	375479	category	ModelVersion	372840	category
CreditScoreEeMini	164541	category	MonthlyPayment	368789	float64
CreditScoreEsEquifaxRisk	12219	category	MonthlyPaymentDay	375479	int64
CreditScoreEsMicroL	343079	category	NewCreditCustomer	375479	bool
CreditScoreFiAsiakasTietoRiskGrade	166120	category	NextPaymentNr	200997	float64
CurrentDebtDaysPrimary	137887	float64	NoOfPreviousLoansBeforeLoan	375466	float64
DebtToIncome	375429	float64	NrOfDependants	35600	category
Default	375479	category	NrOfScheduledPayments	200997	float64
DeltaBidding&DebtOccured	137887	float64	OccupationArea	375388	category
DeltaF&LPayment	365564	float64	PlannedInterestPostDefault	121249	float64
DeltaGracePeriod	99265	float64	PlannedInterestTillDate	370874	float64
DeltaMaturityDate	375477	float64	PlannedPrincipalPostDefault	121249	float64
EAD1	121247	float64	PlannedPrincipalTillDate	54464	float64
EAD2	121247	float64	PreviousEarlyRepaymentsBeforeLoan	62099	float64
Education	375429	category	PreviousEarlyRepaymentsCountBeforeLoan	375466	float64
EmploymentDurationCurrentEmployer	364326	category	PreviousRepaymentsBeforeLoan	232411	float64
EmploymentStatus	375277	category	PrincipalBalance	375479	float64
ExistingLiabilities	375479	int64	PrincipalDebtServicingCost	112009	float64
ExpectedLoss	372840	float64	PrincipalOverdueBySchedule	360621	float64
ExpectedReturn	372840	float64	PrincipalPaymentsMade	375479	float64
Explanation	375479	category	PrincipalRecovery	121249	float64
Fraud	375479	category	PrincipalWriteOffs	112009	float64
FreeCash	375429	float64	ProbabilityOfDefault	372840	float64
Gender	375434	category	Rating	372762	category
HomeOwnershipType	373822	category	RecoveryStage	223093	category
IncomeFromChildSupport	375479	float64	RefinanceLiabilities	375479	int64
IncomeFromFamilyAllowance	375479	float64	Rescheduled	375479	bool
IncomeFromLeavePay	375479	float64	Restructured	375479	category
IncomeFromPension	375479	float64	Status	375479	category
IncomeFromPrincipalEmployer	375479	float64	UseOfLoan	375479	category
IncomeFromSocialWelfare	375479	float64	VerificationType	375429	category
IncomeOther	375479	float64	WorkExperience	36522	category
IncomeTotal	375479	float64	WorseLateCategory	245414	category
Interest	375479	float64			

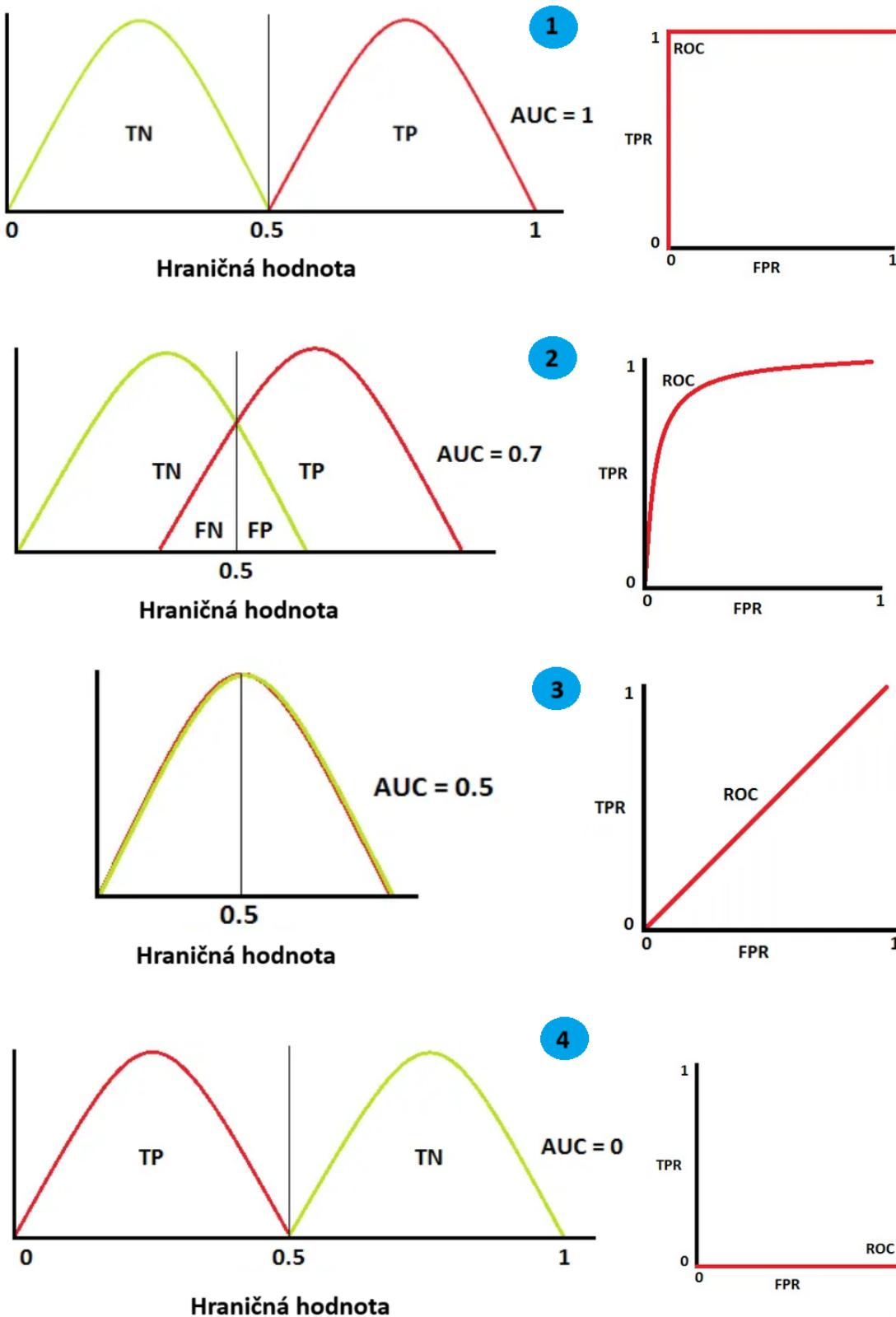
Príloha F | Korelačná matica pre numerické premenné

	ApplicationSignedHour	1.00	-0.04	0.01	0.02	0.01	0.01	-0.07	0.02	0.01	0.01	-0.05	0.00	0.01	0.02	0.04	0.00	-0.01	-0.01	0.02	0.01	0.00	-0.07	0.06	0.01	0.02	-0.03	0.03	-0.02								
ApplicationSignedWeekday	1.00	1.00	0.00	-0.01	0.01	-0.01	0.00	-0.01	0.00	-0.01	0.00	-0.01	0.00	-0.01	0.00	-0.01	0.00	-0.02	0.01	-0.01	0.02	-0.01	0.01	-0.02	0.00	0.00	0.00	0.00	0.01								
Age	0.01	0.00	1.00	0.09	0.09	-0.04	0.06	0.04	-0.01	0.07	-0.02	0.00	-0.01	0.02	0.02	0.07	0.00	-0.02	0.03	-0.04	0.01	0.02	0.05	0.05	0.04	0.06	0.08	0.03	0.09								
AppliedAmount	-0.02	-0.01	0.09	1.00	-0.07	0.00	0.22	0.76	0.07	0.02	0.04	0.02	0.01	0.02	0.03	0.02	-0.12	0.00	0.09	0.11	0.03	-0.01	0.56	0.04	-0.00	0.02	0.04	0.33	0.56	0.50	0.58						
Interest	-0.01	-0.01	-0.04	-0.00	-0.03	1.00	0.00	0.24	0.08	0.04	0.02	0.02	0.01	0.02	0.03	0.01	-0.10	0.00	0.03	0.14	-0.04	-0.09	0.16	0.81	-0.22	0.65	0.78	0.26	-0.01	0.16	-0.04	-0.17	-0.01				
LoanDuration	-0.01	0.00	0.06	0.22	0.23	0.00	1.00	0.15	-0.02	-0.01	-0.00	-0.00	-0.01	-0.01	-0.00	0.03	-0.00	-0.00	-0.04	-0.01	0.02	-0.02	0.03	0.01	-0.03	0.00	0.01	0.13	0.26	0.08	0.13	0.01	0.01				
MonthlyPayment	-0.02	-0.01	0.04	0.76	0.71	0.24	-0.15	1.00	0.08	0.03	0.03	0.03	0.01	0.03	0.05	0.01	-0.15	0.00	0.01	0.15	0.03	-0.03	0.53	0.23	-0.04	0.18	0.24	0.37	0.47	0.41	0.37	-0.20	-0.14	-0.01	-0.02	0.01	
IncomeFromPrincipalEmployer	-0.01	-0.00	-0.01	0.07	0.03	0.08	-0.02	0.08	1.00	0.02	0.13	0.03	0.01	0.06	0.09	0.04	0.10	0.01	0.14	0.37	0.88	0.01	0.10	0.16	0.07	0.07	0.08	0.13	0.05	0.07	-0.03	0.05	0.05	0.04			
IncomeFromPension	-0.01	0.00	0.07	0.02	0.01	0.04	-0.01	0.03	0.02	1.00	0.04	0.10	0.00	0.02	0.02	0.00	0.06	0.00	0.10	0.21	0.07	0.00	0.05	0.09	0.04	0.02	0.04	0.08	0.01	0.02	0.01	-0.03	0.03	0.02			
IncomeFromFamilyAllowance	-0.01	-0.01	-0.02	0.04	0.02	0.00	-0.00	0.03	0.13	0.04	1.00	0.12	0.09	0.29	0.02	-0.00	0.09	0.00	0.12	0.26	0.09	0.02	0.06	0.04	0.03	0.04	0.02	0.08	0.03	0.05	-0.02	-0.05	-0.03	0.04	-0.00	-0.02	0.00
IncomeFromSocialWelfare	-0.00	0.00	-0.00	-0.01	0.01	0.00	-0.01	0.01	0.01	0.00	0.09	0.01	1.00	0.03	0.00	0.00	0.02	0.00	0.03	0.06	0.07	0.00	0.01	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
IncomeFromLeavePay	-0.00	-0.00	-0.01	0.02	0.01	-0.00	0.03	0.06	0.02	0.29	0.09	0.03	1.00	0.02	-0.00	0.05	0.00	0.06	0.15	0.05	0.01	0.04	0.04	0.02	0.03	0.02	0.05	0.02	0.03	-0.01	-0.01	-0.01	0.01	-0.00	-0.01	0.00	
IncomeFromChildSupport	-0.00	-0.00	0.00	0.03	0.01	0.03	-0.01	0.05	0.09	0.02	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
IncomeOther	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
IncomeTotal	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
ExistingLiabilities	-0.05	-0.01	0.07	-0.12	-0.13	-0.10	0.03	-0.15	0.10	0.06	0.09	0.05	0.02	0.05	0.03	-0.00	1.00	0.01	0.26	0.26	0.03	0.07	-0.05	-0.09	-0.01	0.01	-0.03	-0.06	-0.07	-0.02	-0.11	0.55	0.42	0.08	-0.01	0.06	0.11
LiabilitiesTotal	-0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.4		
RefinanceLiabilities	-0.01	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
DebtToIncome	-0.02	-0.01	0.05	0.11	0.06	0.14	-0.05	0.15	0.37	0.21	0.26	0.12	0.06	0.15	0.09	-0.02	0.26	0.01	0.43	1.00	0.14	0.03	0.18	0.25	0.09	0.15	0.15	0.22	0.09	0.13	-0.07	-0.05	0.08	0.07	-0.01	0.03	0.32
FreeCash	-0.01	-0.00	0.00	0.03	0.01	0.04	-0.01	0.03	0.02	0.00	0.02	0.01	0.03	0.01	0.05	0.04	0.07	0.05	0.21	0.05	0.03	0.01	0.14	1.00	0.01	0.05	0.10	0.05	0.02	0.04	0.08	0.02	0.01	0.00	0.00	0.00	
MonthlyPaymentDay	-0.00	0.00	0.01	-0.04	-0.01	0.01	-0.09	0.02	-0.03	0.01	0.00	0.02	0.00	0.01	0.03	0.00	-0.07	0.00	0.01	0.03	0.01	0.00	-0.02	-0.08	0.05	-0.05	-0.05	-0.02	-0.01	-0.03	-0.01	0.00	0.00	0.00	0.00		
PlannedInterestTillDate	-0.01	-0.02	0.03	0.56	0.56	0.16	0.02	0.53	0.10	0.05	0.03	0.01	0.04	0.03	0.00	-0.05	0.00	0.13	0.18	0.05	-0.02	1.00	0.19	-0.18	0.12	0.22	0.50	0.70	0.21	-0.08	-0.15	0.11	0.02	0.01	-0.23	0.15	
ExpectedLoss	-0.02	-0.01	-0.04	0.04	0.00	0.81	-0.02	0.23	0.16	0.09	0.04	0.03	0.02	0.04	0.04	-0.01	0.09	0.00	0.14	0.25	0.10	-0.08	0.19	1.00	0.09	0.25	0.94	0.32	0.00	0.15	0.02	-0.21	0.02	-0.07	-0.09	0.10	0.16
LossGivenDefault	-0.04	0.02	-0.01	0.00	0.01	-0.22	0.03	0.04	0.07	0.04	0.03	0.02	0.00	0.02	0.01	0.03	-0.01	0.00	0.04	0.09	0.09	0.02	-0.01	0.70	-0.00	-0.12	-0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	-0.2		
ExpectReturn	-0.00	-0.01	0.02	-0.02	-0.05	0.65	0.01	0.18	0.07	0.02	0.04	0.02	0.01	0.03	0.03	-0.00	-0.02	0.00	0.12	0.13	0.04	0.01	0.21	0.25	-0.23	0.10	0.04	0.17	0.00	0.14	-0.07	0.00	0.00	0.03	-0.08	0.09	
ProbabilityOfDefault	-0.01	-0.02	-0.05	0.04	0.00	0.78	-0.05	0.24	0.08	0.04	0.02	0.01	0.02	0.03	0.00	-0.03	0.00	0.07	0.05	0.15	0.04	-0.05	0.22	0.84	-0.28	0.34	1.00	0.33	0.04	0.20	-0.07	-0.17	-0.17	-0.01	-0.07	-0.09	0.07
PrincipalOverdueBySchedule	-0.01	-0.01	0.05	0.33	0.31	0.26	0.00	0.37	0.13	0.08	0.08	0.06	0.02	0.05	0.03	-0.00	0.00	0.14	0.22	0.08	-0.02	0.50	0.32	-0.13	0.17	0.32	1.00	0.05	0.06	0.33	-0.12	0.11	0.00	-0.12	-0.11	0.13	0.25
PrincipalPaymentsMade	-0.02	-0.01	0.04	0.56	0.58	-0.01	0.01	0.47	0.05	0.01	0.03	0.01	0.00	0.02	0.00	-0.07	0.00	0.09	0.09	0.02	-0.01	0.70	-0.00	-0.12	-0.00	0.04	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	-0.2		
InterestAndPenaltyPaymentsMade	-0.00	-0.02	0.06	0.50	0.51	0.16	0.13	0.41	0.07	0.02	0.05	0.02	0.01	0.03	0.03	-0.00	-0.02	0.00	0.12	0.13	0.04	0.01	0.21	0.15	-0.15	0.14	0.20	0.06	0.37	1.00	0.07	0.01	0.15	0.15	0.23		
PrincipalBalance	-0.00	-0.00	0.08	0.58	0.61	-0.04	0.26	0.37	-0.03	-0.01	-0.02	-0.00	-0.01	-0.01	-0.02	-0.11	-0.00	-0.02	-0.07	-0.02	-0.01	0.03	0.07	-0.07	0.33	-0.24	0.26	1.00	-0.05	-0.02	-0.04	-0.04	0.34	0.18	-0.05		
NoOfPreviousLoansBeforeLoan	-0.00	-0.00	0.09	0.15	0.05	0.03	0.07	0.20	0.09	0.05	0.03	0.02	0.01	0.01	0.01	0.01	0.00	0.06	0.03	0.09	0.05	0.04	0.23	0.18	0.10	0.11	0.00	0.00	1.00	0.79	0.16	0.01	0.12	0.10	-0.03		
AmountOfPreviousLoansBeforeLoan	-0.00	-0.00	0.08	0.11	0.08	0.13	0.02	0.08	0.19	0.09	0.15	0.05	0.03	0.07	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
PreviousEarlyRepaymentsCountBeforeLoan	-0.00	-0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
NextPaymentNr	-0.02	-0.00	0.05	0.00	0.01	0.01	0.07	0.01	-0.02	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	
NrOfScheduledPayments	-0.03	-0.01	0.02	0.03	-0.11	-0.03	-0.01	-0.02	-0.02	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	
DeltaMaturityDate	-0.03	-0.01	-0.01	0.02	0.03	-0.11	-0.03	-0.01	-0.02	-0.02	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	
DeltaF&LPayment	-0.02	-0.03	0.06	0.11	0.08	0.13	0.02	0.08	0.19	0.09	0.15	0.05	0.03	0.07	0.20	0.00																					

Príloha G | Korelačná matica pre kategorické premenné



Príloha H | ROC krivka a AUC



Príloha I | Nastavenie modelu Logistickej regresie

Randomized Search	Logistická regresia	GridSearch	Logistická regresia
	solver = newton-cholesky		solver = newton-cholesky
	penalty = l2		penalty = l2
	max_iter = 300		max_iter = 275
	class_weight = None		class_weight = None
	C = 10826.36733874054		C = 10826.36733874054
	Výsledné FBeta = 0.91		Výsledné FBeta = 0.91

Príloha J | Nastavenie modelu Rozhodovacieho stromu

Randomized Search	Rozhodovací strom	GridSearch	Rozhodovací strom
	Rozhodovací strom		Rozhodovací strom
	splitter = best		splitter = best
	min_samples_split = 3		min_samples_split = 3
	min_samples_leaf = 10		min_samples_leaf = 11
	min_impurity_decrease = 0.0		min_impurity_decrease = 0.0
	max_depth = 40		max_depth = 40
	criterion = entropy		criterion = entropy
	Výsledné FBeta: 0.95		Výsledné FBeta: 0.95

Príloha K | Nastavenie modelu Náhodného lesa

Randomized Search	Náhodný les	GridSearch	Náhodný les
	n_estimators = 125		n_estimators = 175
	min_weight_fraction_leaf = 0.0		min_weight_fraction_leaf = 0.0
	min_samples_split = 15		min_samples_split = 15
	min_samples_leaf = 1		min_samples_leaf = 1
	min_impurity_decrease = 0.0,		min_impurity_decrease = 0.0
	max_depth = 50		max_depth = 50
	Výsledné FBeta: 0.91		Výsledné FBeta: 0.91

Príloha L | Pamäťová karta

Príloha vo forme pamäťovej karty obsahuje:

- Práca vo formáte PDF
- Datasetsy s použitými dátami
- Zdrojový kód
- Používateľská príručka vo formáte PDF