

EKONOMICKÁ UNIVERZITA V BRATISLAVE
NÁRODOHOSPODÁRSKA FAKULTA

Evidenčné číslo: 101007/I/2020/36097107973709060

MODELOVANIE KREDITNÉHO RIZIKA

Diplomová práca

2020

Bc. Marko Dávid Vateha

EKONOMICKÁ UNIVERZITA V BRATISLAVE
NÁRODOHOSPODÁRSKA FAKULTA

MODELOVANIE KREDITNÉHO RIZIKA

Diplomová práca

Študijný program: Bankovníctvo

Študijný odbor: Ekonómia a manažment

Školiace pracovisko: Katedra bankovníctva a medzinárodných financií

Vedúci záverečnej práce: doc. Ing. Štefan Lyócsa, PhD.

Bratislava 2020

Bc. Marko Dávid Vateha

PodĎakovanie

Táto práca a v tejto podobe, by nebola vznikla bez odborného vedenia môjho školiteľa. Preto ďakujem touto cestou vedúcemu diplomovej práce doc. Ing. Štefanovi Lyócsovi, PhD. za jeho rady a pripomienky, ktorými mi bol nápomocný pri jej písaní. A ďakujem aj Rastovi, osobnému priateľovi, za jeho životné usmernenia a rozhovory, ktoré mi boli vždy výzvou. V neposlednom rade ďakujem svojim rodičom, za podporu a ich čas pri výchove.

Abstrakt

VATEHA, Marko Dávid: *Modelovanie kreditného rizika*. – Ekonomická univerzita v Bratislave. Národohospodárska fakulta; Katedra bankovníctva a medzinárodných financií. – doc. Ing. Štefan Lyócsa, PhD. – Bratislava: NHF EU, 2020, 87 s.

Rozvoj umelej inteligencii a strojového učenia rozšírili ešte viac digitálnu priepasť vo všetkých vedných disciplínach. A finančný sektor v tomto nie je výnimkou. Vzhľadom na vývoj a dosiahnuté výsledky samoučiacich systémov, navrhuje diplomová práca revíziu tradičných techník pre modelovanie kreditného rizika. V záujme spresnenia hodnotenia kreditného rizika by mali byť zvážené aj pokroky v prediktívnej analýze. Na základe prehľadu literatúry sa zdá, že zoznam štúdií o uplatniteľnosti strojového učenia pre potreby kreditnej analýzy je dostatočne dlhý. Napriek tomu len niekoľko z nich vykonalo výskum na takej veľkej vzorke produkčných dát simulujúc úverovú prax, a to najmä z dôvodu ich dôvernosti. Okrem toho vystali určité nejasnosti týkajúce sa metrík použitých na účely hodnotenia úspešnosti modelu. Diplomová práca sa viac zameriava na manažérsky pohľad s dôrazom na faktory špecifické pre banku. V štúdiu boli použité viaceré metódy spracovania údajov a prístupy strojového učenia s cieľom nájsť tie najvhodnejšie. Výsledky signalizujú, že implementácia samoučiacich systémov môže pre úverovú inštitúciu priniesť pridanú hodnotu, a to na úrovni použitých zdrojov aj presnosti odhadov. Záver práce naznačuje, že strojové učenie v kombinácii s neštandardnými údajmi o klientoch sú sľubnou alternatívou k tradičnému modelovaniu úverového rizika. Táto štúdia ponecháva priestor na ďalšie vylepšenie použitím zložitejších, nelineárnych, modelovacích techník.

Kľúčové slová: kreditné riziko, nebankové finančné inštitúcie, úvery, kreditné skórovanie, ekonometria, prediktívne modelovanie, strojové učenie

Abstract

VATEHA, Marko Dávid: *Credit risk modelling*. – University of Economics in Bratislava. Faculty of National Economy; Department of Banking and International Finance. – doc. Ing. Štefan Lyócsa, PhD. – Bratislava: NHF EU, 2020, 87 p.

In the wake of AI and machine learning digital divide widens in all disciplines. Financial sector is no exception. In the light of recent development, this master thesis suggests revision of traditional techniques for credit risk modelling. Technical advances in predictive analytics should be considered when seeking for improvement in credit risk assessment. After reviewing literature on application of machine learning in credit scoring, it might seem that the number of relevant studies is vast. Nonetheless, just a few of those studies have done an extensive research using real-world data at such a scale, mainly due to its confidentiality. Additionally, there are some concerns about metrics which have been used for evaluation purposes. When modelling credit risk, this thesis shifts attention to a managerial view with an emphasis on factors relevant for bank-specific credit risk management purposes. In the study, multiple data processing methods and machine learning approaches were used with the intention of finding the most suitable ones. The result indicates that the implementation of self-learning systems can bring an added value, both in terms of resources (resulting in profits) and increased accuracy. On this basis, it is concluded that the machine learning combined with non-traditional data sources is likely a promising alternative to traditional credit risk modelling techniques. This study leaves some room for further improvement as more complex, non-linear modelling techniques were not utilized.

Key words: credit risk, non-bank financial institutions, loans, credit scoring, econometrics, predictive modelling, machine learning

Obsah

Zoznam skratiek a znakov	10
Úvod	11
1 Finančný systém, finanční sprostredkovatelia a riziká z financovania.	13
1.1 Finančný systém	13
1.2 Finančný trh a jeho miesto vo finančnom systéme	14
1.2.1 Členenie finančných trhov	15
1.3 Nástroje finančného trhu	16
1.3.1 Spotrebný úver	18
1.4 Finančné inštitúcie a ich význam	19
1.4.1 Obchodné banky ako finančné inštitúcie	21
1.4.2 Nebankové subjekty – ostatní finanční sprostredkovatelia	22
1.4.3 Úverové registre ako ochrana pred rizikom	25
1.5 Regulácia finančného systému	26
1.6 Manažment rizika	28
1.6.1 Prístupy k riadeniu rizík	28
1.6.2 Riadenie rizík v bankovom sektore	29
1.6.3 Kreditné riziko v kontexte bankovníctva	29
1.7 Strojové učenie	30
1.7.1 Modelovanie kreditného rizika pomocou strojového učenia	32
1.7.2 Prehľad literatúry	33
2 Cieľ práce	38
3 Metodika a metódy skúmania	39
3.1 Logistická regresia	40
3.1.1 Vstupy do modelu	42
3.1.2 Linková funkcia sigmoid	42
3.1.3 Strata krížovej entropie	44
3.1.4 Optimalizácia stratovej funkcie cez gradient klesania	46
3.1.5 Regularizácia modelu	47
3.2 Support Vector Machine	50
3.2.1 Optimalizačný problém SVM	52
4 Tvorba predikčného modelu pomocou metód strojového učenia	55
4.1 Základný rámec postupu pri úprave a tvorbe modelu	55
4.1.1 Vstupné dáta	56
4.2 Predpríprava dát	58
4.2.1 Izolovanie testovacej vzorky	58
4.2.2 Chýbajúce hodnoty	58
4.2.3 Odstránenie odľahlých hodnôt	60
4.2.4 Feature engineering	61
4.2.5 Encoding	62

4.2.6	Logaritmovanie numerických premenných	63
4.2.7	Škálovanie vstupov	63
4.2.8	Informácie o minulom správaní sa klienta	64
4.2.9	Výber štatisticky významných atribútov	64
4.2.10	Problém multikolinearity	65
4.3	Fáza modelovania a hodnotenia	66
4.3.1	Nastavenie hyperparametrov modelu	66
4.3.2	Metódy hodnotenia	68
4.3.3	Hodnotenie kvality modelu	70
4.3.4	Hodnotenie najlepšieho modelu	75
4.3.5	Technické špecifiká experimentu	77
5	Diskusia	78
	Záver	80
	Zoznam použitej literatúry	82
	Prílohy	88

Zoznam tabuliek:

Tabuľka 1 Medziročná zmena aktív lízingových spoločností, spoločností splátkového financovania a faktoringových spoločností.	25
Tabuľka 2 Zhrnutie údajov o klientskej databáze Application.csv	57
Tabuľka 3 Rozdelenie súboru na tréningovú a testovaciu vzorku	58
Tabuľka 4 Prehľad o závažnosti chýbajúcich dát v registri žiadateľov	58
Tabuľka 5 Prehľad atribútov, ktoré boli dopočítavané metódami strojového učenia	60
Tabuľka 6 Prehľad novovytvorených atribútov a ich vzťah k sledovanej premennej	61
Tabuľka 7 Hlavné charakteristiky datasetov použitých v práci	64
Tabuľka 8 Optimalizácia hyperparametrov individuálnych klasifikačných metód	67
Tabuľka 9 Výsledky modelov na upravenom datasete bez selekcie atribútov	71
Tabuľka 10 Manažérsky pohľad – dataset bez selekcie atribútov	73
Tabuľka 11 Výsledky modelov na upravenom datasete po selekcii atribútov	74
Tabuľka 12 Manažérsky pohľad – dataset po selekcii atribútov	74
Tabuľka 13 Rýchlosť modelov na rôznych datasetoch	75

Zoznam obrázkov:

Obrázok 1 Zobrazenie zastúpenia rôznych foriem financovania nebankového sektoru	23
Obrázok 2 Vývoj aktív spoločností poskytujúcich úvery a pôžičky nebankovým spôsobom (mil. EUR)	24
Obrázok 3 Schematický prehľad sférou umelej inteligencie	31
Obrázok 4 Dizajn realizovaného systému hodnotenia kreditnej kvality	39
Obrázok 5 Priebeh sigmoid funkcie	43
Obrázok 6 Stratová funkcia pri logistickej regresii	45
Obrázok 7 Krížová entropia	45
Obrázok 8 Hľadanie minima gradientom klesania	46
Obrázok 9 Rôzne rozhrania, ktorými je možné klasifikovať pozorovania do tried	51
Obrázok 10 Vytvorenie rozhrania pomocou pomocných vektorov	51
Obrázok 11 Vytvorenie rozhrania s porušením hraníc	52
Obrázok 12 Povolené narušenie rozpätia medzi pomocnými vektormi a rozhraním	54
Obrázok 13 Dizajn realizovaného systému hodnotenia kreditnej kvality	56
Obrázok 14 Príklad label encoding-u	62
Obrázok 15 Príklad one-hot encoding-u	62
Obrázok 16 Výsledok testovania hypotéz o významnosti atribútov	65
Obrázok 17 Príklad uplatnenej krížovej validácie	67
Obrázok 18 Tabuľka klasifikácií a jej derivácie	69
Obrázok 19 Tabuľka klasifikácií pre LR – dataset bez selekcie atribútov	72
Obrázok 20 Tabuľka klasifikácií pre SVM – dataset bez selekcie atribútov	72
Obrázok 21 Kolmogorov-Smirnov test a krivka ROC	76
Obrázok 22 Kumulovaný zisk a krivka zdvihu	76

Zoznam skratiek a znakov

AI	Artificial Intelligence (umelá inteligencia)
AUROC	Area Under the Receiver Operating Characteristics (plocha pod krivkou ROC)
BCBS	Basel Committee on Banking Supervision (Bazilejskí výbor pre bank. dohľad)
BS	Brier Score (brierovo skóre)
EAD	Exposure at Default (expozícia v riziku)
ECB	European Central Bank (Európska centrálna banka)
FNR	False Negative Rate (miera nesprávnych pozitívnych predikcií)
HC	Home Credit Group
L1	Lasso regularizácia
L2	Ridge regularizácia
LGD	Loss Given Default (strata v prípade zlyhania)
LR	Logistická regresia
ML	Machine Learning (strojové učenie)
NBS	Národná banka Slovenska
NFI	Nebankové finančné inštitúcie
PD	Probability to Default (pravdepodobnosť zlyhania)
RVA	Rizikovo-vážené aktíva
SGD	Stochastic Gradient Descent (náhodný gradient klesania)
SVM	Support Vector Machine (metóda pomocných vektorov)
TPR	True Positive Rate (miera skutočne pozitívnych predikcií)
VaR	Value at Risk (hodnota v riziku)

Úvod

Problematika riadenia úverového rizika sa dostáva v posledných rokoch významne do popredia, vďaka čomu vidíme nebyvalý rozvoj tejto oblasti.¹ Udržateľnosť predmetu podnikania bánk závisí od minimalizácie zaknihovaných strát z úverovej činnosti. Prístupy k manažmentu kreditného rizika sa vyvíjali nielen rozvojom teoretického základu ale aj príspevom progresu v oblasti informačných technológií.² Technologické pokroky ako strojové učenie a umelá inteligencia, už dosiahli alebo prekonalí v mnohých aplikačných doménach ľudskú kapacitu. Hovoríme napríklad o klasifikácii obrázkov, rozpoznávaní reči a preklade textov. V oblasti posudzovania úverového rizika však stále existujú metódy založené na expertíze úverového špecialistu, ktorý posudzuje bonitu žiadateľa o úver.³ Väčšie finančné domy sú na tom lepšie a opierajú sa o rigorózne, matematicko-štatistické modelovanie. Samoučiace prediktívne systémy implementujú zatiaľ najmä noví hráči na trhu – fintech spoločnosti.⁴

Aj keby predikcie inteligentných algoritmov neboli oveľa presnejšie ako expertné analýzy, použitie technológií a automatizácia vedie k zníženiu prevádzkových nákladov. Stroj, ktorý po inštalácii naprogramovaní dokáže zachytiť súvislosti a učiť sa novými pozorovaniami, v mnohom šetrí zdroje. Ďalší dobrý dôvod pre banky, prečo by sa mali o túto tému zaujímať. Hoci je stále mimoriadne náročné zostrojiť efektívny model na hodnotenie úverovej bonity klientov, strojové učenie si v ňom pomaly nachádza uplatnenie. Ako podnet k empirickej analýze nás viedol aj rastúci záujem o umelú inteligenciu v slovenskom bankovníctve.⁵ Analyzované dáta pochádzajú z nebankového prostredia, čo situáciu zjednodušuje. Banky sú povinné svoj skóringový model nechať schváliť regulátorom⁶, ktorému záleží, aby boli výstupy modelu interpretovateľné. Nebankový subjekt môže používať modely aj s nižšou až žiadnou mierou interpretovateľnosti.

¹ McKinsey Global Banking Annual Review 2019. *The last pit stop? Time for bold late-cycle moves.*

² MUNKHDALAI, Lkhagvadorj et al. An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. In *Sustainability MDPI Journals*, vol. 11. January, 2019. p. 2.

³ GIETZEN, Thomas. Credit Scoring vs. Expert Judgement – a randomized controlled trial. In *Working papers on finance* No. 2017/09. Swiss Institute Of Banking And Finance. June, 2017. p. 3.

⁴ FUSTER, Andreas et al. *The Role of Technology in Mortgage Lending*. [elektronický zdroj] Federal Reserve Bank of New York Staff Reports No. 863. Dostupné na: https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr836.pdf

⁵ TEMPEST [elektronický zdroj] Dostupné na: <http://www.tempest.sk/index.php/najhorucejsie-technologicke-trendy-roka-2020-s-ktorymi-sa-urcite-stretnete-aj-vy-5c1.html>

⁶ § 14 opatrenia NBS č. 4/2007 o vlastných zdrojoch financovania bánk a požiadavkách na vlastné zdroje financovania bánk a o vlastných zdrojoch financovania obchodníkov s cennými papiermi a požiadavkách na vlastné zdroje financovania obchodníkov s cennými papiermi

V prípade úverových inštitúcií sú systémy kreditného bodovania zamerané na odhad pravdepodobnosti zlyhania (PD) klientov. V intenciách tejto práce sa však jemne odkloníme od tohto štandardu a budeme namiesto výšky rizika vyhodnocovať rozlišovaciu schopnosť modelu medzi tzv. dobrými a zlými úvermi. Práca študuje automatické hodnotenie úverovej kvality ako problém binárnej klasifikácie. Jedným dychom dodávame, že niektoré použité modely prichádzajú aj s kvantifikáciou pravdepodobnosti zlyhania, napr. logistická regresia. Táto vlastnosť by mohla poslúžiť pri stratifikácii klientov podľa ich vnútornej rizikovosti a pomohla by inštitúcii stanoviť cenu za úver, teda úrok. V rovnakom kontexte sa odvíjalo viacero už publikovaných prác z tejto problematiky, čo nám poskytovalo orientačný bod.

Z inventára nástrojov strojového učenia si vyberáme logistickú regresiu a metódu pomocných vektorov. Obe tieto metódy bližšie popíšeme v kapitole 3. *Metodika a metódy skúmania*. V oboch prípadoch sa jedná o lineárne modely. Máme za to, že problém hodnotenia kreditnej kvality môže byť vo svojej podstate oveľa zložitejší a predpoklad o lineárnej separácii medzi pozorovaniami nemusí platiť, respektíve nemusí byť stopercentný. Aplikáciu komplexnejších, nelineárnych modelov nechávame na ďalší výskum v tejto oblasti.

Úspešnosť zostrojených modelov popisujeme v poslednej kapitole. Hodnotená je pomocou štandardných štatistických metrik aj finančnými ukazovateľmi o ušetrenom kapitále investora, čím prinášame aj manažérsky pohľad na vec.

Štúdiou sa snažíme priniesť pridanú hodnotu odborníkovi z praxe aj akademikovi pôsobiacim v oblasti kreditného hodnotenia. Potenciálny prínos strojového učenia v úverovom procese bol diskutovaný vo viacerých publikáciách, počínajúc rokom 2003.⁷ Napriek tomu bolo len málo experimentov realizovaných s produkčnými dátami v takom veľkom rozsahu pozorovaní ako je tomu v tejto práci.⁸ Zároveň pri modelovaní pracujeme s odvetvovým štandardom, ktorý predstavuje logistická regresia⁹ a ktorý rozširujeme o rozmer „strojovosti“. V týchto bodoch vidíme praktický aspekt práce. Z akademického hľadiska, štúdia poskytuje nezávislé hodnotenie najnovších metód klasifikácie žiadateľov o úver a ponúka základ, do ktorého sú vítané budúce vstupy.

⁷ BAESENS, Bart et al. Benchmarking state-of-the-art classification algorithms for credit scoring. In *Journal of the Operational Research Society*, Volume 54. January, 2003. p. 627–635.

⁸ FINLAY, Steven. Credit scoring for profitability objectives. In *European Journal of Operational Research*, Volume 202. April, 2010. p. 528–537.

⁹ BOLTON, Christine. *Logistic regression and its application in credit scoring*. [elektronický zdroj] University of Pretoria, 2019. p. 27. Dostupné na: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1024.2660&rep=rep1&type=pdf>

1 Finančný systém, finanční sprostredkovatelia a riziká z financovania.

V úvodnej časti práce sa budeme venovať opisu finančného systému a jeho prvkov, ktoré tvoria celkovú finančnú infraštruktúru ekonomiky. Dobré fungujúci finančný systém je nevyhnutnou podmienkou ekonomického rozvoja.¹⁰ Cieľom je presunúť zdroje na miesta, kde budú čo najefektívnejšie využité. Tento presun zdrojov v sebe neodlúčiteľne nesie riziko, že subjekt, ktorý prijíma zdroje z trhu nebude schopný ich riadne a na čas vrátiť spôsobujúc tým ekonomickú ujmu veriteľovi. Finančný systém tvorí rad finančných sprostredkovateľov pôsobiacich na rôznych finančných trhoch. Finanční sprostredkovatelia zabezpečujú funkčnosť toku zdrojov od subjektov, ktoré nimi disponujú a majú ho v prebytku k subjektom, ktoré ich potrebujú. Význam sprostredkovania tkvie v redukování informačnej asymetrie, časovej náročnosti a transakčných nákladov. Bližšie sa pozrieme na výhody a nevýhody financovania založeného na bankovom systéme a klasifikujeme jeho alternatívy. Špeciálnu pozornosť venujeme riziku. Riziko je neoddeliteľnou súčasťou ekonomickej aktivity. Ak ekonomicky aktívny subjekt chce byť úspešný, presnejšie byť dlhodobo úspešný, manažment rizika je toho kľúčový predpoklad.

1.1 Finančný systém

Finančný systém je základným stavebným kameňom národného hospodárstva každej krajiny a v literatúre je často označovaný za „jeden z najdôležitejších výdobytkov modernej doby.“¹¹ Reálna ekonomická aktivita by len veľmi ťažko vznikala, ak by tu nebol kanál zabezpečujúci presun vzácneho kapitálu od prebytkových subjektov, teda tých, ktorí sporia, k subjektom deficitným, ktorí si vypožičiavajú kapitál za účelom realizácie spotreby alebo investovania. Pri neexistencii finančného systému a bez uskutočňovania finančného sprostredkovania by sme pravdepodobne žili v inom spoločenskom usporiadaní.¹² „Finančný systém je súbor trhov, inštitúcií, zákonov, regulácií a techník, s ich pomocou sú obchodované obligácie, akcie a iné cenné papiere, určované úrokové sadzby a poskytované finančné služby po celom svete.“¹³ Rose rozdeľuje finančný systém na jeho parciálne zložky. Jednotlivým stavebným kameňom finančného trhu sa budeme konkrétne venovať v ďalšom texte. Začneme

¹⁰ POLOUČEK, Stanislav – KULHÁNEK, Lumír – MIRDALA, Rajmund et al. *Peniaze, banky, finančné trhy*. Bratislava: Iura Edition, 2010, Ekonomia. s. 17. ISBN 978-80-8078-305-1.

¹¹ Tamtiež s. 17.

¹² Tamtiež s. 17.

¹³ ROSE, Peter S. – MAREK, Karel. *Peněžní a kapitálové trhy*. Praha: Victoria Publishing, 1995. s. 16. ISBN 80-58605-52-X.

detailným pohľadom na finančný trh, opornú sústavu ekonomického života a prejdeme k finančným nástrojom, spomedzi ktorých sa zameriame hlavne na úvery. Na záver sa pozrieme na činnosť finančných inštitúcií a prácu s rizikom. Zo všetkých dostupných inštitúcií finančného trhu, budú predmetom nášho záujmu najmä banky a nebankové finančné inštitúcie.

1.2 Finančný trh a jeho miesto vo finančnom systéme

Osobitný význam pre finančné sprostredkovanie má finančný trh, ktorý prepája a uvádza do optima dva základné trhy – trh s výrobnými faktormi a trh s výrobkami.¹⁴ Proces poskytovania finančných prostriedkov pomáha vyťažiť investičné príležitosti práve vtedy, keď sa naskytnú bez nutnosti ich oddiaľovania, či straty, čím zabezpečuje ekonomický rast a zvyšovanie životnej úrovne. Finančný trh nachádza svoje uplatnenie pri umožňovaní toku úspor k subjektom, ktoré potrebujú viac zdrojov ako je množstvo, ktorým práve disponujú. Pričom jedinci, ktorí uvoľňujú svoje disponibilné prostriedky, sú za túto činnosť odmenení formou úroku.

Finančný trh je všeobecný ekonomický termín a nereflektuje konkrétne miesto. Tento pojem agreguje všetky miesta, kde prebieha transfer zdrojov vo vzťahu veriteľ-dlžník.¹⁵ Jednotlivé konkrétne trhy sa však môžu líšiť veľkosťou či povahou. Niektoré finančné trhy sú malé s malou aktivitou, a iné, napríklad burza cenných papierov v New Yorku – NYSE, obchodujú miliardy dolárov denne.¹⁶

Finančný trh plní viacero funkcií.¹⁷ Dostatočne sme zatiaľ poukázali len na kreditnú funkciu, ktorá významne podporuje ekonomický rast tým, že umožňuje ekonomickým subjektom presúvať budúcu spotrebu a investície do súčasnosti. Subjekty získavajú prostriedky potrebné na túto aktivitu z trhu, kde sú ponúkané aktuálne voľné zdroje. Nadväznou funkciou je depozitná funkcia, ktorá umožňuje ukladanie dočasne voľných prostriedkov na finančný trh. Vďaka tejto akumulácii sa vytvorí kapitál, ktorý je možné transformovať do formy investície. Tým sa realizuje iná funkcia, konkrétne funkcia premeny peňazí na finančnú investíciu. Tento vzťah platí aj naopak, a to premenou investičného

¹⁴ HRVOLOVÁ, Božena - BADURA, Peter - MARKOVÁ, Jana. *Analýza finančných trhov*. 3. prepac. a dopl. vyd. Praha: Wolters Kluwer, 2015. s. 17. ISBN 978-80-7478-948-9.

¹⁵ DE HAAN, Jakob - OOSTERLOO, Sander - SCHOENMAKER, Dirk. *European financial markets and institutions*. Cambridge: Cambridge University Press, 2009. p. 5. ISBN 978-0-511-50686-4.

¹⁶ NYSE [elektronický zdroj] Dostupné na: <https://www.nyse.com/data/transactions-statistics-data-library>

¹⁷ REJNUŠ, Oldřich. *Finanční trhy: učebnice s programem na generování cvičných testů*. Praha: Grada Publishing, 2016. s. 40-41. ISBN 978-80-247-5871-8.

nástroja na peniaze. Premenu peňazí na finančný investičný nástroj sa realizuje ochrana ich hodnoty, príp. aj zhodnotenie. Peniaze sú vo svojej čistej forme vystavené inflácii. Inflácia sa však premieta do nominálnej hodnoty finančného inštrumentu a pri spätnej premene na peniaze sa prejaví v prírastku ich objemu. V tomto procese sa realizuje funkcia ochrany bohatstva a funkcia likvidity. Okrem toho, trh zabezpečuje aj platobnú funkciu alebo skrz diverzifikáciu aj funkciu ochrany pred rizikom.

Podľa povahy a predmetu činnosti rozlišujeme niekoľko typov finančných trhov. Členeniu finančných trhov sa venujeme v nasledujúcej podkapitole. Ako si ukážeme, neexistuje jeden uniformný pohľad na ich klasifikáciu. Je viacero rovín, ktoré predeľujú množinu finančných trhov a záleží len od toho, ktorá perspektíva je pre nás zaujímavá alebo dôležitá.

1.2.1 Členenie finančných trhov

Aj napriek svojej charakteristickej inštitucionálnej podobe nie je skupina finančných trhov dokonale vnútorne homogénna. Ponúka sa viacero hľadísk členenia finančných trhov. V práci uplatňujeme pohľad Roseho.¹⁸

Azda najdôležitejším pohľadom je rozdelenie finančných trhov na peňažné a kapitálové. Pri tomto členení slúži ako kritérium doba splatnosti. Peňažný trh je miestom, kde subjekty uvoľňujú a vypožičiavajú krátkodobé zdroje. Finančné nástroje tohto trhu sú emitované so splatnosťou do jedného roka. Na kapitálovom trhu sa naopak realizujú finančné transakcie s dobou splatnosti nad jednoročný horizont. Kapitálový trh slúži na financovanie dlhodobých investičných príležitostí, ktoré dlžníci realizujú a splácajú z budúcich peňažných tokov. Investori tento trh vyhľadávajú z dôvodu, že im ponúka veľmi lákavé zhodnotenie finančných prostriedkov. Časový aspekt však robí tieto transakcie rizikovejšie.

Ak je emisia finančného nástroja vykonaná prvýkrát, ide o obchod na primárnom trhu. V prípade, že daný finančný nástroj už emitovaný bol a teraz prebieha jeho následný obchod, hovoríme o sekundárnom trhu.

Trhy uľahčujú kupujúcim a predávajúcim meniť svoje aktuálne finančné pozície obchodovaním s finančnými inštrumentami. Na finančnom trhu môžu byť umiestnené rôzne finančné inštrumenty. Bežne sa obchoduje s dlhopismi, akciami, komoditami, devízami

¹⁸ ROSE, Peter S. – MAREK, Karel. *Peněžní a kapitálové trhy*. Praha: Victoria Publishing, 1995. s. 18-34. ISBN 80-58605-52-X.

a/alebo ich derivátmi. Na trhu s dlhovými cennými papiermi dominujú spravidla emisie štátu. Akciový trh je dynamický a vyžaduje si skúsenosť a precízny manažment rizika. Pri komoditách sa na finančných trhoch najčastejšie stretávame so zlatom, drahými kovmi či energetickými surovinami a devízový trh je spravidla neorganizovaný trh, kde prebieha výmena rozdielných národných mien vo veľkých objemoch.

1.3 Nástroje finančného trhu

V predchádzajúcom texte sme si pripomenuli diverzitu finančných trhov. Menované trhy sú voči sebe viac či menej izolované, no existujú sily, ktoré tieto trhy prepájajú. Jedným z takýchto činiteľov je úmysel racionálneho ekonomického subjektu využívať také finančné nástroje, ktoré najlepšie odpovedajú jeho požiadavkám a prinášajú najlepšie podmienky.

Fabozzi klasifikuje finančné nástroje podľa druhu pohľadávky, ktorú má majiteľ voči emitentovi.¹⁹ V prípade, že je záväzok vyjadrený pevnou sumou, finančný nástroj sa považuje za dlhový nástroj. Napríklad štátne dlhopisy, dlhopisy podnikov alebo vklady v bankách sú dlhové nástroje, ktoré si vyžadujú pevné platby. Druhou kategóriou sú majetkové nástroje (equity instruments). Majetkové nástroje zaväzujú emitenta finančného nástroja vyplatiť jeho držiteľovi sumu, ktorá vychádza z prípadného zisku, výnosov a pod. Pri majetkových nástrojoch nevieme vopred kvantifikovať navrátenú sumu, pretože tieto nástroje vyjadrujú vlastnícky podiel. Kmeňové akcie sú príkladom majetkovej pohľadávky. Niektoré nástroje však môžu spadať do oboch kategórií. Napríklad, prednostné akcie sú majetkovým nástrojom, ktorý oprávňuje investora na prijatie pevnej sumy, čo je charakteristika dlhového nástroja.

Druhé delenie, ktoré prezentuje Fabozzi vychádza zo spôsobu odvodenia hodnoty.²⁰ Primárne, nazývané aj hotovostné nástroje, sa oceňujú priamo na trhu. Tieto nástroje majú hodnotu samé o sebe a sú ľahko prevoditeľné. Radíme sem bežné nástroje ako vklady a úvery. Do druhej skupiny spadajú nástroje, ktorých hodnota je odvodená od hodnoty jedného alebo viacerých podkladových komponentov. Podkladovou zložkou v takomto prípade môže byť aktívum, index alebo úroková sadzba, prípadne niečo iné. Pre tieto nástroje je charakteristické, že keď cena podkladovej zložky stúpa alebo klesá, ovplyvňuje to aj hodnotu derivátových nástrojov, avšak nie nevyhnutne o rovnaké percento.

¹⁹ FABOZZI, Frank J. *The handbook of financial instruments*. John Wiley & Sons, Inc. 2002. p. 2-3. ISBN 0-471-22092-2

²⁰ Tamtiež s. 749

Rose označuje všetky hore spomenuté nástroje ako investičné finančné nástroje.²¹ Tie sú podľa jeho názoru právnym dokladom dokumentujúcim, že ich majitelia poskytnú svoje prostriedky na finančnom trhu nejakému subjektu a tento záväzok stále trvá. Doplnkovým faktom je, že Rose rozlišuje aj neinvestičné nástroje. To sú také nástroje, ktoré neplnia kreditnú funkciu ale majú iné odôvodnenie svojej existencie. Do tejto kategórie radí platobné karty, poisťné zmluvy a iné právne dokumenty. Bližšie sa budeme venovať investičnému finančnému nástroju – úveru.

Úvery sú rozhodujúcim predmetom produktového portfólia bánk.²² Pri pohľade na súvahy bánk zistíme, že tvoria veľkú časť ich aktív. Úver môžeme popísať ako „*ekonomicko-právnu skutočnosť prejavujúcu sa poskytnutím finančnej hodnoty na konkrétny čas a za odplatu.*“²³ Suma, o ktorú dlžník splatí svojmu veriteľovi na konci obdobia viac, je odmenou za vzdanie sa likvidity i podstúpenie rizika. Cena úveru sa vyjadruje percentuálne ako časť pôvodnej sumy a nazýva sa úrok. Úrok však nie je jediný náklad, ktorý s čerpaním úveru súvisí. Náklady spôsobujú aj dodatočné administratívne poplatky či poistenie. Na vyjadrenie celkových nákladov bola zavedená veličina – RPMN – ročná percentuálna miera nákladov. RPMN vyjadruje aké percento zo zapožičanej sumy bude klient splácať na ročnej báze, započítavajúc všetky náklady spojené so zaobstaraním úveru.²⁴

Pri úverovaní si kreditná inštitúcia uvedomuje prepojenosť zisku s rizikom. Očakávaná úverová miera r_L sa líši od zmluvnej úrokovej miery na úver r_C , keďže niekoľko dlžníkov nebude schopných vyrovnať svoje záväzky.²⁵ Uvažujúc o riziko-neutrálnej inštitúcii, rozdiel medzi zmluvnou mierou a očakávanou mierou je matematicky vyjadrený Rovnicou 1.1a.

$$E(1 + r_C) = (1 + r_C) * (1 - p) + (1 + r_C) * p * \gamma = 1 + r_L \quad (1.1a)$$

Malé p je pravdepodobnosť zlyhania klienta a γ je miera návratnosti (recovery rate), ktorá sa vyjadruje ako zlomok istiny a úroku vymožitelný po nastatí úpadku dlžníka. Rovnicu môžeme ilustrovať na príklade kde je zmluvná úroková miera 9%, pravdepodobnosť zlyhania

²¹ ROSE, Peter S. – MAREK, Karel. *Peněžní a kapitálové trhy*. Praha: Victoria Publishing, 1995. s. 34-40. ISBN 80-58605-52-X.

²² CHEN, Jiaqian – VERA, Giuseppe. Bank Balance Sheets and the Value of Lending. In *IMF Working Paper*, No. 17/111, May, 2017. ISSN 1018-5941.

²³ POLOUČEK, Stanislav – KULHÁNEK, Lumír – MIRDALA, Rajmund et al. *Peniaze, banky, finančné trhy*. Bratislava: IuraEdition, 2010, Ekonomie. s. 342. ISBN 978-80-8078-305-1.

²⁴ NBS [elektronický zdroj] <https://www.nbs.sk/sk/dohlad-nad-financnym-trhom/ochrana-financneho-spotrebitela/spotrebitelske-prava/pozicky-a-uvery>

²⁵ DE HAAN, Jakob - OOSTERLOO, Sander - SCHOENMAKER, Dirk. *Financial markets and institutions: european perspective*. 3rd ed. Cambridge: Cambridge University Press, 2015. p. 321. ISBN 978-1-107-53936-5.

0,05% a miera návratnosti z nesplateného úveru 80%. Potom je očakávaná úverová miera na úrovni 7,91%. Výpočet prezentuje *Rovnica 1.1b*.

$$(1,09) * (0,95) + (1,09 * 0,05 * 0,80) = 1,0791 \quad (1.1b)$$

S úverom sa spája viacero špecifických vlastností. Tieto produkty majú nízku likviditu a pomerne vysokú rizikovosť, vyplývajúcu z informačnej asymetrie, kompenzovanú zvýšenými výnosmi. Minimálne vo svete nenulových úrokových sadzieb tento záver o výnosnosti dokonale platil. Banka alebo iný kreditný ústav operuje s cudzími zdrojmi, podlieha zásadnej regulácii a snaží sa generovať zisk, ktorý bude uspokojivý pre investorov. Z týchto dôvodov musí byť takáto inštitúcia vo svojom podnikaní veľmi obozretná a schopná rozlíšiť žiadateľov o úver, ktorí budú schopní riadne a na čas splatiť poskytnuté peniaze od uchádzačov, ktorí pri napĺňaní svojho záväzku zlyhajú. Inštitúcia poskytujúca úvery ale aj iné kreditné produkty zvažuje riziko obsiahnuté v tej-ktorej žiadosti o financovanie. V opísanom prípade hovoríme o úverovom riziku. Úverové riziko sa naplní keď dlžník nebude schopný splniť svoj záväzok v plnej miere a v dohodnutom čase. Dôvodom v tomto prípade nemusí byť nutne len jeho zlá finančná situácia, ale aj jeho vedomá neochota vyrovnať vzniknuté záväzky.

1.3.1 *Spotrebný úver*²⁶

Úver nie je len finančný nástroj ale aj právny inštitút čo robí ekonomickú realitu ešte o trochu komplexnejšou.²⁷ Banky a nebankové finančné inštitúcie ponúkajú celé portfólio rôznych foriem úveru. Na najväčšie nominálne hodnoty znejú spravidla dlhodobé úvery ako podnikateľské úvery, hypotekárne úvery či komunálne úvery. Najfrekvencovanejšou položkou sú však pôžičky pre neinštitucionálnych klientov alebo malé podnikateľské jednotky. Tieto subjekty, teda domácnosti a živnostníci, si financujú svoje spotrebné výdavky cez špecifický produkt – spotrebný úver. Využitie takéhoto úveru je často na obstaranie statku, služieb alebo na vyrovnanie predchádzajúcich záväzkov. Úver sa spláca v pravidelných intervaloch (vo všeobecnosti mesiac) splátkami, ktoré zahŕňajú istinu aj úrok. Spotrebný úver, pri ktorom nie je nutné preukazovať účel a peniaze môže dlžník minúť na čokoľvek, sa nazýva bezúčelový.

²⁶ V tejto podkapitole chceme upozorniť na odlišnosti v ekonomickej a právnej terminológii a poukázať na rozdiel v pojmoch spotrebný úver a spotrebiteľský úver.

²⁷ POLOUČEK, Stanislav – KULHÁNEK, Lumír – MIRDALA, Rajmund et al. *Peniaze, banky, finančné trhy*. Bratislava: Iura Edition, 2010, Ekonómia. s. 342. ISBN 978-80-8078-305-1.

Spotrebné úvery, ktoré smerujú domácnostiam a jednotlivcom (kde dlžník vystupuje ako fyzická osoba a zdroje nepoužije na realizáciu podnikateľskej činnosti) sú od 1. januára. 2020 regulované novým legislatívnym ustanovením. Zákon o spotrebiteľských úveroch²⁸ cieľi na väčšiu ochranu finančného spotrebiteľa v odvetví úverovania. Na Slovensku máme negatívne skúsenosti s nebankovými úvermi alebo pôžičkami pre obyvateľstvo.²⁹ Tento fakt nútil zákonodarné orgány k prísnejšej regulácii tohto odvetvia.

V doterajšom texte sme už stihli špecifikovať finančný systém aj finančný trh. V tejto kapitole sme definovali finančné nástroje (s dôrazom na úver) a v ďalšej kapitole sa zameriame na inštitúcie, ktoré sprostredkujú finančné služby.

1.4 Finančné inštitúcie a ich význam

Uviedli sme, že finančný systém sa skladá z trhov, inštitúcií, zákonov, regulácií a s ich pomocou sú obchodované najrôznejšie finančné inštrumenty. V texte tejto kapitoly rozoberieme význam a činnosť finančných inštitúcií. Spomedzi všetkých inštitúcií finančného trhu sa neskôr primárne zameriame na banky a nebankové subjekty.

„Finančné inštitúcie – subjekty, ktoré vytvárajú finančné nástroje, obchodujú s nimi a sprostredkujú pohyb zdrojov medzi účastníkmi trhu.“ Takto definuje Polouček³⁰ finančné inštitúcie. Finančné inštitúcie sú piliermi finančného systému, a preto je nutné zabezpečiť ich zdravie, hladké fungovanie a efektívnosť. Bez týchto predpokladov by finančný systém vykazoval funkčnú poruchu. Existuje viacero typov inštitúcií, no spoločným menovateľom všetkých z nich je striktná regulácia a intenzívny dohľad, aby nevznikali nedostatky v systéme.³¹

Prakticky každý, kto žije v rozvinutej ekonomike, má neustále alebo aspoň periodickú potrebu využiť služby finančných inštitúcií.³² Finančné inštitúcie sa zaoberajú obchodom s finančnými a peňažnými transakciami ako sú vklady, pôžičky, investície a obchody s národnými menami. Tieto spoločnosti dosahujú rôznu veľkosť od lokálnych úverových zväzov, až po nadnárodné investičné banky. Zatiaľ čo sa niektoré finančné inštitúcie

²⁸ Zákon č. 129/2010 Z. z. Zákon o spotrebiteľských úveroch a o iných úveroch a pôžičkách pre spotrebiteľov a o zmene a doplnení niektorých zákonov

²⁹ INEKO. *Pôsobenie nebankových spoločností na trhu spotrebiteľských úverov na Slovensku*. Marec 2016. s. 5. [elektronický zdroj] Dostupné na: <http://www.ineko.sk/clanky/publikacie>

³⁰ POLOUČEK, Stanislav – KULHÁNEK, Lumír – MIRDALA, Rajmund et al. *Peniaze, banky, finančné trhy*. Bratislava: Iura Edition, 2010. s. 318. ISBN 978-80-8078-305-1.

³¹ BABOUČEK, Ivan a kol. *Regulace činnosti bank*. Praha: Bankovní institut, a.s., 2005. s. 107. ISBN 80-7256-071-8.

³² Tamtiež s. 318.

zameriavajú na poskytovanie služieb pre širokú verejnosť, iné, špecializované, obsluhujú len určitú klientelu. Aby sme vedeli, ktorá finančná inštitúcia je najvhodnejšia na uspokojenie konkrétnych potrieb, je dôležité pochopiť rozdiel medzi typmi inštitúcií a účelmi ktorým slúžia. V ďalšej časti kapitoly prezentujeme Rejnušov³³ pohľad na univerzum finančných inštitúcií.

Slovenská banková sústava je dvojstupňová, kde jeden celý stupeň tvorí jediná inštitúcia – centrálna banka. Ide o inštitúciu, do ktorej individuálni sporitelia nemajú priamy prístup. Centrálna banka nevykonáva klasické komerčné bankovníctvo, ale aktívne spolupracuje s ostatnými obchodnými bankami pri poskytovaní produktov a služieb širokej verejnosti. Naša centrálna banka je integrovaná v nadnárodných štruktúrach, v ktorých zodpovednosť za výkon menovej politiky a dohľad nad finančnými inštitúciami prebrala Európska centrálna banka. Povoľovaciu právomoc na Slovensku má stále v pôsobnosti Národná banka Slovenska. NBS rozhoduje o udelení tak bankovej licencie, ako aj licencie pre nebankové kreditné inštitúcie.

Pod dohľadom centrálnej banky pôsobia ostatné komerčné banky, ktoré sú vnútorne členené podľa predmetu činnosti. Retailové banky tradične ponúkajú produkty individuálnym spotrebiteľom, zatiaľ čo korporátne banky priamo spolupracujú s podnikmi. Hypotekárne banky majú ako predmet podnikania financovanie nehnuteľností. Spravidla však tieto činnosti zastrešuje každá komerčná banka. Cieľom obchodnej banky na rozdiel od centrálnej, je tvorba zisku za účelom uspokojenia investorov.

Množinu finančných inštitúcií rozširujú aj nebankové kreditné subjekty. Tieto spoločnosti sú lákavé hlavne pre klientov, ktorí sa nevedia dostať ku kreditným produktom cez kanály štandardného bankovníctva. Na spektre finančných inštitúcií sa nachádzajú aj investičné banky. Tie neprijímajú vklady ale namiesto toho pomáhajú jednotlivcom, podnikom a vládám získať kapitál prostredníctvom emisie cenných papierov. Vývoj posledných rokov a digitálna doba prispeli k vytvoreniu novej špecifickej bankovej spoločnosti – internetovej banky. Internetová banka netvorí sieť kamenných pobočiek, ale ponúka svoje služby cez mobilné aplikácie a internetové bankovníctvo.

Ďalšou kategóriou sú poisťovne. Subjekty trhu vyhľadávajú poisťovacie spoločnosti na zabezpečenie ochrany pred finančnými stratami v dôsledku úmrtia, zdravotného

³³ REJNUŠ, Oldřich. *Finanční trhy: učebnice s programem na generování cvičných testů*. Praha: Grada Publishing, 2016. s. 54-58. ISBN 978-80-247-5871-8.

postihnutia, nehôd, poškodenia majetku a iných nešťastí. Nemenej dôležitou inštitúciou finančného trhu je sprostredkovateľská spoločnosť. Maklérske firmy pomáhajú jednotlivcom a inštitúciám pri nákupe a predaji cenných papierov. Ekosystém finančných inštitúcií je vskutku rôznorodý a zahŕňa ďalšie spoločnosti ako penzijné či správcovské fondy, lízingové spoločnosti, obchodníkov s cennými papiermi, finančných poradcov ale i zmenárne či záložne a mnohé iné.

1.4.1 Obchodné banky ako finančné inštitúcie

Pri snahe stratifikovať subjekty a inštitúcie tvoriace finančný systém z hľadiska významnosti prichádzame na to, že banky sú subjektami, ktorých činnosť má zásadný vplyv na funkčnosť finančného systému. „Banky významnou mierou ovplyvňujú aktivity jednotlivých ekonomických subjektov v národnom hospodárstve,“³⁴ a preto sú podľa Tkáčovej práve banky najdôležitejšie inštitúcie spomedzi všetkých, ktoré tvoria finančný systém.

Bankovú sústavu popri centrálnej banke so svojím charakteristickým postavením tvorí rad ostatných obchodných bánk. Tieto inštitúcie sa zameriavajú prioritne na klasické bankové produkty, medzi ktoré radíme prijímanie vkladov, poskytovanie úverov a výkon platobného styku. Rovnako majú skrz svoje úverové linky prístup k likvidite z centrálnej banky, čo im ponúka komparatívnu výhodu oproti ostatným finančným inštitúciám.

Obchodné banky zabezpečujú tri podstatné funkcie, a to finančné sprostredkovanie, emitovanie bezhotovostných peňazí a platobný styk.³⁵ Sprostredkovanie sme už v predchádzajúcom texte definovali ako presun peňažných prostriedkov od deficitných k prebytkovým subjektom. V kontexte bankového systému tento transfer prebieha premenou prijatých depozít na emitované úvery. V nadväznosti na tento proces hovoríme aj o funkcii emitovania bezhotovostných peňazí, keďže proces úverovania a prijímania vkladov sa prejavuje digitálnymi zápismi na rôznych účtoch bez potreby vytvorenia hotovostných peňazí. Po prijatí úveru dlžník disponuje sumou, ktorá patrí pôvodnému vkladateľovi, pričom vkladateľ má stále nárok na vložený obnos peňazí. Bankový sektor týmto spôsobom emituje dodatočné prostriedky. Plynulý chod hospodárstva podporuje aj elektronický platobný styk. Bezhotovostný styk odbúrava náročnosť presunu veľkých objemov hotovosti od subjektu k subjektu.

³⁴ TKÁČOVÁ, Dana a kol. *Finančné trhy a bankovníctvo*. Bratislava. Wolters Kluwer, 2017. s. 137. ISBN 978-80-7552-528-4.

³⁵ REJNUŠ, Oldřich. *Finanční trhy: učebnice s programem na generování cvičných testů*. Praha: Grada Publishing, 2016. s. 57. ISBN 978-80-247-5871-8.

Banka poskytuje prostriedky vo forme úverov z prijatých vkladov od klientov a nie nevyhnutne z vlastných zdrojov. Úvery majú maturitu spravidla na dlhšie obdobie, čo pridáva na rizikovosti celej tejto schémy. Na zabezpečenie udržateľnosti vytvoreného modelu boli banky nútené vyvinúť a implementovať nástroj na ohodnotenie kreditnej kvality žiadateľa v snahe znížiť informačnú asymetriu.³⁶ Pod informačnou asymetriou rozumieme, že jedná zo zmluvných strán, v tomto prípade klient má viac alebo presnejšie informácie o podstúpenom riziku. Naliehavosť práce s rizikom dokladuje aj intenzita monitorovania tohto aspektu bankového podnikania regulačnými subjektmi. Spomenuli sme, že nutným predpokladom k poskytnutiu úveru je preskúmanie kreditnej kvality žiadateľa o úver. Úver je emitovaný až keď klient spĺňa minimálne podmienky odpovedajúce rizikovému apetítu banky. Pri významných úveroch často kontrola pokračuje a banka skúma, či klient nepodstupuje nadmerné riziká. Klient úver v čase spláca pravidelnými splátkami, čím zabezpečuje pre banku príjem. Rozdiel medzi cenou, za ktorú poskytne banka úver a cenou, ktorú platí banka za zdroje, je maržou banky.

Banka vykonáva niekoľko funkcií a takmer všetky sú smerované ku generovaniu zisku. Náplňou tradičného bankového podnikania je úverovanie. Dostupné sú však aj iné možnosti ako generovať zisk, či už sú to poplatkové úkony pre klientov alebo obchodovanie na vlastné meno. Celý diapazón obchodných aktivít banky je neoddeliteľne spojený s rizikom. Banky preto využívajú na monitorovanie svojej celkovej činnosti súbor modelov riadenia rizík.

1.4.2 Nebankové subjekty – ostatní finanční sprostredkovatelia

Nebankové finančné inštitúcie (NFI) hrajú tiež dôležitú úlohu vo finančnom systéme bez ohľadu na to, že názor verejnej mienky je pomerne negatívny.³⁷ Svojím podnikaním zdokonaľujú finančnú infraštruktúru tým, že zjednodušujú prístup k finančným prostriedkom. NFI vedia totiž obslúžiť aj subjekty, ktorých potreby nie sú uspokojené konvenčným bankovým systémom.

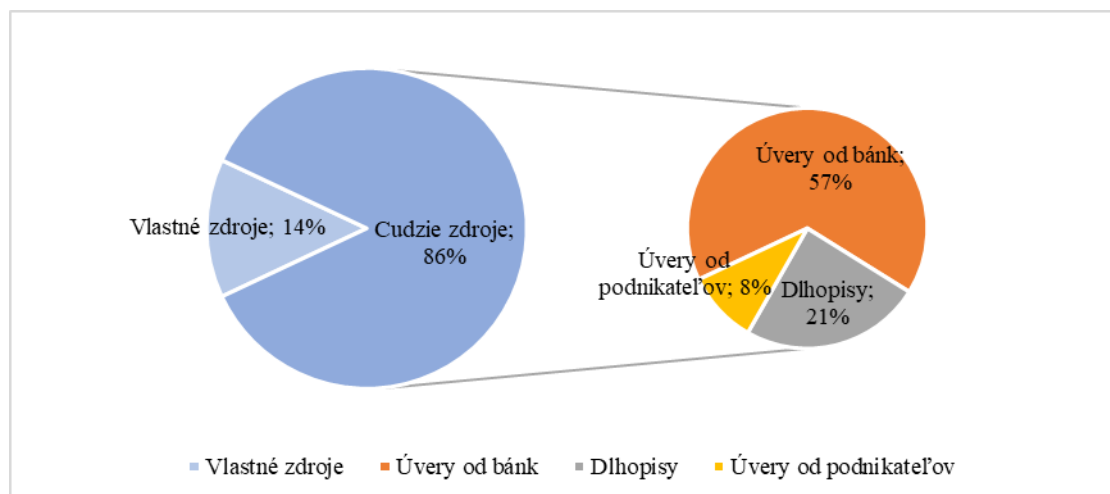
Spoločnosti realizujúce finančné sprostredkovanie nebankovým spôsobom definujeme ako „*finančné korporácie, ktorých predmetom podnikania je finančné sprostredkovanie na základe prijímania záväzkov v iných formách ako je hotovosť, vklady a/alebo blízke náhrady*

³⁶ DE HAAN, Jakob - OOSTERLOO, Sander - SCHOENMAKER, Dirk. *European financial markets and institutions*. Cambridge: Cambridge University Press, 2009. p. 8-9. ISBN 978-0-511-50686-4.

³⁷ INEKO. *Pôsobenie nebankových spoločností na trhu spotrebiteľských úverov na Slovensku*. Marec 2016. s. 5. [elektronický zdroj] Dostupné na: <http://www.ineko.sk/clanky/publikacie>

vkladov od inštitucionálnych jednotiek.“³⁸ Z uvedenej definície vyplýva, že primárny rozdiel medzi bankovým a nebankovým poskytovateľom úverov a pôžičiek je vo forme získavania zdrojov. Banky sú predmetom špecifickej právnej úpravy³⁹ a podliehajú prísnejšej regulácii a intenzívnejšiemu dohľadu. Banky sú na rozdiel od NFI oprávnené prijímať vklady a tým získavať disponibilné zdroje. Model financovania NFI je preto o niečo odlišný, pričom cudzie zdroje stále hrajú kľúčovú úlohu, hoci už iným spôsobom. Za tretí kvartál roku 2019⁴⁰ tvorili vlastné zdroje v nebankovom sektore na Slovensku iba 14%, pričom sa skladali zo základného imania iných majetkových účasti či nerozdeleného zisku. Zvyšných 86% tvorili cudzie zdroje, ktoré mali prevažne charakter prijatého úveru od obchodných bánk. Prijaté úvery tvorili 65,86% z celkových cudzích zdrojov. Ostatné cudzie zdroje boli vytvorené emisiou dlhopisov (24,39%) alebo úvermi od podnikateľských subjektov (9,75%).

Obrázok 1 Zobrazenie zastúpenia rôznych foriem financovania nebankového sektoru



Zdroj: Vlastné spracovanie podľa štatistického bulletinu za 3. štvrťrok 2019 publikovaného NBS

Nebankové subjekty pôsobiace v podmienkach Slovenskej republiky sú lízingové a faktoringové spoločnosti a spoločnosti splátkového financovania. Najvýznamnejšie postavenie, z pohľadu objemu poskytnutého finančného krytia, majú lízingové spoločnosti. Lízingové služby pomáhajú klientom zaobstarať si hnutel'ný majetok, ak aktuálne nedisponujú dostatočnou sumou na priamy nákup. Najčastejším hnutel'ným majetkom sú motorové vozidlá, výrobné stroje alebo výpočtová technika. V tomto odvetví podnikajú nepriamo aj banky spravidla cez špecializované právnické osoby, ktoré vystupujú ako

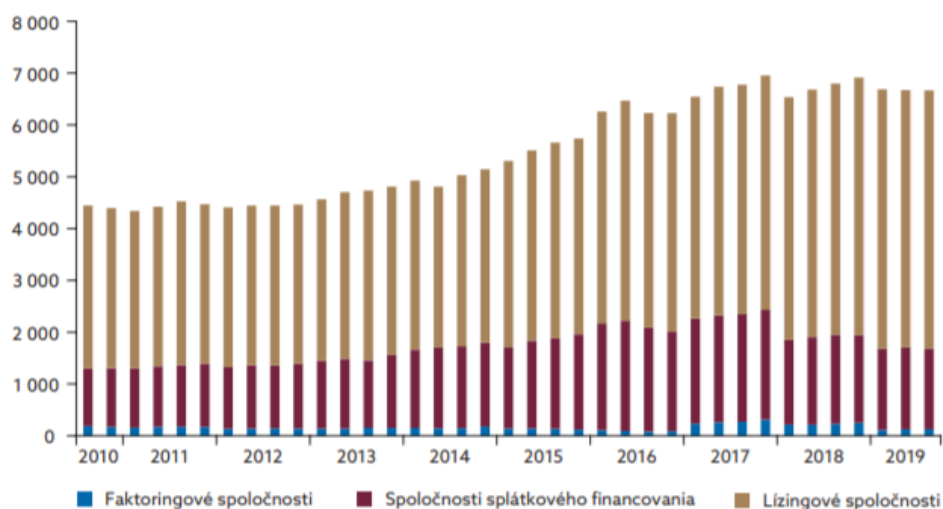
³⁸ Bod 2.86 z dokument NARIADENIE EURÓPSKEHO PARLAMENTU A RADY (EÚ) č. 549/2013 z 21. mája 2013 o európskom systéme národných a regionálnych účtov v Európskej únii. Online: http://www.nbs.sk/_img/Documents/_Statistika/manual_ESA_2010_SK.pdf

³⁹ Zákon č. 314/2011 Z. z. ktorým sa mení a dopĺňa zákon č. 483/2001 Z. z. o bankách

⁴⁰ Štatistický bulletin 3. štvrťrok 2019. [elektronický zdroj] Národná banka Slovenska. ISBN 1338-6328. Dostupné na: http://www.nbs.sk/_img/Documents/_Publikacie/BulletinMFS/2019/BulletinMFS_092019.pdf

nebankové spoločnosti, no súčasne sú zložkou bankového holdingu.⁴¹ Druhými v poradí v objeme emitovaných úverov sú spoločnosti splátkového financovania, ktorých predmetom podnikania je poskytovanie úverov, kreditných kariet alebo iných kreditných produktov, ktoré majú podobu úveru. Ich celková expozícia je približne 1,5 mld. eur. Najmenší podiel na trhu majú nebankové faktoringové spoločnosti, ktoré sa venujú odkupovaniu obchodných pohľadávok.

Obrázok 2 Vývoj aktív spoločností poskytujúcich úvery a pôžičky nebankovým spôsobom (mil. EUR)



Zdroj: Prevzaté zo štatistického bulletinu za 3. štvrťrok 2019 publikovaného NBS

Pri pohľade na údaje pochádzajúce z Národnej banky Slovenska⁴² vidíme jasný klesajúci trend v objeme poskytnutých úveroch splátkovými spoločnosťami. Tento stav zachytáva *Tabuľka 1* s medziročnými zmenami aktív.

⁴¹ Napríklad VÚB leasing, ktorý je súčasťou skupiny okolo VÚB banky, či Tatra leasing ako zložka Tatra banka Group.

⁴² Štatistický bulletin 3. štvrťrok 2019. [elektronický zdroj] Národná banka Slovenska. ISBN 1338-6328. Dostupné na: http://www.nbs.sk/_img/Documents/_Publikacie/BulletinMFS/2019/BulletinMFS_092019.pdf

Tabuľka 1 Medziročná zmena aktív lízingových spoločností, spoločností splátkového financovania a faktoringových spoločností.

Celkové aktíva	Medziročná zmena v %								
	IX.	XII.	III.	VI.	IX.	XII.	III.	VI.	IX.
	2017	2017	2018	2018	2018	2018	2019	2019	2019
Lízing	6.88	7.42	9.54	8.30	9.39	9.99	6.89	4.01	3.03
Faktoring	225.71	251.11	-7.64	-16.62	-13.88	-19.82	-46.62	-40.86	-45.13
Splátkové financovanie	3.79	9.87	-19.47	-18.37	-17.21	-20.20	-4.55	-6.50	-10.00

Zdroj: Vlastné spracovanie podľa štatistického bulletinu za 3. štvrťrok 2019 publikovaného NBS

1.4.3 Úverové registre ako ochrana pred rizikom

Pred úverovým rizikom sa kreditné inštitúcie chránia rôznymi spôsobmi. Klient má voči úverovej inštitúcii vždy výhodu vo forme informačnej asymetrie. Klient podrobnejšie pozná riziko, ktoré mieni podstúpiť, vie či je ochotný daný záväzok aj skutočne vysporiadať a má najlepší prehľad o všetkých úveroch, ktoré čerpá.⁴³ Práve posledný bod, prehľad o aktuálnych úveroch klienta, sa snaží adresovať inštitút úverového registra.

Existuje bankový úverový register, ktorý agreguje údaje o záväzkoch jedincov voči bankovým domov, o ručiteľoch aj poskytnutých zárukách. Oficiálny názov registra je Spoločný register bankových informácií (SRBI).⁴⁴ Bankové inštitúcie sú povinné informovať dohľad o vzniku dlžníckeho vzťahu.⁴⁵ Banky touto cestou zdieľajú medzi sebou informácie o platobnej schopnosti a disciplíne. Banky môžu vstúpiť a vyhľadávať osoby v tomto registri iba so súhlasom daného klienta/klientov, keďže informácie v ňom sú vysoko citlivé.

Ekvivalentným nástrojom na zmiernenie rizika v nebankovom sektore je Nebankový register klientskych informácií (NRKI).⁴⁶ Register obsahuje údaje o úspešnosti splácania úverov a pôžičiek poskytnutých NFI. Tento register si vytvorili nebankové subjekty, aby boli medzi sebou lepšie informované o bonite klientov a ich schopnosti splácať. NRKI je vedený a

⁴³ DE HAAN, Jakob - OOSTERLOO, Sander - SCHOENMAKER, Dirk. *European financial markets and institutions*. Cambridge: Cambridge University Press, 2009. p. 8-9. ISBN 978-0-511-50686-4.

⁴⁴ Slovak Banking Credit Bureau [elektronický zdroj]. Dostupné na: <http://www.sbc.sk/o-spolocnosti/spolocny-register-bankovych-informacii-srbi/>

⁴⁵ BAČA, Jozef. *Bankový dohľad*. Vydavateľstvo EKONÓM, Bratislava, 2001, s. 105, ISBN 80-225-1377-6

⁴⁶ HYRÁNEK, Eduard - BIKÁR, Miloš., *Manažment úverového financovania a úverových obchodov*. Vydavateľstvo EKONÓM, Bratislava, 2010, s. 152, ISBN 978-80-225-3042-2

editovaný lízingovými spoločnosťami. Splátkové spoločnosti majú svoj samostatný register, ktorý má názov SOLUS.⁴⁷ Tento register zaznamenáva iba negatívne informácie o klientoch.

Nevýhodou systému úverových registrov je nulový informačný tok medzi bankovými a nebankovými registrami. Slovenská právna úprava nedovoľuje informáciám z bankového sektora prienik von. Na druhú stranu, v Českej republike takáto spolupráca úspešne funguje.

1.5 Regulácia finančného systému

Nezvládnutá práca s rizikom je častou príčinou situácií, v ktorých banky zaznamenali závažné straty.⁴⁸ Takýto stav môže viesť v krajnom prípade až k uzavretiu podnikateľskej činnosti. Ani prostredie finančného sveta nie je odolné voči pôsobeniu rizík. Druhým dôležitým bodom pri nutnosti zaviesť reguláciu je sociálny aspekt zlyhania finančných inštitúcií. Tie operujú prevažne s cudzím kapitálom, či už zo súkromného alebo podnikateľského prostredia. Úpadok takejto entity by viedol k stratám na strane subjektov, ktoré neprevádzkovali takúto činnosť a dôvera vo finančný systém by bola narušená.⁴⁹ Pri poskytovaní úverov je rovnako nutná aj ochrana spotrebiteľa – klienta. Tá smeruje k regulácií podmienok, aby dlžník nečelil neodôvodnene vysokým úrokom, nevýhodným podmienkam či komplikovaným a nezrozumiteľným podmienkam.

Protipólom k silnejúcej regulácií sú napríklad závery Revendu⁵⁰, ktorý poukazuje na finančnú náročnosť regulácie. Banky stoja pred voľbou, či premietnuť zvýšené náklady do zdraženia svojich služieb, alebo do poklesu ziskovosti. Otázna je podľa neho aj efektívnosť regulácie. Medzery vo formulácií pravidiel paradoxne viedli k vzrastajúcim preferenciám kupovať rizikovejšie aktíva.

Modernú podobu bankovej regulácie na globálnej úrovni definuje primárne Bazilejský výbor pre bankový dohľad (BCBS), ktorý si dal za cieľ dosahovať finančnú stabilitu prudentnou reguláciou nad úverovými a investičnými inštitúciami. V podmienkach Európskej Únie vykonáva jednotný dohľad ECB v spolupráci s národnými regulátormi, napr. NBS. Nad finančným trhom v EÚ komplexne dohliada The European system of financial supervision

⁴⁷ ONUFEROVÁ, Marianna [elektronický zdroj] Denník N. Dostupné na: <https://e.dennikn.sk/63329/zoznamy-neplaticov-ako-sa-dostanete-dnu-ako-von/>

⁴⁸ BECK, Thorsten – LAEVEN, Luc. Resolution of failed banks by deposit insurers : cross-country evidence. In Policy, Research working paper ; no. WPS 3920. Washington, DC: World Bank.. May, 2006. p. 3

⁴⁹ BABOUČEK, Ivan a kol. *Regulace činnosti bank*. Praha: Bankovní institut, a.s., 2005. s. 107. ISBN 80-7256-071-8.

⁵⁰ REVENDA, Zbyněk. *Centrální bankovníctví*. 3. aktualiz. vyd. Praha: Management Press, 2011. s. 369. ISBN 978-80-7261-230-7.

(ESFS), ktorý bol uvedený do pôsobnosti v roku 2010. Problematika bankovej regulácie je nesmierne obsiahla a vystupuje nad rámec tejto práce.

V praktickej časti analyzujeme dáta o úveroch pochádzajúce z nebankovej spoločnosti Home Credit. V tomto odseku sa preto podrobnejšie zameriame na reguláciu NFI v podmienkach Slovenskej republiky. V roku 2015 Národná banka Slovenska prevzala kontrolu nebankových subjektov pod svoju pôsobnosť.⁵¹ NFI, ktoré chceli aj naďalej prevádzkovať svoju podnikateľskú činnosť týmto museli požiadať o pridelenie licencie. Pred licenčným procesom pôsobilo na trhu viac ako 270 takýchto inštitúcií.⁵² Licenčný proces podstúpilo nakoniec len pár desiatok z nich. Následkom toho dnes na trhu nepôsobí viac ako 32 subjektov.⁵³ Jedným z dôvodov je prísna právna regulácia v segmente spotrebiteľských úverov, ako je napríklad strop miery odplaty veriteľovi. Nebankové subjekty na rozdiel od bankových domov, výrazne častejšie poskytovali malé sumy v stovkách eur. Po zavedení novej právnej úpravy by výnosy z takýchto úverov predstavovali len niekoľko centov. Nebankový sektor sa hlbkovej právnej regulácií dlho bránil argumentmi, že na rozdiel od bankových subjektov, oni nehazardujú s vkladmi klientov, ale investujú svoj vlastný podnikateľský kapitál. Táto námietka bola do určitej miery pravdivá, no neuspela.

Proces vystavovania povolení pre činnosť NFI sa síce výrazne sťažil, no stále je tu pomerne veľký rozdiel v regulácií a dohľade medzi bankou a nebankovou inštitúciou. Banky, ako sme spomenuli, používajú na úverovú činnosť prevažne vklady obyvateľstva, čo vedie k potrebe regulovať aj mieru rizík, ktorú banky podstupujú. Úmyslom je chrániť záujmy vkladateľov. Bazilejské dohody, konkrétne aktuálne platný Basel III, diktuje bankám aké riziká a do akej výšky majú kryť vlastnými zdrojmi. Banka smie modelovať vývoj rizík iba regulátorom schválenými technikami. Tieto techniky musia spĺňať požiadavky na jednoduchú interpretáciu výsledku, ktorým dospeli k záveru o bonite klienta. Pre nebankové splátkové inštitúcie takéto náležitosti nie sú nutné a môžu používať modely aj s nižšou až žiadnou mierou interpretovateľnosti. Tento fakt je kľúčový pre praktickú časť diplomovej práce. Neobmedzuje nás totiž vylúčiť akýkoľvek nástroj z portfólia strojového učenia.

⁵¹ NBS [elektronický zdroj]. Správa o činnosti Útvary dohľadu nad finančným trhom Národnej banky Slovenska. Dostupný na: https://www.nbs.sk/_img/Documents/_Dohlad/Makropolitika/SC-UDF_2015_web.pdf

⁵² FUSEK, Roman. [elektronický zdroj] Televízna stanica TA 3, Ekonomika, téma: Sú nebankovky "čistejšie"? Dostupné na: http://www.nbs.sk/_img/Documents/_Rozhovory/2018/Fusek_TA3_19_9.pdf

⁵³ NBS [elektronický zdroj] Dostupné na: https://subjekty.nbs.sk/&ll=en?aa=select_sector&bb=7&cc=&qq=

1.6 Manažment rizika

Ekonomická aktivita neprináša subjektom, ktoré ju vykonávajú len benefity ale zákonite aj široké spektrum rizík. Nevyhnutným predpokladom úspechu je schopnosť efektívne pracovať s rizikom. Termín, ktorý označuje prácu s rizikom je pojem manažment rizika. „*Manažment rizika je proces identifikácie, ohodnotenia a kontrolovania rizík, ktoré ohrozujú investora a jeho zisky.*“⁵⁴ Tieto riziká majú spravidla rôznych pôvodcov.

Dôraz na optimálnosť práce s rizikom sa kladie z dvoch dôvodov. Jedným z nich je snaha samotného subjektu zlepšiť výkonnosť svojej ekonomickej činnosti. Druhým dôvodom je snaha o vytvorenie si dobrého mena a reputácie silného subjektu smerom ku klientom a obchodným partnerom.⁵⁵

Bežný život prináša viacero typov rizík. V ekonómii sa pri uvažovaní o riziku dostávame skôr k spojeniu finančné riziko.⁵⁶ Tento pojem odkazuje na potenciálnu stratu v budúcnosti, ktorá však pramení z aktuálne podstupovaných rizík. Podstatou rizika je, že strata môže a nemusí nastať, z čoho následne vychádza proces riadenia rizík.

1.6.1 Prístupy k riadeniu rizík

Medzi najznámejšie prístupy k riadeniu rizík radíme tieto štyri: vyhýbanie sa riziku, redukcia rizika, transfer rizika a nesenie rizika.⁵⁷ Vyvarovanie sa riziku je jedným z najjednoduchších spôsobov boja proti negatívnym udalostiam. Zastaviť všetky činnosti, ktoré by mohli ohroziť podnikanie, žiaľ, nie je možné. Aktivita, ktoré sú súčasťou podnikania a ktoré nie je možné zastaviť je potrebné vykonávať formou, ktorá minimalizuje riziká. Zároveň však musíme zvážiť rentabilitu takýchto obmedzení činnosti. Prenos rizika na iný subjekt za odplatu je realistický pohľad na vec, ktorý uznáva, že sa incidenty niekedy vyskytnú. Pri prenose rizika vzniknuté škody znáša predajca ochrany, pričom je nutné zvážiť náklady na transfer rizika. Možnosť znášať riziko individuálne si často volia tí, ktorí považujú náklady na prenos alebo zníženie rizika za neprimerané alebo zbytočné. Realizované škody sú plne v rézii nositeľa.

⁵⁴ ROUSE, Margaret [elektronický zdroj] Dostupné na: <http://searchcompliance.techtarget.com/definition/risk-management>

⁵⁵ SIVÁK, Rudolf - GERTLER, Ľubomíra - KOVÁČ, Urban. *Riziká a modely vo financiách a v bankovníctve*. 2. vyd. Bratislava: Sprint dva, 2010. s. 71. ISBN 978-80-89393-44-2.

⁵⁶ SIVÁK, Rudolf – GERTLER, Ľubomíra – KOVÁČ, Urban. *Riziká vo financiách a v bankovníctve*. Bratislava: Sprint dva, 2009. s. 54. ISBN 978-80-89393-03-9

⁵⁷ REIM, Wiebke. et al. Risk management for product-service system operation. In *International Journal of Operations & Production Management*. June, 2016. ISSN 0144-3577.

1.6.2 Riadenie rizík v bankovom sektore

Preberanie rizika je fundamentálny predmet podnikania v bankovníctve.⁵⁸ Iba preberaním kalkulovaného finančného rizika môže banka vyprodukovať mieru návratnosti nad úrovňou bezrizikovej miery. Prvým krokom je rozkladať riziká na jednotlivé prvky a podľa toho ich riadiť. Napríklad, úver s fixnou úrokovou mierou je vhodné rozdeliť na riziko plynúce zo zmeny úrokovej miery a na riziko kreditné. Následne vieme tieto parciálne riziká riešiť individuálne. Riziko zmeny úrokovej miery môžeme eliminovať nákupom úrokových derivátov (napr. swapov) a kreditné riziko minimalizovať založeným kolaterálom.

Hlavné typy rizík v bankovom portfóliu sú kreditné, trhové, operačné a likvidné riziko. Kreditnému riziku sa budeme venovať v samostatnej časti. Trhové riziko definuje Sivák ako „riziko, že podnikateľský subjekt zaregistruje stratu na základe opačného pohybu cien finančných aktív na trhu ako očakával.“⁵⁹ Trhové riziko zahŕňa ďalšie špecifické riziká ako menové, úrokové, akciové a komoditné. Operačné riziko nie je jednoduché zadefinovať, vo všeobecnosti však „vyplýva zo zložitosti obchodných a prevádzkových operácií finančnej inštitúcie.“⁶⁰ Riziko likvidity sa naplní keď inštitúcia nebude schopná získať zdroje v požadovanej mene a dostatočnom objeme pri rozumnej úrovni nákladov, aby mohla vyrovnať svoje záväzky vtedy, keď nastanú splatné.

1.6.3 Kreditné riziko v kontexte bankovníctva

Kreditné riziko⁶¹ je finančné riziko vyplývajúce z možnej neschopnosti alebo neochoty dlžníka splácať svoje záväzky. Vlastníctvo aktív, ktoré sú splatné v budúcnosti, znamená pre veriteľa neistotu. V období medzi poskytnutím úveru a splatením existuje možnosť, že dlžník nebude schopný naplniť podmienky zmluvy a veriteľ zaknihuje stratu. K strate môže dôjsť v dôsledku:

- Omeškania dlžníka k určitému času v určitej výške (priame kreditné riziko).
- Zmenou úverovej kvality (ratingu) dlžníka a následnou zmenou trhovej ceny aktíva.

⁵⁸AVDJIEV, Stefan – GIUDICI, Paolo – SPELTA, Alessandro. Measuring contagion risk in international banking. [elektronický zdroj] BIS Working Papers, No. 796. June, 2019. Dostupné na: <https://www.bis.org/publ/work796.pdf>

⁵⁹SIVÁK, Rudolf – GERTLER, Ľubomíra – KOVÁČ, Urban. *Riziká vo financiách a v bankovníctve*. Bratislava: Sprint dva, 2009. s. 183. ISBN 978-80-89393-03-9

⁶⁰Tamtiež s. 217.

⁶¹Tamtiež s. 151-181.

V globálnom svete existujú prípady, keď kreditné riziko nie je odvodené iba od protistrany, ale aj od krajiny, v ktorej dlžník pôsobí. Vyhodnocujeme dva typy rizík:

- Platby od dlžníka sú oneskorené (úverové riziko).
- Dlžník zažije makroekonomický stres (systémové riziko).

Banka povinne musí držať určitý objem kapitálu na absorbovanie potenciálnych strát. Veľkosť tejto kapitálovej požiadavky priamo úmerne závisí od úrovne a veľkosti rizík, ktoré podstupuje. Veľká pozornosť sa kreditnému riziku venuje najmä preto, že hrá dôležitú úlohu pri výpočte požadovaného objemu regulačného kapitálu. Keďže hlavnou činnosťou banky je poskytovanie úverov klientom, kreditné riziko má vo výpočte rizikovo-vážených aktív (RVA) vysoký podiel. Druhý pojem, ktorý je nutné si predstaviť je ekonomický kapitál (economic capital). Pod týmto termínom rozumieme objem kapitálu, ktorý banka sama považuje za adekvátny na absorbovanie všetkých potenciálnych strát v určitom časovom horizonte a na definovanom intervale spoľahlivosti. O úrovni ekonomického kapitálu na rozdiel od regulačného kapitálu (regulatory capital) sa rozhoduje banka individuálne po zohľadnení všetkých svojich špecifik. Podmienky vyčíslenia regulačného kapitálu sú naopak, presne definované regulátorom a ich držba je podmienená možnosťou odobrania bankovej licencie. Pravidlá výpočtu regulačného kapitálu sú plošne dané a záväzné pre všetky bankové domy.

Ak chce byť banka zisková potrebuje zabezpečiť dve veci. Prvou je držba ekonomického kapitálu na minimálnej nutnej úrovni, aby mohla svoje zdroje zhodnocovať. Na dosiahnutie tohto cieľa potrebuje predikčný model strát, ktorý ich priblíži čo najbližšie k realite. Druhým krokom je minimalizácia samotných strát na portfóliu. V tomto bode banka vyvíja modely, ktoré odhadujú rizikovosť jednotlivých obchodných aktivít.

1.7 Strojové učenie

Strojové učenie známe skôr pod anglickým termínom Machine Learning (ML) sa vo zvýšenej miere dostáva do centra spoločenskej debaty či už odbornej obci alebo širokej verejnosti. Pomerne populárne pojmy („buzz words“) ako umelá inteligencia a strojové učenie stále znejú ako z budúcnosti, no už dnes si vo veľkej miere nachádzajú uplatnenie v praxi a pomáhajú ľuďom zjednodušovať a zdokonaľovať svoje procesy.

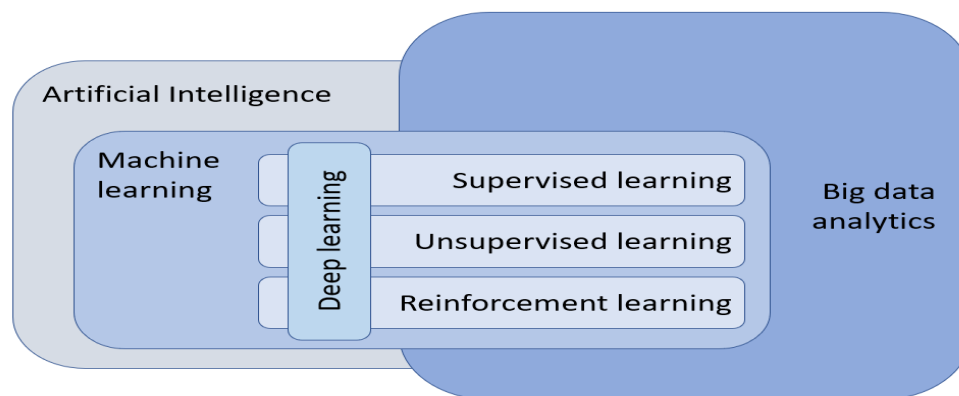
Strojové učenie vyrástlo ako oblasť matematiky a štatistiky a za podpory rozvoja IT technológií sa stalo jedným z motorov priemyslu 4.0. Účelom jeho využitia je získať poznatky

z veľkých a ťažko interpretovateľných súborov údajov, ktoré nazývame Big Data.⁶² Tieto techniky vykonávajú (lepšie povedané simulujú) činnosť predtým typickú len pre človeka – učenie sa zo skúseností. Počítače majú oproti ľuďom komparatívnu výhodu, keďže odhaľovať vzťahy a súvislosti v dátach dokážu násobne rýchlejšie, vďaka veľkej výpočtovej sile.

Aplikácia výpočtovej techniky na riešenie úloh, ktoré si nutne vyžadovali ľudský intelekt sa všeobecne nazýva umelá inteligencia (artificial intelligence – AI). AI ako vedná disciplína existuje niekoľko desiatok rokov, no jej potenciál bol dlho nevyužitý.⁶³ Až nedávna revolúcia v raste výkonnosti techniky, dostupnosti údajov a klesajúcej cene za ich uskladnenie, viedla k obnoveniu záujmu o sféru umelej inteligencie.⁶⁴ S AI sa stretávame pri diagnostike chorôb, prekladaní jazykov či riadení vozidiel a významne preniká aj do finančného sektora.

Strojové učenie na vykonávanie konkrétnej úlohy nepotrebuje mať vopred explicitne definované kroky, ale iba základný rámec ako úlohu riešiť. Takéto modely sa približujú k optimálnemu nastaveniu parametrov na základe skúseností, ktoré načerpajú z predložených dát s obmedzeným alebo žiadnym ľudským zásahom.⁶⁵

Obrázok 3 Schematický prehľad sférou umelej inteligencie



Zdroj: Spracované podľa FSB (2017). Artificial intelligence and machine learning in financial services. November.

⁶² „Big Data“ je výraz, pre ktorý neexistuje jednotná a konzistentná definícia, ale vo všeobecnosti sa používa na opis veľkých a/alebo komplikovaných súborov údajov a ich analýzu pomocou rôznych techník vrátane AI.

⁶³ Architektúra umelých neurónových sietí bola prvýkrát prezentovaná v roku 1943 v diele „A logical calculus of ideas immanent in neural activity.“ V publikácii bol prezentovaný zjednodušený matematický model opisujúci spoločnú prácu biologických neurónov pri vykonávaní komplexných operácií s využitím výrokovej logiky. Od tohto momentu vznikalo mnoho iných neurónových architektúr až pokým v roku 1960 nebolo evidentné, že tento potenciál nedokáže byť v praxi využitý. Zvrat nastal až začiatkom 21. stor. s rozvojom technickej infraštruktúry.

⁶⁴ Financial Stability Board.[elektronický zdroj] *Artificial intelligence and machine learning in financial services*. November, 2017. Dostupné na: <https://www.fsb.org/wp-content/uploads/P011117.pdf>

⁶⁵ KOZA, John R. et al. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In *Artificial Intelligence in Design '96*. Springer, Dordrecht. p. 151–170. ISBN 978-94-009-0279-4

Existuje niekoľko disciplín strojového učenia, ktoré delíme podľa miery zásahu človeka:⁶⁶

- Pri „supervised learning“ sa algoritmu predložia predpripravené dáta, ktoré obsahujú cieľové označenia. Napríklad model, ktorý má za cieľ rozpoznávať podvodné transakcie potrebuje dáta, v ktorých sú fraudulentné transakcie unikátne označené. Model prisudzuje váhy jednotlivým charakteristikám s cieľom detegovať podvodné transakcie a tieto váhy použije na klasifikáciu nových transakcií, ktoré predtým nevidel.
- „Unsupervised learning“ je metódou učenia, kedy sa modelu nedodávajú označené dáta. Algoritmus vytvára zhľuky pozorovaní, ktoré závisia od podobných základných charakteristík. Algoritmus napríklad môže rozdeliť cenné papiere do klastrov podľa ich podobností. Všetky inštrumenty v klastri potom vieme oceniť rovnakou cenou.
- „Reinforcement learning“ sa nachádza na pomedzí týchto dvoch tried. Modelu sú poskytnuté neoznačené dáta a algoritmus sa rozhodne vykonať nejakú akciu pre každý dátový bod. Na vykonanú akciu sa vzťahuje spätná väzba (pravdepodobne od človeka), ktorá pomáha algoritmu učiť sa. Tento model je kľúčový pre samoriadiace automobily.
- „Deep learning“ je metóda tvorby modelu inšpirovaná štruktúrou a funkciou mozgu. Programujú sa umelé neurónové siete, ktoré následne vedú zvládať úlohy z ktorejkoľvek hore uvedenej kategórie. Táto metóda je veľkým prísľubom do budúcnosti.

Je dôležité uvedomiť si aj nedostatky a limitácie týchto techník. Jednou z nich je to, že model nedokáže určiť kauzalitu vzťahov. Vzťahy v dátach, ktoré strojové učenie identifikuje, sú iba korelácie, z ktorých niektoré sú ľudským okom nerozpoznatelné. Ekonomovia a finančníci však stále viac používajú strojové učenie, aby lepšie pochopili zložité vzťahy spolu s inými nástrojmi a svojou odbornou znalosťou danej oblasti.

1.7.1 Modelovanie kreditného rizika pomocou strojového učenia

Strojové učenie diskutované v predchádzajúcej časti dosiahlo v mnohých doménach výkonnosť na ľudskej úrovni. V oblasti poskytovania úverov však stále dominujú modely

⁶⁶ GÉRON, Aurelien. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. 2st Edition, O'Reilly Media, Inc. p. 8-15. ISBN 978-1-492-03264-9.

založené na expertíze (rule-based models). Podľa Munkhdalaia⁶⁷ je predpokladom k zavádzaniu nových metód stanovenie zmysluplného porovnávacieho štandardu, ktorým by bol model založený na ľudskej expertíze. Následne by sa zostrojený ML model porovnával voči tomuto štandardu na dátach, ktoré odrážajú úverovú realitu.

Najsilnejšou stránkou ML je schopnosť spracovať veľký objem dát za pomerne krátky čas, čo by mohlo pomôcť zvládnuť širší rozsah vstupov. Druhým pozitívom ML je zvládanie neštruktúrovaných dát, napr. dáta zo sociálnych sietí, čo umožní aj ľuďom bez úverovej histórie dostať sa k úveru a zvýši sa počet jednotlivcov, pre ktorých je možné merať úverové riziko.

Zložité algoritmy, ktoré sú postavené na prácu s veľkými balíkmi dát, môžu mať za následok zníženú transparentnosť. Tento aspekt „čiernych skriniek“ (black-box algoritmov) spravidla sťažuje možnosť poskytnúť spotrebiteľom, audítom a orgánom dohľadu vysvetlenie výsledného úverového rozhodnutia. K tomu sa pridávajú aj obavy či kombinácie dodatočných charakteristík dlžníka nemôžu viesť k určeniu rasy alebo pohlavia žiadateľa. Ide o faktory, ktoré podľa niektorých legislatív nie je možné použiť v modelovaní. Takéto algoritmy, by mohli ohodnotiť dlžníka z etnických menšín nižším ratingom a navrhnúť menej priaznivé podmienky úveru.

FSB⁶⁸ považuje za náročné vyniesť všeobecný súd o úspešnosti modelov využívajúcich strojové učenie, keďže predikčná schopnosť týchto modelov bola testovaná spravidla len v experimentálnych podmienkach v rámci vedeckých štúdií. Rovnako je táto oblasť dosť mladá nato, aby sme určili výkonnosť modelov v rôznych ekonomických cykloch.

1.7.2 *Prehľad literatúry*

Snaha aplikovať výdobytky samoučiaceho modelovania do oblasti kreditného skórovania nie je úplne nová. Prvé snahy sa objavili v roku 2003, kedy Baesens a kolektív⁶⁹ začali analyzovať účinnosť vybraných klasifikačných techník. Z preštudovanej literatúry sa zdá, že inteligentné systémy spĺňajú predpoklady na prekonanie tradičných techník.

⁶⁷ MUNKHDALAI, Lkhagvadorj et al. An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. In *Sustainability MDPI Journals*, vol. 11. January, 2019. p. 2

⁶⁸ Financial Stability Board.[elektronický zdroj] *Artificial intelligence and machine learning in financial services*. November, 2017. Dostupné na: <https://www.fsb.org/wp-content/uploads/P011117.pdf>

⁶⁹ BAESENS, Bart et al. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. In *Journal of the Operational Research Society*, Vol. 54. June, 2003. p. 627–635.

K aplikácii strojového učenia majú sklon hlavne fintech spoločnosti, ktoré „reprezentujú oblasť technologických inovácií aplikovaných v praxi vo vzťahu k poskytovaní rôznorodých finančných služieb.“⁷⁰ Medzinárodná banka pre zúčtovanie (BIS) publikovala štúdiu, v ktorej hodnotí praktickú úspešnosť využitia strojového učenia a neštandardných dát v kreditnom hodnotení. Gambacorta⁷¹ poukazuje hlavne na dva primárne rozdiely oproti tradičnému bankovníctvu. Prvým rozdielom je technická vybavenosť fintech spoločností umožňujúca spracovať veľké množstvo dát a pripravenosť spracúvať aj netradičné údaje (napr. zo sociálnych sietí). Druhým dôvodom je aktivácia potenciálu techník strojového učenia. Primárne poukazuje na význam nelineárnych a neparametrických predikčných modelov, ktoré môžu z dát vyťažiť dodatočné informácie, ktoré sú lineárnym modelom ako je logistická regresia nedostupné. Gambacorta vo svojej práci porovnáva tri modely. Prvý, tradičný lineárny prístup so štandardnými bankovými údajmi. Druhý model je lineárny model rozšírený o netradičné dáta a tretím je model strojového učenia s dátami z oboch zdrojov. Rozšírenie lineárneho modelu o netradičné dáta vylepšuje AUROC modelu o 2,2 %, pričom aplikácia machine learning-u v treťom modeli prekonáva počiatočný model o 5,3 %. Ďalej poukazuje aj na robustnosť pokročilého modelu voči exogénnym zmenám, ktoré zasiahli čínsky bankový sektor po zmene regulačných zákonov v roku 2017.

Khandani⁷² podobne ako Gambacorta deklaruje potrebu riešiť problém posudzovania kreditnej kvality nelineárnymi technikami. Opiera sa o výsledky empirického testu, pri ktorom pracoval so zápismi z úverového registra a transakčnými dátami. Vo výsledku preukázal, že takéto nové modely môžu prekonať tradičné lineárne prístupy o 6 až 23 % z pohľadu minimalizácie realizovaných strát, čím prináša aj finančný rozmer. K totožnému záveru dospel aj Berg⁷³, ktorý si rovnako myslí, že za úspechom fintech spoločnosti stojí agilná dátová architektúra a komplexné prediktívne modelovanie.

Chang⁷⁴ odporúča použiť pokročilé neparametrické modely, pričom aplikuje eXtreme gradient boosting (XGBoost) na pozorovaniach tchajwanských korporátnych úverov. Navrhnutý prístup dosahuje AUROC 0,94 a prekonáva iné lineárne prístupy ako logistickú

⁷⁰ Národná banka Slovenska [online] <https://www.nbs.sk/sk/dohlad-nad-financnym-trhom/fintech>

⁷¹ GAMBACORTA, Leonardo. How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. In *BIS Working Papers* No 834. December, 2019.

⁷² KHANDANI, Amir E. et al. Consumer credit-risk models via machinelearning algorithms. In *Journal of Banking & Finance*, vol. 34, issue 11. November 2010. p 2767–2787.

⁷³ BERG, Tobias. On the rise of fintechs – credit scoring using digital footprints. In *The Review of Financial Studies*. April, 2018.

⁷⁴ CHANG, Yung-Chia et al. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. In *Applied Soft Computing Journal*, vol. 73. October, 2018. p. 914-920.

regresiu s 0,78 alebo metódu pomocných vektorov (SVM) s 0,87 AUROC. V posledných rokoch sa dostávajú do popredia práve prístupy skupinového učenia (ensemble learning), ktoré predstavujú súbor spoločne sa rozhodujúcich individuálnych klasifikátorov. Tento prístup podporujú aj výsledky neskoršej štúdie Lessmanna⁷⁵, v ktorej porovnáva 41 metód strojového učenia a v hornej polovici sa umiestnili práve techniky skupinového učenia. Spomedzi testovaných individuálnych klasifikačných algoritmov preukázali najlepšie výsledky logistická regresia (16. miesto), lineárna diskriminačná analýza (20. miesto) a lineárny SVM (27. miesto).

Logistickej regresii sa podrobnejšie venoval Nikolic.⁷⁶ Konkrétne ju aplikoval na podnikateľské úvery v Srbsku a dosiahol kvalitný výsledok 0,835 AUROC. Munkhadalai⁷⁷ porovnáva vo svojej komplexnej štúdií viacero, v súčasnosti populárnych modelov, ako neurónové siete s AUROC 0,87 a XGBoost s 0,86. V zozname nájdeme aj logistickú regresiu, ktorá za týmito metódami významne nezaostáva a dosahuje 0,85 AUROC. Podobný obraz nachádzame aj v záveroch práce Finlaya⁷⁸. Logistická regresia zaostáva len o 1% za schopnosťami pokročilej neurónovej architektúry. Finlayove nastavenia experimentu sú spomedzi nami preštudovaných prác najviac podobné podmienkam praxe čo sa týka počtu pozorovaní a počtu atribútov.

Ako sme už naznačili, v literatúre často diskutovanou témou je veľkosť použitého súboru. Lessmann analyzuje 45 štúdií, ktoré boli vykonané v rozmedzí rokov 2003-2014. Zisťuje, že takmer polovica z odborných príspevkov nepracovala so vzorkou väčšou ako 1 000 pozorovaní, pričom sa často jedná o rovnaké zdroje dát (Austria credit approval⁷⁹ a German credit data⁸⁰). Okrem malého počtu klientov tieto súbory obsahujú aj malý počet nezávislých premenných, väčšinou do 20. Tieto skutočnosti predstavujú doposiaľ najväčšie limity výskumu vplyvu samoučiacich strojov na ohodnocovanie kreditnej kvality. Z nám známych prác, len Finlay⁸¹ a Kruppa⁸² pracujú so vzorkou nad 50 000 pozorovaní. V našej

⁷⁵ LESSMANN, Stefan et al. Benchmarking state-of-the art classification algorithms for credit scoring: An update of research. In *European Journal of Operational Research*, vol. 247, issue 1. May 2015. p. 124-136.

⁷⁶ NIKOLIC, Nebojsa. The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. In *Expert Systems with Applications*, vol. 40, issue 15. November, 2013. p. 5932-5944

⁷⁷ MUNKHDALAI, Lkhagvadorj et al. An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. In *Sustainability MDPI Journals*, vol. 11. January, 2019. p. 15.

⁷⁸ FINLAY, Steven. Multiple classifier architectures and their application to credit risk assessment. In *European Journal of Operational Research*, vol. 20, issue 2. April, 2011. p. 368-378.

⁷⁹ [https://archive.ics.uci.edu/ml/datasets/statlog+\(australian+credit+approval\)](https://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval))

⁸⁰ [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))

⁸¹ FINLAY, Steven. Multiple classifier architectures and their application to credit risk assessment. In *European Journal of Operational Research*, vol. 20, issue 2. April, 2011. p. 368-378.

práci vzorka číta viac ako 300 000 žiadateľov o úver so 120 atribútmi, čo sa oveľa viac podobá bankovej praxi.

Druhým bodom, ktorý predstavuje priestor na zlepšenie, je identifikácia ideálnej metriky, podľa ktorej by sa posudzovala úspešnosť modelu. V texte sme doteraz porovnávali práce podľa hodnoty AUROC, ktorá je bežnou praxou v štatistike na hodnotenie rozlišovacej schopnosti modelu binárnej klasifikácie. Medzi ďalšie postupy patrí vytvorenie tabuľky klasifikácie (confusion matrix) alebo výpočet F1 skóre. F1 skóre je váženým priemerom precision a recall. Recall (senzitivita modelu) vyjadruje dispozíciu modelu odhaľovať úvery, ktoré nebudú splatené. Precision definuje koľko zo zamietnutých žiadostí, by skutočne neboli splatené. Nízka hodnota precision by znamenala vysoké náklady obetovanej príležitosti. V štatistike je F1 používaná metrika, avšak neodráža určité bankové špecifiká. Snahou pri udeľovaní úverov je predvídať, ktorý úver nebude splatený, aby sa predišlo zaknihovaniu strát. Máme zato, že banková prax kladie väčší dôraz na recall ako na precision, a preto navrhujeme používať F2 skóre pre hodnotenie modelov na podporu úverového rozhodovania. Problematika posudzovania účinnosti modelu intenzívne rezonuje hlavne v práci Lesmanna, ktorý navrhuje, aby štandardom v kreditnom skórovaní bolo Brier score (BS score). Každý model predpovedá zaradenie klienta do určitej triedy s istou pravdepodobnosťou. Čím je hodnota pravdepodobnosti bližšie k hraničným hodnotám 0 alebo 1, tým istejší si model je svojim rozhodnutím. Pri oceňovaní úveru a stanovovaní úroku je dôležité, aby bola kreditná inštitúcia schopná pochopiť aj vnútorné riziko transakcie. Banka potrebuje model, ktorý okrem zatriedenia do správnej kategórie ohodnotí aj implicitnú rizikovosť klienta. BS numericky vyjadruje rozdiel medzi istotou modelu a skutočným stavom. Čím nižšia hodnota, tým presnejší model.

Ďalej existujú dva názorové tábory na úpravu vstupných dát do modelu strojového učenia. Z preštudovanej literatúry máme dojem, že preferovanejší je názor o potrebe redukcie počtu premenných. Tento názor obhajuje aj Belloni⁸³, ktorý prezentuje nutnosť preriediť nezávislé premenné (sparse-model development) a ponechať iba tie, ktoré majú zásadný vplyv na sledovanú premennú. Belloni vo svojej práci využíva Lasso techniku na elimináciu

⁸² KRUPPA, Jochen et al. Consumer credit risk: Individual probability estimates using machine learning. In *Expert Systems with Applications*, vol. 40, issue 13. October, 2013. p. 5125–5131.

⁸³ BELLONI, Alexandre – CHERNOZHUKOV, Victor – HANSEN, Christian B. Inference for high-dimensional sparse econometric models, In *Advances in Economics and Econometrics, 10th World Congress of the Econometric Society*. December, 2011.

prediktorov, ktorej úspešnosť otestujeme aj my. V priamom protiklade je Giannone⁸⁴, ktorý poukazuje na potrebu používať techniky „hustého“ modelovania (dense-model development). Vychádza z predpokladu, že všetky možné vysvetľujúce premenné sú dôležité pre predpoveď, hoci ich individuálny vplyv môže byť malý. Práve to považuje za podstatu strojového učenia, keď modelu poskytneme všetky dostupné dáta a ten individuálne vyvodí záver.

⁸⁴ GIANNONE, Domenico – LENZA, Michele – PRIMICERI, Giorgio E. Economic predictions with big data: The illusion of sparsity. In *Federal Reserve Bank of New York Staff Reports*, No. 847. April, 2018.

2 Cieľ práce

Cieľom našej práce je predstaviť modelovanie kreditného rizika pomocou vybraných metód a postupov strojového učenia. Na dosiahnutie tohto cieľa je nutné popri teoretickom pohľade priniesť aj perspektívu praktickej aplikácie na dátach. Experimentálne nastavenie zahŕňa reálne produkčné dáta od úverovej inštitúcie Home Credit. Podľa výsledku na týchto dátach bude možné urobiť kvalifikované rozhodnutie o tom, či prístup k modelovaniu kreditného rizika umelou inteligenciou má svoje uplatnenie v praxi.

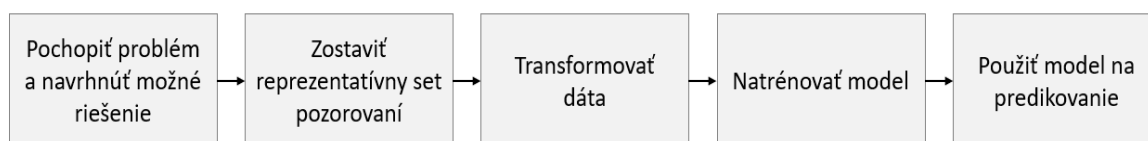
Hlavný cieľ sme doplnili o čiastkové ciele. Ich realizáciou adresujeme otázky, ktoré vystali z predchádzajúcich výskumov. Mnohé z nich sme načrtli v kapitole *1.7.2 Prehľad literatúry*. Hovoríme konkrétne o:

- Ciele nastaviť podmienky experimentu, ktoré budú čo najvernejšie reprezentovať kreditnú prax. V tomto bode kladieme dôraz na veľkosť súboru pozorovaní, ich pôvod a produkčný charakter.
- Ciele používať také hodnotiace kritéria, ktoré by boli výpovedné pre bankovú prax. Z pohľadu kreditnej inštitúcie nie je dôležitá celková presnosť modelu, ale jeho schopnosť identifikovať žiadateľov, ktorí by nedokázali/nechceli splatiť úver.
- Ciele posúdiť prínos samoučiacich systémov nielen štatistickými metrikami ale aj ich finančným rozmerom. Manažérsky pohľad bol často opomínaný, no práve on je kľúčový pri rozhodovaní v praxi.
- Ciele odpovedať na otázku či predikčné modelovanie s využitím väčšieho počtu aj individuálne menej významných atribútov (dense modelling) je schopné priniesť dodatočnú hodnotu v porovnaní s užším počtom významných prediktorov (sparse modelling).

3 Metodika a metódy skúmania

V záverečnej práci sa snažíme zodpovedať otázku, či metódy strojového učenia dokážu priniesť pridanú hodnotu do kreditného modelovania. Strojové učenie nám pomáha nájsť vzťahy a závislosti v dátach, ktoré sú kľúčové pri vytváraní predpovedí o výsledku nových pozorovaní. Géron⁸⁵ navrhuje týchto päť krokov pri aplikácii strojového učenia na riešenie problému.

Obrázok 4 Dizajn realizovaného systému hodnotenia kreditnej kvality



Zdroj: Spracované podľa Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition

Potom, čo máme zadefinovaný a pochopený problém – odhadovať kreditnú kvalitu žiadateľa o úver a aj zozbierané dáta – Home Credit (skr. HC) dataset, môžeme sa zamerať na tvorbu modelu. Každý model pracuje s určitými predpokladmi o podobe dát alebo ich rozdelení. Reálne dáta spravidla nespĺňajú dané predpoklady, a preto je nutná ešte ich transformácia.

Výber štatistického modelu považuje Gerón za kľúčový problém. V ďalšom texte priblížime štatistické a matematické metódy, ktoré sa využívajú na pozadí modelov strojového učenia aplikovaných v práci. Prvou je logistická regresia, ktorá zároveň predstavuje zlatý štandard kreditného skórovania⁸⁶ a druhou je metóda pomocných vektorov, ktorá je obľúbená najmä v akademickej obci.

Táto kapitola je inšpirovaná publikáciou Jurafského a Martina⁸⁷, z ktorej čerpáme poznatky zo štatistiky. Sekciu 3.1 *Logistická regresia* dopĺňame aj o výstupy Coss.⁸⁸ Informácie o konkrétnej transformácii daných konceptov do podoby strojového učenia

⁸⁵ GÉRON, Aurelien. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. 2st Edition, O'Reilly Media, Inc. p. 37. ISBN 978-1-492-03264-9.

⁸⁶ BOLTON, Christine. *Logistic regression and its application in credit scoring*. University of Pretoria, 2019. p. 27

⁸⁷ JURAFSKY, Daniel – MARTIN, James H.. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd Edition, Stanford University. p. 72-94. ISBN 978-0130950697

⁸⁸ COSS, Soňa. Modely s binárne závislou premennou.[elektronický zdroj] In *Ekonomika Informatika, vedecký časopis FHI EU v Bratislave a SSHI*. Ročník 2015, náklad 15, číslo 2. Dostupné na: <http://ei.fhi.sk/index.php/EAI/article/view/18>

získavame z diel Geróna⁸⁹ a VanderPlasa⁹⁰. Z praktických dôvodov budeme v ďalšom texte zásadné fakty citovať, iba ak budú pochádzať z iných zdrojov.

3.1 Logistická regresia

Ako sme uviedli v úvode do strojového učenia, dve najznámejšie a najrozšírenejšie metódy používané v rámci supervised learningu (učenia pod dohľadom) sú regresia (predikcia hodnôt) a klasifikácia (predikcia kategórie).

Základným a východiskovým modelom pre regresné úlohy je lineárna regresná analýza. Cieľom v lineárnom regresnom modeli je nájsť najvhodnejšiu nadrovinu, ktorá je schopná predikovať vplyv nezávislých premenných na hodnotu závislej premennej. Táto téza sa spája s predpokladom, že sledovaná premenná je spojitá.⁹¹ V kontexte nastavenia nášho experimentu sa snažíme zaradiť žiadateľa o úver do jednej z dvoch kategórií – splatí alebo nesplatí úver. Modelujeme vplyv vysvetľujúcich premenných (parametrov žiadosti o úver) na kategoriálnu závislú premennú, ktorá nenadobúda spojité ale diskkrétne obmeny znaku (splatí, resp. nesplatí úver). Takýto typ problému, v ktorom modelujeme kategoriálnu závislú premennú spadá do okruhu klasifikačných úloh.

Niektoré regresné modely môžu byť po menších úpravách použité na klasifikáciu. Príkladom takejto úpravy je logistická regresia. Názov v tomto prípade môže byť klamlivý, keďže za účelom klasifikácie potrebujeme zostaviť lineárny model zo vstupných premenných a pomocou linkovej funkcie prevádzame jeho výsledky na kategoriálnu závislú premennú. Logistickú regresiu môžeme považovať za osobitný prípad lineárnej regresie, keď je výsledná premenná kategorizovaná, pričom ako závislú premennú používame log pomeru pravdepodobnosti. V Rovnici 3.1 je premenná z výsledok lineárneho modelu, ktorý môže nadobúdať spojité hodnoty od $-\infty$ po ∞ .

$$z = \beta_0 + \beta_1 * x_1 + \dots + \beta_n * x_n \quad (3.1)$$

Výsledok lineárneho modelu z je pomocou linkovej funkcie *logit* prevedený na p , teda pravdepodobnosť. Rovnica 3.2 zachytáva tento prevod. V Rovnici 3.3 je p vyjadrené samostatne na ľavej strane rovnici.

⁸⁹ GÉRON, Aurelien. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. 2st Edition, O'Reilly Media, Inc. p. 113-176. ISBN 978-1-492-03264-9.

⁹⁰ VANDERPLAS, Jake. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc. p. 390-420 ISBN 978-1491912058.

⁹¹ Pre spojitú premennú platí, že od hodnoty k hodnote sa mení spojitou, teda počet jej možných hodnôt je veľmi veľký, často závislý od presnosti merania.

$$z = \ln \left(\frac{p}{1-p} \right) \quad (3.2)$$

Pravdepodobnosť p pochádza z intervalu 0 až 1 a rozhoduje a pridelení pozorovania do jednej z kategórií. Ak p presiahne vopred stanovanú hranicu (v nemodifikovanom tvare 0,5) je pozorovanie zaradené do pozitívnej kategórie.

$$p = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \quad (3.3)$$

Logistická regresia sa najčastejšie aplikuje v prípadoch dvojkategoriálnej (binárnej) vysvetľovanej premennej, napr. dobrý/zlý úver. Tento model si však poradí aj viackategoriálnou závislou premennou. V takom prípade ide o multinomickú logistickú regresiu, ktorá je nad rámec našej práce. Tento úvod zo štatistiky doplníme v ďalšom texte o špecifiká príznačné pre strojové učenie.

Algoritmus strojového učenia využívajúci logistickú regresiu, potrebuje súbor pozorovaní M pozostávajúci z tzv. vstup-výstup párov $x^{(i)}, y^{(i)}$, kde horné indexy v zátvorkách označujú jednotlivé prípady (pozorovania). X je množina všetkých charakteristík (atribútov, resp. nezávislých premenných) jednotlivých prípadov a Y je množina výsledných tried daných pozorovaní. Systém strojového učenia má potom 4 komponenty:

1. Vstup. Pre každé vstupné pozorovanie $x^{(i)}$ to bude vektor prvkov $[x_1; x_2; \dots; x_n]$, hovoríme aj o množine charakteristík vstupu, príp. atribútoch. Zvyčajne budeme odkazovať na charakteristiku i pre vstup $x^{(i)}$ označenú ako $x^{(i)}_i$, niekedy zjednodušene ako x_i .
2. Klasifikačná (linková) funkcia, ktorá vypočíta odhadovanú triedu $\hat{y}$. V jednoduchей LR sa za týmto účelom používa sigmoid funkcia. Pri multinomickej LR sa pristupuje k použitiu softmax funkcie.
3. Stratová funkcia učenia (loss function), ktorá sa zameriava na minimalizáciu chýb pri učení z pozorovaní z množiny M . Pri logistickej regresii nepoužívame rovnakú stratovú funkciu ako pri lineárnej regresii. Lineárna regresná analýza využíva strednú kvadratickú chybu odhadu (mean squared error). Namiesto nej používame v LR krížovú entropiu (cross-entropy).⁹²

⁹² NIELSEN, Michael. [elektronický zdroj] *Neural Network and Deep Learning*. Dostupné na: <http://neuralnetworksanddeeplearning.com/chap3.html>

4. Optimalizačný algoritmus stratovej funkcie. Štandardne sa používa nejaký variant gradientu klesania (GD).⁹³ V práci konkrétne využívame SAGA algoritmus.

3.1.1 Vstupy do modelu

Majme jedno vstupné pozorovanie $x^{(i)}$, ktoré reprezentuje vektor charakteristík $[x_1; x_2; \dots; x_n]$. Výstup $\hat{y}$ klasifikačného algoritmu môže byť 1, teda zaradenie do pozitívnej triedy alebo 0, ktorá odpovedá negatívnej triede. V kontexte kreditného modelovania je výsledok $P(y = 1|x)$ indikátorom, že klient úver nesplatí a $P(y = 0|x)$, že klient úver splatí.

Na základe pozorovaní z tréningového datasetu sa odhaduje vektor váh $[w_1, \dots, w_n]$ a konštantný člen b , pre ktoré nadobúda stratová funkcia najmenšiu hodnotu. Tréningový dataset tvoria usporiadané dvojice $x^{(i)}$ a $y^{(i)}$, vďaka čomu sa dá odhadnúť aké charakteristiky sú kľúčové pre pozorovania z rovnakej triedy. Parametre sa odhadujú v priebehu minimalizácie stratovej funkcie. Proces odhadovania parametrov modelu sa v žargóne nazýva ako „učenie“ alebo „trénovanie“. Každá váha w_i je reálne číslo a prináleží jednému zo vstupných atribútov x_i . Čím vyššiu absolútnu hodnotu w_i nadobudne, tým významnejší je tento atribút. Ak je w_i kladná, znamená to, že atribút je pozitívne závislý k triede. Naopak, záporná hodnota w_i napovedá o negatívnej závislosti k triede. V úlohe kreditného rizika by sme teda mohli očakávať, že charakteristika o minulom zlyhaní klienta bude mať vysokú pozitívnu váhu a veličina príjem klienta bude mať zápornú váhu. Konštantný člen b (bias term), tiež nazývaný priesečník (intercept) alebo lokujúca konštanta, je ďalšie reálne číslo, ktoré sa pripočíta k váženým vstupom. Model klasifikácie najskôr vynásobí každý atribút x_i jeho príslušnou váhou w_i . Následne sčíta vážené atribúty a pridá člen b . Výsledné číslo z vyjadruje vážený súčet vstupných parametrov. Tento postup vyjadruje *Rovnica 3.4*.

$$z = \left(\sum_{i=1}^n w_i * x_i \right) + b \quad (3.4)$$

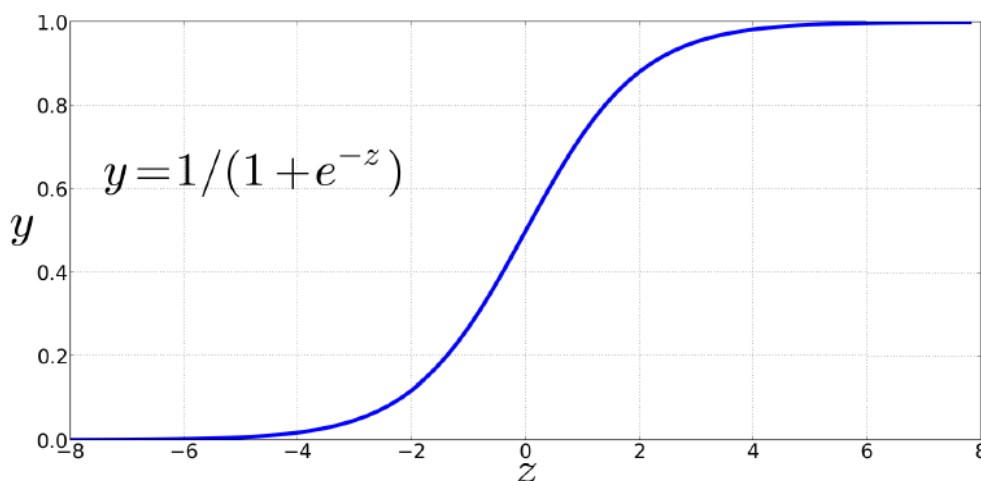
3.1.2 Linková funkcia sigmoid

V doterajšej časti sme uviedli, že výsledok logistickej regresie leží v uzatvorenom intervale $[0,1]$, čo odráža hodnotu pravdepodobnosti. Avšak vzťah, ktorým sme definovali pre potenciálne hodnoty premennej z , nespĺňa túto podmienku. Keďže váhy pochádzajú z množiny všetkých reálnych čísel R , hodnota z pochádza z intervalu $-\infty$ až ∞ . Tu aplikujeme

⁹³ Pri použití GD (gradient descent) je vhodné, aby všetky atribúty mali rovnakú škálu (škálované vstupy), inak algoritmus bude bežať veľmi dlho kým nájde optimum.

linkovú funkciu, pomocou ktorej prevádzame výsledky lineárneho modelu na kategoriálnu závislú premennú.

Obrázok 5 Priebeh sigmoid funkcie



Zdroj: Prevzaté z Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition

V tomto prípade je vhodnou linkovou funkciou, funkcia sigmoid so zápisom $\sigma(z)$. Sigmoid funkcia má priebeh podobný písmenu „S“, čo dokumentuje *Obrázok 5*. Nazýva sa tiež logistická funkcia (logistic function), od čoho je odvodený názov logistickej regresie. *Rovnica 3.5* vyjadruje vzťah na prevod z . S týmto vzťahom sme sa už stretli pri *Rovnici 3.3*.

$$y = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (3.5)$$

Sigmoid funkcia mapuje akékoľvek reálne číslo z k hodnote z intervalu $[0,1]$, teda pravdepodobnosti. Použitie tejto funkcie implikuje niekoľko dodatočných pozitív. Ostrý sklon funkcie pre hraničné pravdepodobnosti (blízke 0 a 1) pomáha minimalizovať negatívne účinky extrémnych hodnôt na model. Ďalšou výhodou je, že funkcia je spojitá a diferencovateľná po celej svojej dĺžke, čo umožňuje využiť GD pri tréňovaní modelu.⁹⁴

Model štandarde (*Rovnica 3.6*) predikuje $\hat{y}$ pri použití hraničnej pravdepodobnosti 0,5. Túto hranicu (threshold) je možné modifikovať, k čomu sa často pristupuje pri nebalancovaných údajoch. Nebalancovanými označujeme dáta, v ktorých jedna trieda prevažuje. Napríklad vo vzorke poskytnutých úverov je prevaha splatených úverov.

⁹⁴ Diferencovateľná funkcia zabezpečuje, aby bolo možné pre odhadované parametre dopočítavať gradienty určujúce smer, ktorým je potrebné sa uberať pri minimalizovaní stratovej funkcie (loss function) v priebehu tréňovania.

$$\hat{y} = \begin{cases} 1, & \text{ak } P(y = 1|x) > 0,5 \\ 0, & \text{ak } P(y = 0|x) \leq 0,5 \end{cases} \quad (3.6)$$

Proces tréovania, ktorý vedie k určeniu parametrov W a b si vyžaduje 2 komponenty. Prvým je stratová funkcia a druhým je jej optimalizačný algoritmus. Stratová funkcia zisťuje vzdialenosť medzi výstupom modelu a skutočným výsledkom. Optimalizačný algoritmus zabezpečuje iteratívnu aktualizáciu váh. V ďalšom texte predstavíme stratovú funkciu, ktorá sa bežne používa na riešenie logistickej regresie, stratu krížovej entropie.

3.1.3 Strata krížovej entropie

Hľadáme takú stratovú funkciu, ktorá pre každé pozorovanie $x^{(i)}$ vyjadří ako blízko je predpoveď modelu $\hat{y}$ od pravdivého výsledku $y^{(i)}$. Slovné tento vzťah vyjadruje *Rovnica 3.7*. Okrem pojmu stratová funkcia ($L(\hat{y}, y)$ - loss function) sa v literatúre často stretávame aj s označením nákladová funkcia (cost function). Odtiaľ tento dvojtvár $Cost(\hat{y}, y)$ alebo $L(\hat{y}, y)$.

$$Cost(\hat{y}, y) = L(\hat{y}, y) = 0 \text{ keď sa } \hat{y} \text{ líši od skutočného } y \quad (3.7)$$

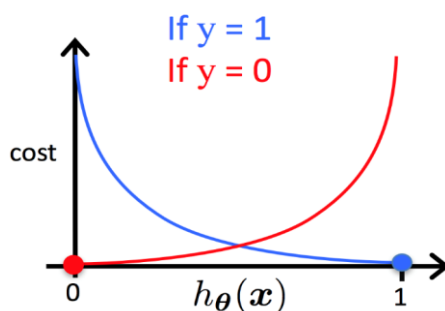
Stratová funkcia nespôsobuje pre správne predikcie tried žiadne náklady, naopak u nesprávnych označení tried vedie k veľkým prírastkom na hodnote. Cieľom fázy „učenia“ je minimalizovať stratovú funkciu. Opísaný prístup sa nazýva metóda maximálnej vierohodnosti odhadu (conditional maximum likelihood estimation). Vyberáme také parametre W a b , ktoré maximalizujú pravdepodobnosť správneho pridelenia do triedy $y^{(i)}$ pre pozorovanie $x^{(i)}$. Maximalizovať pravdepodobnosť je to isté ako maximalizovať logaritmus pravdepodobnosti a to isté ako minimalizovať záporný logaritmus pravdepodobnosti (negative-log-likelihood), čo odpovedá minimalizácii nákladovej funkcie. Pre klasifikačné úlohy sa tento prístup volá krížová entropia (cross-entropy).

$$Cost(p(y|x), y) = \begin{cases} -\log(p(y|x)), & \text{ak } y = 1 \\ -\log(1 - p(y|x)), & \text{ak } y = 0 \end{cases} \quad (3.8)$$

- $p(y|x)$ je odhadnutá pravdepodobnosť modelom, označujeme aj $\hat{p}$

Táto funkcia dáva zmysel, pretože vzťah $-\log(\hat{p})$ zabezpečuje rastúcu penalizáciu ak sa pre pozorovanie $y^{(i)} = 1$ odhadnutá pravdepodobnosť $\hat{p}$ blíži k 0. V opačnom prípade, ak model odhaduje pravdepodobnosť blížiacu sa 1, bude penalizácia veľmi nízka. Podobne postupujeme pre pozorovanie $y^{(i)} = 0$. Ak sa predikované $\hat{p}$ blíži k 1, tak vzťah $-\log(1 - \hat{p})$ bude nadobúdať rastúce kladné hodnoty. Pre $\hat{p}$ blížiacu sa k 0 bude $-\log(1 - \hat{p})$ nadobúdať malé hodnoty. Na vysvetlenie si pomôžeme graficky *Obrázkom 6*.

Obrázok 6 Stratová funkcia pri logistickej regresii



Zdroj: Prevzaté z www.datascience.stackexchange.com/questions/40982/logistic-regression-cost-function/40983

Stratová funkcia pre celý tréningový set dát je priemerom nákladov skrz všetky pozorovania. Môže byť zapísaná vo forme jednej rovnice.

$$L_{CE}(W, b) = J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(\hat{p}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{p}^{(i)})] \quad (3.9)$$

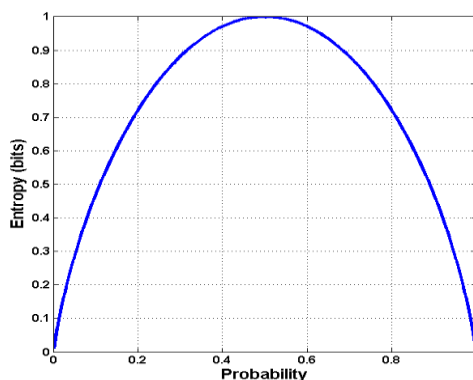
- pre $y^{(i)} = 0$ vypadáva prvý člen v zátvorke
- pre $y^{(i)} = 1$ vypadáva druhý člen v zátvorke
- $J(\theta)$ je konvenčné označenie stratovej funkcie, v ktorom θ znamená súhrny vektor parametrov W a b , teda $\theta = [b, w_1, w_2, \dots, w_n]$.

Rovnica 3.9 definuje našu stratovú funkciu a zároveň je po úprave aj rovnicou krížovej entropie (Rovnica 3.10) medzi skutočným rozdelením pravdepodobnosti y a odhadnutým rozdelením $\hat{y}$.

$$H(y, \hat{y}) = y \log \hat{y} + (1 - y) \log(1 - \hat{y}) \quad (3.10)$$

Ak si je model veľmi istý, že predikovaná hodnota patrí do danej skupiny, miera neistoty bude veľmi malá, čo dokazuje aj priložený graf na Obrázku 7.

Obrázok 7 Križová entropia



Zdroj: Prevzaté z www.matlabdatamining.blogspot.com/2006/11/introduction-to-entropy.html

Proces učenia sa končí keď optimalizačný algoritmus nájde minimum stratovej funkcie.⁹⁵ Užitočnou vlastnosťou stratovej funkcie krížovej entropie je jej konvexnosť (neplatí pre Lasso regularizáciu). Konvexnosť znamená, že úsečka spájajúca ľubovoľné dva body na krivke nepretína funkciu. To znamená, že funkcia má len jedno minimum a to globálne. Táto vlastnosť pomáha predísť nežiadúcemu stavu, kedy sa proces optimalizácie (hľadania minima) ukončí predčasne v lokálnom minime, ktoré nie je globálnym minimom funkcie.⁹⁶ Pri hľadaní minima používame gradient klesania (GD), resp. niektorý jeho variant.

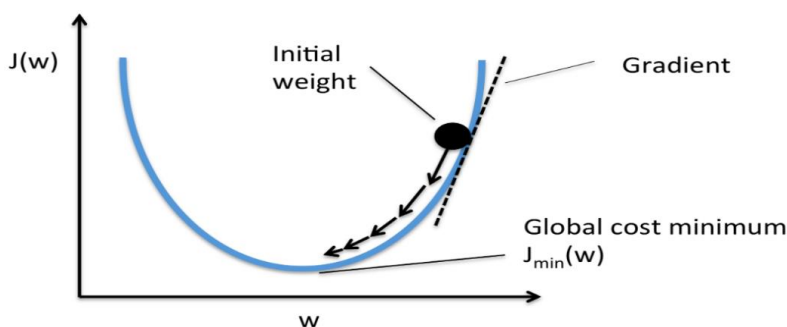
3.1.4 Optimalizácia stratovej funkcie cez gradient klesania

Optimalizácia modelu logistickej regresie nie je tak priamočiara ako pri lineárnej regresnej analýze. Pre LR neexistuje uzatvorená forma rovnice (matematická rovnica, ktorá priamo vráti výsledok – vektor θ , minimalizujúci stratovú funkciu). Preto je nutné pristúpiť k iteratívnemu hľadaniu minima.

Gradient klesania je populárny algoritmus, ktorý je schopný dospieť k najvhodnejšiemu riešeniu veľkého počtu optimalizačných problémov. Myšlienkou GD je iteratívne (postupne) vylepšovať parametre s cieľom minimalizovať stratovú funkciu.

GD je metóda, ktorá speje k minimu funkcie tým, že hľadá sklon funkcie ako to zobrazuje *Obrázok 8*. GD identifikuje smer pohybu (v priestore parametrov funkcie W a b), v ktorom sklon funkcie stúpa najstrmšie a parametre modifikuje v opačnom smere. Keď gradient nadobudne hodnotu nula, dotyčnica k stratovej funkcii (prvá derivácia) je rovnobežná s osou x . V tomto momente GD konverguje v minime tejto funkcie.

Obrázok 8 Hľadanie minima gradientom klesania



Zdroj: Prevzaté z www.rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization/

⁹⁵ Nájdenie globálneho minima a konvergencia v ňom je želaný stav, niekedy sa však proces učenia ukončí tým, že sa vyčerpá maximálny počet iterácií. O význame iterácií budeme hovoriť v ďalšej časti.

⁹⁶ Na porovnanie, stratová funkcia pre viacvrstvové neurónové siete nie je konvexná, čo často vedie k tomu, že sa gradient klesania zasekne v lokálnom minime a nedosiahne globálne minimum funkcie. V tom prípade model nedosiahne najoptimálnejšie nastavenie parametrov. To však nemusí nutne znamenať, že model je chybný.

Odhadovanie parametrov začína inicializáciou s nenulovými náhodnými hodnotami pre W a b (random initialization). Parametre sú postupne upravované (zväčšované, zmenšované) príslušne k tomu, aby sa optimalizačný algoritmus pohyboval po stratovej funkcii smerom nadol. Parameter w_i sa teda zmenší alebo zväčší v závislosti od toho, aby táto zmena viedla k poklesu hodnoty $J(\theta)$. Postupná úprava parametrov sa vykonáva v iteráciách, t.j. krokoch, ktoré sú na *Obrázku 8* znázornené ako čierne plné šípky. GD algoritmus hľadá gradient stratovej funkcie vždy pre aktuálne nastavenie parametrov, v ktorom sa nachádza.

Čitateľa odkazujeme na Jurafského a Martina⁹⁷, kde sa dozvie viac o matematickom odvodení GD a na Geróna⁹⁸, kde sa dočíta viac o rôznych typoch GD.

V tejto práci sme aplikovali optimalizačný algoritmus SAGA.⁹⁹ SAGA v jednotlivých iteráciách počíta gradienty iba pre jedno náhodne vybrané pozorovanie z množiny M , čím urýchľuje proces hľadania minima. Aj napriek tejto malej vzorke dosahuje stabilný priebeh konvergenzie k minimu, pretože začleňuje do pamäte predchádzajúce hodnoty gradientov.

3.1.5 Regularizácia modelu

V procese tréningovania je model vystavený iba určitému súboru dát, ktoré tvoria tréningový set. Výkonnosť modelu na týchto dátach však nie je relevantná pre prax. Model musí v praxi robiť správne rozhodnutia o pozorovaniach, ktoré predtým nevidel a sú pre neho nové. Proces odhadovania parametrov je úspešný, ak sa mu podarí dosiahnuť stav, v ktorom je model schopný generalizovať a nie učiť sa „naspamäť“ špecifiká pozorovaní danej vzorky. Nedostatočne zovšeobecnenie má tri hlavné príčiny:

1. Skreslenie (Bias¹⁰⁰) – táto chyba pramení z nesprávnych predpokladov, napríklad ak predpokladáme, že dáta sú lineárne ale v skutočnosti sú kvadratické. Model je skreslený svojimi predpokladmi a nedokáže dostatočne zohľadniť reálny stav. Skreslenie je výsledkom nedostatočného prispôsobenia sa tréningovým dátam – model nedostatočne identifikoval špecifiká súboru.

⁹⁷ JURAFSKY, Daniel – MARTIN, James H.. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd Edition, Stanford University. p. 82-87. ISBN 978-0130950697

⁹⁸ GÉRON, Aurelien. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. 2st Edition, O'Reilly Media, Inc. p. 119-130. ISBN 978-1-492-03264-9.

⁹⁹ SCHMIDT, Mark – LE ROUX, Nicolas – BACH, Francis. [elektronický zdroj] Minimizing Finite Sums with the Stochastic Average Gradient. In *Mathematical Programming B*, Springer, vol. 162. p.83-112. Dostupné na: <https://hal.inria.fr/hal-00860051/document>

¹⁰⁰ Pojem skreslenie (anglické slovo bias) v tomto kontexte nesúvisí s konštantným členom v lineárnom modeli (po anglicky bias term)

2. Rozptyl (variance) – tento problém naopak pramení z prílišnej senzitivity modelu k malým a nedôležitým variáciám v tréningovom sete dát. Hovoríme, že sa model extrémne prispôbil dátam, na ktorých prebiehalo učenie a nedokáže tak účinne generalizovať.
3. Neredukovateľná chyba – dáta, v ktorých je hluk (obsahujú chyby), neodrážajú dostatočne skutočný stav. Učenie na chybných dátach následne vedie k nesprávnym záverom z tréningu. Model teda nie je schopný presnej predikcie.

Zvyšovanie komplexnosti modelu vedie k zníženiu skreslenia, naopak spôsobuje väčší rozptyl. Tento vzťah platí aj naopak, zjednodušovanie modelu umožňuje znížiť variabilitu za cenu zvýšenia skreslenia. V práci sme aplikovali tri známe regularizačné techniky, ktoré majú prispieť k dostatočnej generalizácii. Pre lineárne modely sa regularizácia spravidla dosahuje obmedzením váh W modelu.

Prvým typom je Ridge regresia¹⁰¹ (známa ako L_2 regularizácia), ktorá je regularizáciou pridávajúcou člen $\lambda \sum_{i=1}^n \theta_i^2$ k stratovej funkcii. Konkrétne, l_2 norm¹⁰² je sumácia váh umocnených na druhú. Toto obmedzenie núti model nielen vhodne sa prispôbiť dátam, na ktorých prebieha tréning, ale aj držať váhy W čo najmenšie. Regularizačný člen je pridávaný k stratovej funkcii len počas tréningu. Keď sa model dotrénuje, jeho výkonnosť je posudzovaná neregularizovaným meraním.¹⁰³ Mieru regularizácie ovplyvňuje veľkosť hyperparametra λ .¹⁰⁴ Ak je $\lambda=0$ nedochádza k žiadnej regularizácii a váhy ostávajú na pôvodnej úrovni. Ak je naopak λ veľmi veľká, hodnoty W , teda váhy, sa blížia k nule a výsledkom je rovná čiara, ktorá ide cez priemer dát. Rovnica 3.11 dokumentuje upravenú stratovú funkciu pre Ridge regresiu.

$$J(\theta) = L_{CE}(\theta) + \lambda \sum_{i=1}^n \theta_i^2 \quad (3.11)$$

¹⁰¹ Pri Ridge je nevyhnutné, aby všetky atribúty boli z rovnakej škály (škálované vstupy), inak by niektoré atribúty boli bezdôvodne neúmerne penalizované.

¹⁰² Norm je označenie pre meranie vzdialenosti medzi dvoma vektormi, v tomto prípade medzi vektorom predikcií a vektorom očakávaných výsledkov.

¹⁰³ Bežným postupom je, že sa nákladová funkcia použitá pri tréningu líši od kritéria, ktorým meriame výkonnosť. V práci sme model trénovali na stratovej funkcii krížovej entropie a výkonnosť testovali cez iné metriky, o ktorých píšeme v záverečnej kapitole.

¹⁰⁴ Lambda (λ) má svoju implementáciu v *Pythone* ako hyperparameter C , ktorý je definovaný vzťahom $C=1/\lambda$. Odkiaľ C , je prevrátená hodnota lambdy. Znižovaním hodnoty hyperparametra C sa zosilňuje regularizácia.

Jediný parameter z θ , ktorý nie je ovplyvnený regularizáciou je priesečník (konštantný člen $-b$), keďže sumácia začína od $i=1$.

Druhým variantom regularizácie je Lasso Regresia (prípadne L1 regularizácia). Podobne ako Ridge aj pri Lasso pridávame k stratovej funkcii penalizačný člen, tentokrát $\lambda \sum_{i=1}^n |\theta_i|$. Tento vzťah sa nazýva aj l_1 norm a je sumáciou váh bez umocnenia. Upravená nákladová funkcia pre Lasso má podobu Rovnice 3.12.

$$J(\theta) = L_{CE}(\theta) + \lambda \sum_{i=1}^n |\theta_i| \quad (3.12)$$

Lasso je schopné eliminovať váhy najmenej významných atribútov. Inými slovami, Lasso vykonáva výber atribútov (feature selection) a vracia preriedený model (sparse model). Eliminácia atribútov prebieha tým, že im model prisúdi nulové váhy. Lasso bola pre nás prirodzená alternatíva, keďže skúmaný dataset obsahoval veľa stĺpcov (atribútov).

Tretia možnosť ako regularizovať lineárny model je elastická sieť (Elastic Net), ktorá agreguje vlastnosti Ridge aj Lasso regresie. Táto charakteristika sa odráža v hyperparametri r , ktorý definuje pomer kombinácie Ridge a Lasso. Ak je r blízke nule prevažuje Ridge, a naopak, ak je r blízke 1, majoritne sa prejavuje Lasso. Funkčný vzťah vieme vyjadriť Rovnicou 3.13.

$$J(\theta) = L_{CE}(\theta) + r * \lambda \sum_{i=1}^n |\theta_i| + (1 - r) * \lambda \sum_{i=1}^n \theta_i^2 \quad (3.13)$$

Regularizačný parameter λ a pomerový parameter r sú po miere učenia η (ovplyvňuje gradient klesania) ďalšími hyperparametrami nášho modelu. Mieru učenia nastavíme cez SAGA optimalizačný algoritmus. Hodnoty λ a r však musíme definovať. Netrúfame si nastaviť tieto hodnoty arbitrárne, preto použijeme techniku krížovej validácie. Tejto technike sa podrobne venujme v kapitole 4.3.1 *Nastavenie hyperparametrov modelu*. V krátkosti, najprv preddefinujeme interval, povedzme $\lambda=1$, $\lambda=10$, $\lambda=100$, na základe ktorého sa natrénujú (tri) rôzne modely. Následne sa porovnáva úspešnosť predikcií týchto modelov na nových pozorovaniach a za optimálne nastavenie sa označí model s najlepším výsledkom.

3.2 Support Vector Machine

Technika support vector machine (SVM) alebo v slovenskom preklade metóda pomocných vektorov je bežne používanou technikou strojového učenia. Jej počiatky siahajú do 60. rokov 20. stor. Mnohé z nápadov, ktoré sa teraz vyvíjajú v rámci SVM prvýkrát navrhli V. N. Vapnik a A. Y. Chervonenkis.¹⁰⁵ Najsilnejšou stránkou mechanizmu je jeho všestrannosť. Okrem toho, že technika SVM dokáže riešiť klasifikačné aj regresné úlohy, vieme ju nastaviť do podoby lineárneho aj nelineárneho modelu. SVM sa transformuje do nelineárnej formy pomocou tzv. triku jadra (kernel trick), pri ktorom implicitne mapuje vstupné dáta do viacrozmerného priestoru. SVM spadá do kategórie učenia pod dohľadom (supervised learning), preto ako nutná súčasť vstupu do modelu pri tréovaní je aj hodnota závislej premennej. Algoritmus support vector clustering zase predstavuje kategóriu učenia bez dohľadu a používa techniku pomocných vektorov vyvinutých v algoritme SVM na kategorizáciu neoznačených údajov a je jedným z najpoužívanějších zoskupovacích (clustering) algoritmov.¹⁰⁶ Miera flexibility SVM je unikátna a definuje jej špecifické postavenie v oblasti strojového učenia. Častejšie sa však s aplikáciou SVM stretávame pri akademických prácach ako v praxi. Dôvodom je výpočtová náročnosť SVM, ktorá naráža na svoje limity pri veľmi veľkých súboroch dát.¹⁰⁷ Povedzme, pri súboroch, ktoré sú podobne rozsiahle ako naše.

Klasifikácia pozorovaní je bežnou úlohou strojového učenia. Predpokladajme, že niektoré pozorovania patria do jednej z dvoch tried, napr. splatený a nesplatený úver, a cieľom je rozhodnúť, do ktorej kategórie zaradiť nové pozorovanie. V prípade SVM sa pozorovanie chápe a zaznačuje ako p -rozmerný vektor (zoznam p čísel). Hodnoty p -rozmerného vektora sú tvorené hodnotami jednotlivých atribútov pozorovania. Pozorovania sú rozmiestnené v p -rozmernom priestore a ich súradnice odpovedajú hodnotám ich atribútov. Úlohou SVM je oddeliť tieto dve triedy pomocou $(p-1)$ -dimenzionálneho rozhrania – nadroviny (hyperplane). Opísaná metóda je lineárnou technikou. Existuje veľa rozhraní, ktoré môžu rozdeľovať pozorovania do príslušiacich tried. Obrázok 9 poukazuje na dilemu výberu ideálneho rozhrania. Známené sú tri veľmi odlišné oddelovacie rozhrania, ktoré medzi

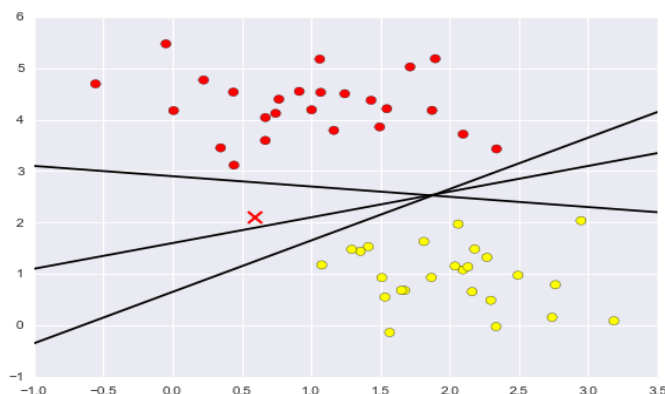
¹⁰⁵ SCHÖLKOPF, Bernhard – LUO, Zhiyuan – VOVK, Vladimir. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Springer-Verlag Berlin Heidelberg, 2013. p. 4. ISBN 978-3-642-41136-6.

¹⁰⁶ VAPNIK, Vladimir N. et al. Support vector clustering. In *Journal of Machine Learning Research*, vol. 2. 2001. p. 125–137.

¹⁰⁷ YU, Hwanjo – YANG, Jiong – HAN, Jiawei [elektornický zdroj] *Classifying Large Data Sets Using SVMs with Hierarchical Clusters*. University of Illinois. Dostupné online: http://hanj.cs.illinois.edu/pdf/kdd03_scalesvm.pdf

zakreslenými skupinami dokonale rozlišujú. Problém je, že v závislosti od toho, ktoré rozhranie zvolíme, bude nové pozorovanie (bod označený červeným „X“) zaradené do jednej alebo druhej triedy. Tým dokumentujeme význam výberu správneho oddeľovača – rozhrania.

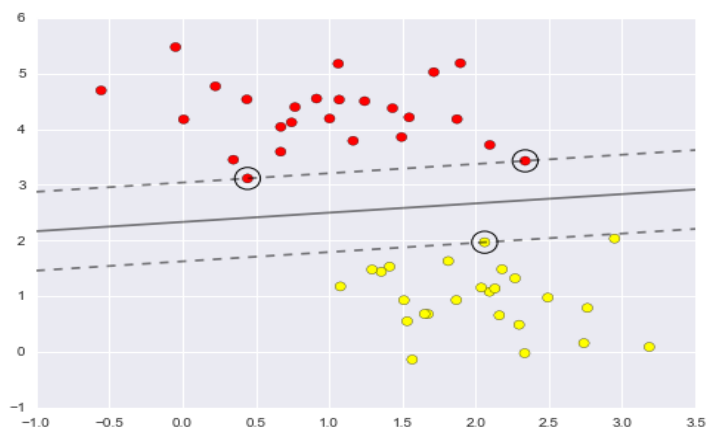
Obrázok 9 Rôzne rozhrania, ktorými je možné klasifikovať pozorovania do tried



Zdroj: Prevzaté z Python Data Science Handbook: Essential Tools for Working with Data

Najvhodnejšie rozhranie podľa princípov SVM je také, ktoré predstavuje najväčšie oddelenie (rozpätie) medzi týmito dvoma triedami. Preto vyberieme rozhranie tak, aby bola dodržaná maximálna vzdialenosť (margin) od tohto rozhrania k najbližšiemu pozorovaniu z jednej aj druhej kategórie, teda oboch strán. Pozorovania, ktoré sa nachádzajú najbližšie k rozhraniu z jednej či druhej strany sa nazývajú pomocné vektory (support vectors). Odtiaľ pochádza názov – metóda pomocných vektorov. Ak takéto rozhranie existuje, je známe ako rozhranie s maximálnym rozpätím (maximum-margin hyperplane). Pomocné vektory sú na *Obrázku 10* zaznačené v čiernych krúžkoch.

Obrázok 10 Vytvorenie rozhrania pomocou pomocných vektorov

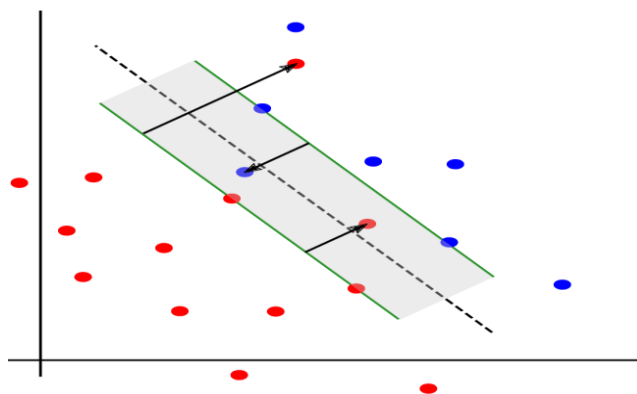


Zdroj: Prevzaté z Python Data Science Handbook: Essential Tools for Working with Data

Kľúčom k úspechu tohto klasifikačného algoritmu je fakt, že z hľadiska vhodnosti záleží iba na polohe pomocných vektorov. Body ďalej od okraja, ktoré sú na správnej strane, neovplyvňujú výsledok tréningu modelu. Technicky vzaté, tieto body neprispievajú k funkcii straty použitej pri tréningu modelu, takže na ich polohe a počte nezáleží, pokiaľ nepresiahnu definované rozpätie k deliacemu rozhraniu (margin).

V praxi sa nestretávame iba s dátovými súbormi, u ktorých vieme jednoducho a ľahko rozhodnúť o rozdelení do tried neporušenou hranicou. Spravidla je v dátach nejaký prienik, čo znemožňuje jednoduché rozdelenia. Pozorovania z jednej kategórie sa charakteristicky podobajú na pozorovania z druhej kategórie a naopak.

Obrázok 11 Vytvorenie rozhrania s porušením hraníc



Zdroj: Prevzaté z <https://www.semanticscholar.org/paper/Training-Support-Vector-Machines-using-Coresets-Baykal-Liebenwein/>

SVM preto implementuje faktor, ktorý zjemňuje túto hranicu. Tento faktor umožňuje niektorým bodom v tréningovej fáze priblížiť sa k rozhraniu, alebo ho aj prekročiť, ak to zabezpečí lepšie prispôsobenie. Neprístupnosť do rozpätia medzi pomocným vektorom a rozhraním sa riadi hyperparametrom, najčastejšie známym ako C . Pri veľmi veľkých C je táto zóna málo priepustná a body v nej nemôžu ležať (hard margin). Pri menších C je narušenie zóny prípustné (soft margin) tak, ako to ilustruje Obrázok 11.

3.2.1 Optimalizačný problém SVM

Určovanie rozhrania matematicky prebieha nasledovne.¹⁰⁸ Nech x je pozorovanie vybrané z dát. Zároveň je x vektor s hodnotami jednotlivých atribútov definujúci toto pozorovanie. Rozhranie je vyjadrené Rovnicou 3.14.

¹⁰⁸ GÉRON, Aurelien. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. 2st Edition, O'Reilly Media, Inc. p. 166-176. ISBN 978-1-492-03264-9.

$$w^T * x + b = 0 \quad (3.14)$$

Lineárny klasifikačný model SVM predpovedá triedu novej inštalácie x jednoduchým výpočtom rozhodovacej funkcie, prezentovaným v Rovnici 3.15.

$$w^T * x + b = w_1 * x_1 + \dots + w_n * x_n + b \quad (3.15)$$

Ak je výsledok kladný, predpovedaná trieda $\hat{y}$ je pozitívna trieda (trieda 1), v opačnom prípade je to negatívna trieda (trieda 0). Tento vzťah pri soft margin vyjadríme Rovnicou 3.16.

$$\hat{y} = \begin{cases} 0, & \text{ak } w^T * x + b \leq -margin \\ 1, & \text{ak } w^T * x + b \geq +margin \end{cases} \quad (3.16)$$

A pri hard margin Rovnicou 3.17.

$$\hat{y} = \begin{cases} 0, & \text{ak } w^T * x + b < 0 \\ 1, & \text{ak } w^T * x + b \geq 0 \end{cases} \quad (3.17)$$

Vo vzťahu je w vektor váh atribútov $[w_1, \dots, w_n]$ a premenná b je konštantný člen. Rozpätie (margin) je vyjadrené nasledovne.

$$\frac{|w^T * x_i + b|}{\|w\|} \quad (3.18)$$

$\|w\|$ je označenie Euklidovej metriky, ktoré prevádza vektor na jeho dĺžku v Euklidovom priestore. V jednoduchosti povedané, tento parameter vyjadruje sklon funkcie. Práve $\|w\|$ je optimalizačný problém, ktorý SVM rieši pri tréningu. Čím je $\|w\|$ menšie, tým väčšie je rozpätie medzi pomocným vektorom a deliacim rozhraním. Avšak vzťah, ktorý je predmetom optimalizácie je $\frac{1}{2} w^T * w$, kde w je vektor váh. Tento výraz je rovný $\frac{1}{2} \|w\|^2$, pričom vo svojej podstate ide o $\|w\|$, ktorý však optimalizovať sám o sebe nemôžeme. $\|w\|$ nie je diferencovateľné v nule. Optimalizačný algoritmus pracuje oveľa lepšie na diferencovateľných funkciách, a preto pristupujeme k takejto úprave.

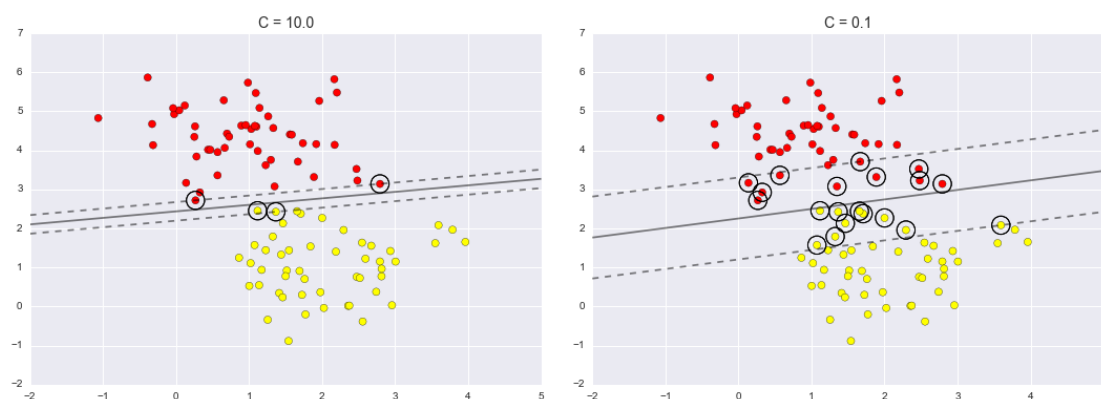
$$\min_{w,b} \left| \frac{1}{2} w^T * w \right| \quad (3.19)$$

Pri optimalizácii jemného rozpätia (soft margin) musíme pre každé pozorovanie zaviesť novú premennú ζ , ktorá zmierňuje zákaz narušania rozpätia. ζ musí byť vždy ≥ 0 . $\zeta^{(i)}$ pre premennú x_i definuje do akej miery môže toto pozorovanie narušiť rozpätie.

$$\min_{w,b,\zeta} \left| \frac{1}{2} w^T * w + C \sum_{i=1}^m \zeta^{(i)} \right| \quad (3.20)$$

Teraz máme pri optimalizácii dva protichodné ciele. V rovnakom čase sa snažíme dosiahnuť hodnotu ζ čo najnižšiu, aby sa znížil počet porušení cieleného rozpätia a zároveň minimalizovať $\|w\|$, aby bolo rozpätie maximálne. Kľúčovým v tomto bode je C , ktoré umožňuje definovať kompromis medzi týmito dvoma cieľmi. Vytvára sa nový optimalizačný problém vyjadrený v Rovnici 3.20. Ilustrujeme ho aj na Obrázku 12.

Obrázok 12 Povolené narušenie rozpätia medzi pomocnými vektormi a rozhraním



Zdroj: Prevzaté z Python Data Science Handbook: Essential Tools for Working with Data

Matematické riešenie problému optimalizácie týchto dvoch veličín je nad rámec tejto práce. Všeobecný záver z tejto časti by mohol byť nasledovný. Trénovanie lineárneho klasifikátora SVM znamená nájsť hodnoty W a b , ktoré maximalizujú vzdialenosť pomocných vektorov od rozhrania, pričom sa treba vyhnúť narušeniu rozpätia (hard margin) alebo tieto neželané narušenia minimalizovať (soft margin).

4 Tvorba predikčného modelu pomocou metód strojového učenia

V tejto kapitole popíšeme proces úpravy vstupných dát a samotnú tvorbu predikčného modelu s využitím techník strojového učenia. Zostrojených bude niekoľko modelov, a preto sa v závere kapitoly budeme venovať podrobnému vyhodnoteniu výkonu jednotlivých posudzovaných modelov.

Klientske dáta o úverových dlžníkoch majú produkčnú povahu a boli sprístupnené verejnosti spoločnosťou Home Credit Group (skr. HC). Úprava dát a stavba modelu bola realizovaná pomocou programovacieho jazyka *Python*. Na tvorbu modelu sme použili voľne dostupné nástroje knižnice *scikit-learn*.¹⁰⁹

Výsledný kód, natrénované modely a bližší popis dát je voľne dostupný na odkaze: https://github.com/MarkoVateha/Credit_Risk_Home_Credit

4.1 Základný rámec postupu pri úprave a tvorbe modelu

Celková architektúra nášho navrhovaného systému hodnotenia úverovej kvality, ktorú zachytáva *Obrázok 13* pozostáva z troch fáz.

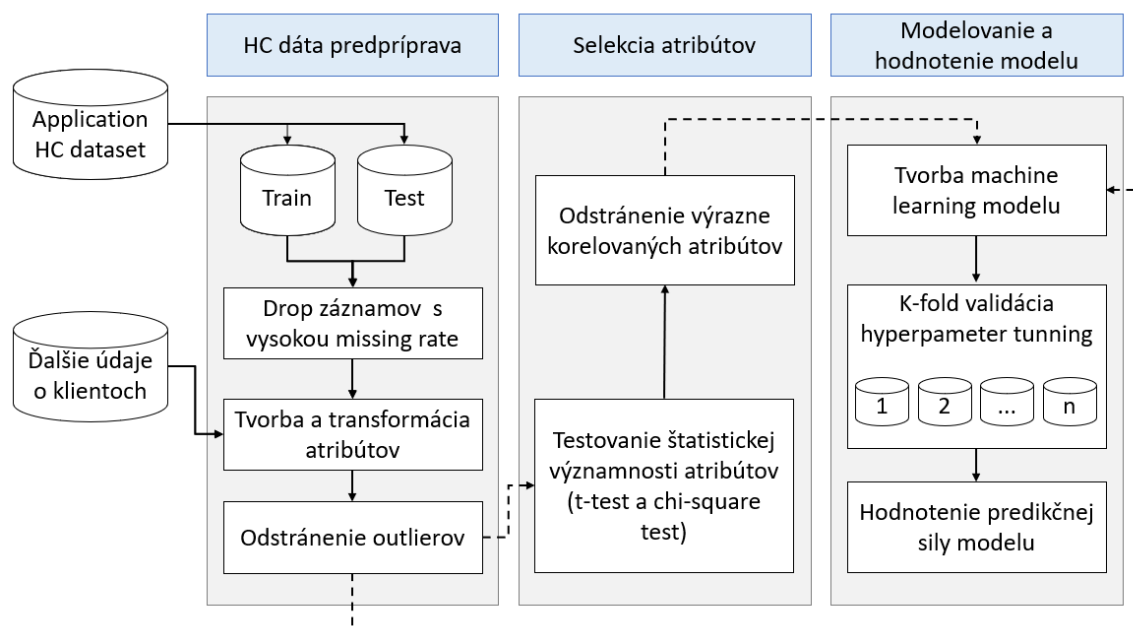
V prvej fáze spracúvame získané dáta. Pozornosť sme konkrétne venovali chýbajúcim záznamom, extrémnym hodnotám a extrakcii dodatočných informácií o žiadateľoch z historických databáz. Rovnako sme sa snažili v tejto etape prispieť aj našou expertízou z danej problematiky a vylepšiť predikčnú silu modelov vytváraním nových atribútov (feature engineering). Úplne prvým bodom je rozdelenie súboru na tréningovú a testovaciu vzorku. Na tréningovom datasete prebieha odhadovanie parametrov (učenie modelu). Model s týmto nastavením sa testuje pri klasifikácii nových pozorovaní z testovacej vzorky.

V druhej fáze používame algoritmus výberu štatisticky významných atribútov. Konkrétne ide o funkciu filtrovania a výberu atribútov pozostávajúcu z t-testu a testu chí-kvadrát. Navyše pristupujeme aj ku eliminácii silne korelovaných atribútov. Cieľom týchto úprav je vybrať najlepšie, najúčinnnejšie a najmenej redundantné premenné.

V záverečnej tretej fáze sa takto predpripravené dáta použijú na trénovanie algoritmov strojového učenia, pričom následne posudzujeme ich dosiahnutú predikčnú silu. Sledujeme aj to, či selekcia štatisticky významných atribútov skutočne pomôže k zlepšeniu predikčnej sily modelu, alebo len zníži celkový čas trénovania modelu.

¹⁰⁹ <https://scikit-learn.org/stable/>

Obrázok 13 Dizajn realizovaného systému hodnotenia kreditnej kvality



Zdroj: Vlastné spracovanie

4.1.1 Vstupné dáta

Dáta, ktoré využívame v diplomovej práci pochádzajú z platformy [kaggle.com](https://www.kaggle.com/c/home-credit-default-risk/overview)¹¹⁰, na ktorej ich zverejnila spoločnosť Home Credit Group. Klientske údaje boli anonymizované. Jedná sa o produkčné dáta, ktoré HC žiadnym spôsobom neupravoval. Produkčná povaha dát zvýšila náročnosť manipulácie s nimi, no zároveň pridala výsledkom práce na relevantnosti. Komplexnosť poskytnutých záznamov je popísaná nižšie. Klientske dáta boli zverejnené v niekoľkých zložkách:

- Application.csv – hlavný register, ktorý obsahoval informácie o klientoch v čase podania žiadosti a informáciu o konečnom stave úveru – splatený alebo zlyhaný.
- Bureau.csv – dáta z centrálného registra úverov poskytnutých fyzickým osobám rôznymi kreditnými inštitúciami.
- Bureau_balance.csv – behaviorálne dáta o vývoji mesačných splátok pri úveroch zaznamenaných v Credit Bureau.
- Previous_application.csv – história úverových žiadostí v spoločnosti HC. Pri každej žiadosti je aj záznam o prijatom rozhodnutí a v prípade schválenia kreditného produktu aj o výslednom stave – splatený alebo zlyhaný.

¹¹⁰ <https://www.kaggle.com/c/home-credit-default-risk/overview>

- POS_CAS_balance.csv – behaviorálne dáta na mesačnej báze o zostatku istiny poskytnutých úverov zo zdrojov HC.
- Credit_card_balance.csv – behaviorálne dáta o zostatku na kreditných kartách klientov Home Credit.
- Instalments_paments.csv – behaviorálne dáta o priebehu splátok klientov HC ku emitovaným kreditným produktom.

Pre účely modelovania bol východiskovým zdrojom informácií primárne hlavný register žiadateľov o úver – Application.csv. V ďalšom texte k nemu referujeme pod označením „klientske dáta“. Historický prehľad poskytnutých úverov HC nám pomohol doplniť obraz o žiadateľovi aj o jeho minulé kreditné správanie. Podobnú informáciu o žiadateľoch aj z ostatných inštitúcií poskytoval register spravovaný Credit Bureau. Z dôvodu nízkej kvality dát a náročnosti úpravy sme tento zdroj nevyužili.

Podrobnejší pohľad na databázu klientov poskytuje *Tabuľka 2*. Daným žiadateľom bol poskytnutý úver¹¹¹ zo zdrojov HC a výsledný stav bol doplnený ako sledovaná premenná. Každá žiadosť má 120 atribútov (vysvetľujúcich premenných), ktoré možno zaradiť do štyroch skupín. Prvou skupinou sú finančné informácie, ako napríklad mesačný príjem či výška úveru. Druhú kategóriu tvoria demografické údaje, ako napríklad pohlavie, počet detí či vek. Treťou skupinou sú externé ratingy klienta, ktorých pôvod nepoznáme. Celkovo HC pracuje s tromi externými ratingami. Do štvrtej kategórie radíme ostatné údaje, označujeme ich aj ako netradičné dáta. Netradičné zdroje dát sú príznačné pre modely strojového učenia. Patrí sem napríklad vek auta, rozloha bytu či hodina a deň kedy bola žiadosť o úver vyplnená. Úspešne splatených úverov vo vzorke je 282 686. U zvyšných 24 825 nastal problém so splácaním. Miera zlyhaných úverov je približne 8,8%.

Tabuľka 2 Zhrnutie údajov o klientskej databáze Application.csv

Inštitúcia	Počet pozorovaní	Nezávislé premenné	Splatené úvery	Nesplatené úvery	Miera zlyhania
Home Credit	307 511	120	282 686	24 825	0.088

Zdroj: Vlastné spracovanie

Počet pozorovaní, ktoré vstúpili do tvorby modelu je o niečo menší než pôvodný záznam. Nie však o veľa. Niektoré pozorovania boli odstránené pre chyby alebo extrémne hodnoty. V ďalšom texte približujeme ako sme postupovali pri ich úprave.

¹¹¹ Súbor neobsahuje len úvery ale aj kreditné karty. Od tohto faktu abstrahujeme a oba tieto kreditné produkty označujeme súhrnným názvom „úver“.

4.2 Predpríprava dát

V prípravnej fáze pred tvorbou modelu je nutné pristúpiť k úprave vstupných dát do podoby, ktorá podporí jeho správne fungovanie. Zle upravené vstupy vedú k nesprávnym výstupom z modelu. V ďalšom texte kapitoly priblížime náš prístup ku transformácii dát.

4.2.1 Izolovanie testovacej vzorky

Pred samotnou transformáciou musíme pozorovania rozdeliť na tréningovú a testovaciu sadu čo zachytáva *Tabuľka 3*. Tréningový dataset obsahuje 80 % pozorovaní a je teda dostatočne robustný, aby obsiahol všetky klientske špecifiká. Objem testovacích dát ostáva stále relatívne veľký na posúdenie úspešnosti modelu bez skreslenia.

Tabuľka 3 Rozdelenie súboru na tréningovú a testovaciu vzorku

Dáta	Splatené úvery	Nesplatené úvery	Celkový počet klientov	Miera zlyhania úverov
Tréning	220 808	19 565	240 373	8.86 %
Testovanie	55 162	4 898	60 060	8.88 %

Zdroj: Vlastné spracovanie

4.2.2 Chýbajúce hodnoty

Najväčším problémom pri transformácii dát sú spravidla chýbajúce informácie. Práve preto začíname úpravu klientskych dát analýzou tohto problému. Sledovali sme relatívnu početnosť chýbajúcich záznamov pre jednotlivé atribúty a výsledky sme zaznamenali do *Tabuľky 4*. Analýza odhalila závažne nedostatky u 48 atribútov z celkového počtu 120.

Tabuľka 4 Prehľad o závažnosti chýbajúcich dát v registri žiadateľov

Percento chýbajúcich hodnôt	0 %	< 1 %	1 – 20 %	20 – 45 %	45 – 70 %	> 70 %
Počet atribútov	54	10	7	2	48	0

Zdroj: Vlastné spracovanie

Problémovú skupinu tvoria primárne atribúty vzťahujúce sa na nehnuteľnosť, ktorú žiadateľ obýva. Medzi takéto atribúty patrí napr. veľkosť obytnej a neobytnej plochy, počet výťahov v budove, počet vstupov do budovy a pod.

Menované atribúty sú zdanlivo málo výpovedné, no v snahe zachovať povahu strojového učenia, sme sa rozhodli ponechať ich. Predpokladáme, že model má možnosť vyťažiť z týchto premenných dodatočnú informáciu. K zvládnutiu práce s atribútmi rôznej

kvality má napomôcť regularizácia. Tá zabezpečí, že najmenej výpovedné atribúty budú minimalizované v procese tréningovania modelu. Usudzujeme, že model sa pri tréningovaní dokáže vysporiadať s vysokým počtom atribútov a sám odhadne optimálnu váhu pre tieto vstupy. Zároveň budeme hodnotiť modely tréňované na dvoch rôzne upravených vstupoch. Jeden dataset si ponechá všetky vytvorené premenné. Druhý, bude obsahovať iba štatisticky významné a nekorelované atribúty. Predpokladáme, že zámerne ponechané atribúty sa neocitnú v druhom variante úpravy. V závere práce zistíme, ktorý prístup je správnejší.

Pri atribútoch s chybovosťou do 1 % postupujeme nasledovne. Pozorovania, ktoré pre tieto charakteristiky nenadobúdajú hodnoty vymažeme. Týmto krokom nestratíme veľa pozorovaní a ani neskreslíme výpovednú hodnotu atribútu. Prázdne záznamy pre atribúty s vysokou chybovosťou, teda nad 1% musíme v záujme zachovania vzorky upraviť inak (v opačnom prípade by sme stratili 75% súboru). Prvým krokom je vytvorenie nových zrkadlových, nula jednotkových atribútov. Tieto zrkadlové atribúty nadobúdajú hodnotu 1 pre pozorovania, u ktorých chýbal záznam v pôvodnom atribúte. Následne dochádza u pôvodných atribútov k dopĺňaniu prázdnych záznamov. Dopĺňanie bolo prevedené dvoma rôznymi spôsobmi, pričom sme rozlišovali či ide o atribúty s veľkým alebo malým vplyvom na sledovanú premennú. Vplyv na sledovanú premennú sa určoval výpočtom korelačného koeficientu Kendall tau-b medzi atribútom a vysvetľovanou premennou. Následne sme vybrali šesticu atribútov s najvyššou hodnotou koeficientu. Kendall tau-b pre ostatné parametre už bol z nášho pohľadu veľmi malý.

Prázdne záznamy v kategórii menej významných sme dopĺňali metódou mediánu. Vypočítali sme medián z tréningového datasetu a použili ju na doplnenie chýbajúcich hodnôt v tréningovej aj v testovacej sade.¹¹² Ak by sa takýto systém aplikoval do praxe, použili by sme nameraný medián pre všetkých nových žiadateľov s nevyplnením záznamom.

¹¹² Medián sa vypočítava výlučne len na tréningovej vzorke. Z testovacieho datasetu nezískavame žiadne informácie, keďže tieto pozorovania nám simulujú budúcich žiadateľov o úver.

Tabuľka 5 Prehľad atribútov, ktoré boli dopočítavané metódami strojového učenia

Názov premennej	Korelácia s predikovanou veličinou	Počet chýbajúcich záznamov	Percentuálne vyjadrenie chybovosti
Vek auta	0.037	202 929	65.99 %
Externý rating 1	- 0.155	173 378	56.38 %
Rozloha bytu	- 0.044	153 020	49.76 %
Externý rating 3	- 0.179	60 965	19.83 %
Zamestnaný v dňoch	0.075	54 852	18.03 %
Zamestnaný ako percento dĺžky života	- 0.068	54 852	18.03 %

Zdroj: Vlastné spracovanie

V kategórii významných prediktorov sme hodnoty chýbajúcich záznamov dopočítavali použitím modelu strojového učenia postaveného na princípoch lineárnej regresie. Zoznam premenných definovaných ako významné prezentuje *Tabuľka 5*. Cieľom je doplniť tieto parametre čo možno najpresnejšími potenciálnymi hodnotami, pretože arbitrárna hodnota medián nezohľadňuje vnútornú heterogenitu klientov. Vstupmi, z ktorých sa model učil boli ostatné klientske údaje. Model odhaduje lineárny vzťah medzi vstupmi a zvoleným atribútom ako závislou veličinou. Podobne ako pri metóde medián je potrebné odložiť si tieto natrénované modely a dopočítavať nimi prázdne hodnoty parametrov v testovacej sade. V praxi by sa týmito modelmi dopočítavali chýbajúce záznamy u nových žiadateľov.

4.2.3 Odstránenie odľahlých hodnôt

Identifikácia odľahlých hodnôt (outliers) je kľúčová pri odhaľovaní chýb alebo šumu vo veľmi obsiahlych súboroch dát, pri ktorých nie je v ľudských silách overiť si relevantnosť nameraných hodnôt každej jednej premennej. Odstránenie pozorovaní, ktoré v sebe nesú nepravdepodobné, extrémne hodnoty má priniesť analýze konzistentnosť. V snahe predísť skresleniu parametrov žiadostí pristupujeme k ich detekcii a následnému odstráneniu.

Existuje viacero štatistických metód na odhaľovanie odľahlých hodnôt. Vzhľadom na špecifiká nášho súboru dát uprednostňujeme výpočet Z-skóre. Z-skóre vyjadruje počet štandardných odchýlok, o ktoré je hodnota konkrétneho pozorovania nad strednou hodnotou javu, ktorý pozorujeme. Úlohou Z-skóre je opísať ľubovoľný bod cez jeho vzťah k smerodajnej odchýlke a priemeru skupiny bodov, do ktorej patrí. Štandardne sú kritickou hranicou tri štandardné odchýlky nad alebo pod priemerom. V snahe minimalizovať počet

odstránených záznamov definujeme našu kritickú hodnotu na štyri štandardné odchýlky. Aplikovaním tejto techniky odstraňujeme 3 655 pozorovaní.

4.2.4 Feature engineering

Zaujímavou možnosťou ako vylepšiť schopnosť modelu rozlišovať medzi žiadateľmi o úver je vytvoriť dodatočné atribúty z informácií, ktorými aktuálne disponujeme. Cieľom je vytvoriť premenné, ktoré pomôžu odhaliť vzťah medzi nezávislou a závislou premennou.

Používanou metódou je vytváranie polynomických atribútov prenásobovaním charakteristík medzi sebou. Tým vieme v rámci lineárnej regresii modelovať aj nelinearitu medzi premennými a sledovanou veličinou. Pre zachovanie jednoduchosti a ekonomickej interpretovateľnosti veličín nevyužívame túto možnosť. Namiesto toho pristupuje k tvorbe nových nezávislých premenných pomocou expertízy (tzv. domain knowledge feature engineering). *Tabuľka 6* prináša prehľad novovytvorených premenných.

Tabuľka 6 Prehľad novovytvorených atribútov a ich vzťah k sledovanej premennej

Nová premenná	Korelácia s predikovanou veličinou
Pomer medzi príjmom žiadateľa a výškou úveru	- 0.002
Pomer medzi príjmom žiadateľa a anuitou	- 0.004
Pomer medzi príjmom a počtom členov v domácnosti	- 0.006
Pomer medzi výškou úveru a pravidelnou anuitou	- 0.032
Anomálie v informácii o počte odpracovaných dní	- 0.046
Odpracované obdobie ako percento prežitého života	- 0.068

Zdroj: Vlastné spracovanie

Významný vzťah k sledovanej premennej majú dva novovytvorené atribúty. Jedným z nich je atribút o anomáliách v počte odpracovaných dní. Niektorí žiadatelia o úver totiž uviedli, že majú odpracovaných viac dní ako je ich vek. Druhý novovytvorený atribút predstavuje pomer medzi počtom odpracovaných dní a vekom žiadateľa. Vychádzali sme z predpokladu, že je rozdiel ak 10 rokov odpracoval 30 alebo 50 ročný človek. Táto skutočnosť môže mať vplyv na skryté finančné rezervy žiadateľa či jeho pracovnú morálku. Záporná korelácia len potvrdzuje tento úsudok. Čím väčšiu časť svojho života človek odpracoval, tým skôr sa vyhne problémom so splácaním. Ostatné premenné vzišli ako

pomerové veličiny medzi finančnými údajmi, napr. pomer medzi príjmom a výškou úveru a pod.

4.2.5 Encoding

V našom súbore dát sa okrem numerických atribútov nachádzajú aj atribúty kategorické. Tieto premenné nadobúdajú diskkrétne hodnoty. Radíme sem napr. pohlavie, zamestnanie a pod.

Model strojového učenia, žiaľ, nedokáže spracovávať kategorické veličiny (s výnimkou Light Gradient Boosting Machine – LightGBM). Preto musíme nájsť spôsob ako tieto premenné zakódovať (reprezentovať) do numerickej podoby predtým, než ich použijeme ako vstup do modelu. Ponúkajú sa dve ekvivalentné možnosti:

- **Label encoding:** kódovanie štítkov priradí každej jedinečnej diskkrétnej hodnote v kategorickej premennej celé číslo. Nevytvárajú sa žiadne nové stĺpce (atribúty).

Obrázok 14 Príklad label encoding-u

	zamestnanie		zamestnanie
0	bankovníctvo	0	1
1	administratíva	1	2
2	bankovníctvo	2	1
3	šport	3	4
4	priemysel	4	0

Zdroj: Vlastné spracovanie

- **One-hot encoding:** nula jednotkové kódovanie vytvorí nový stĺpec pre každú jedinečnú diskkrétne hodnotu premennej. Všetkým pozorovaniam bude priradená 1 v stĺpci zodpovedajúcej kategórie a 0 v ostatných stĺpcoch.

Obrázok 15 Príklad one-hot encoding-u

	zamestnanie		bankovníctvo	administratíva	šport	priemysel
0	bankovníctvo	0	1	0	0	0
1	administratíva	1	0	1	0	0
2	bankovníctvo	2	1	0	0	0
3	šport	3	0	0	1	0
4	priemysel	4	0	0	0	1

Zdroj: Vlastné spracovanie

Rôznym modelom vyhovujú odlišné techniky zakódovania kategorických premenných. Vzhľadom na modely, ktoré využívame v práci je výhodne pristúpiť k nula jednotkovému kódovaniu. Aby sme sa vyhli nula jednotkovej pasci (dummy variable trap) musíme pre každú premennú vynechať jednu diskretnú hodnotu. Inými slovami, ak máme premennú, ktorá nadobúda dve hodnoty, napr. pohlavie (muž alebo žena). Na zakódovanie takejto premennej potrebujeme len jeden stĺpec. Pre premennú nadobúdajúcu n hodnôt potrebujeme $n-1$ nových stĺpcov. Lineárne modely majú ako súčasť svojho odhadu aj priesečník (lokujúcu konštantu). Bez takéhoto opatrenia by mohlo dôjsť k nepresnostiam modelu.

4.2.6 Logaritmovanie numerických premenných

Špeciálnu pozornosť si vyžadujú aj niektoré kvantitatívne atribúty, ktoré nadobúdajú numerické hodnoty. Podľa Paraliča¹¹³ tieto atribúty majú v oblasti strojového učenia špecifické postavenie, pretože „*tento typ atribútov je najčastejší a mnohé algoritmy predpokladajú práve takýto typ dát.*“

Takmer všetky modely sú postavené na predpoklade, že kvantitatívne pozorovania pochádzajú z normálneho rozdelenia. Nanešťastie, v praxi sa stretávame len s malým počtom procesov, ktoré generujú hodnoty s normálnym rozdelením. Pri vizuálnej kontrole našich dát zisťujeme, že sú charakteristické rozdelením s ťažkým chvostom (heavy-tailed distribution). Ak by sme dáta neupravili, model by neopodstatnene preferoval jeden výsledok pred iným. Štandardom pri riešení tohto problému je logaritmovanie vstupov, ktoré sme využili aj my.

4.2.7 Škálovanie vstupov

Jednou z najdôležitejších transformácií, ktorú musíme aplikovať na skúmané údaje je škálovanie vstupov. Až na pár výnimiek, algoritmy strojového učenia nefungujú správne, ak majú numerické atribúty vstupu veľmi odlišné mierky. Naše atribúty nadobúdajú hodnoty z veľmi odlišných rozpätí. Logaritmovanie čiastočne, hoci nepriamo, pomohlo znížiť tieto rozdiely. Ak by model pri trénovaní pracoval s neškálovanými dátami, neodôvodnene by penalizoval parametre, ktoré nadobúdajú väčšie hodnoty. A navyše by konvergoval k minimu veľmi dlho. Stav sme vyriešili aplikáciou metódy štandardizovaného škálovania vstupov.

¹¹³ PARALIČ, Ján [elektronický zdroj] *Pochopenie dát*. Technická univerzita v Košiciach. Dostupné na: <http://people.tuke.sk/jan.paralic/prezentacie/OZ/Pochopenie-dat.pdf>

Štandardizácia je celkom priamočiara metóda, pri ktorej sa od pozorovania najprv odpočíta priemerná hodnota (čiže štandardizované hodnoty majú vždy nulovú strednú hodnotu) a potom sa vydolí rozptylom tak, aby výsledné rozdelenie malo št. odchýlku 1. Výsledkom je normalizované normálne rozdelenie.

4.2.8 Informácie o minulom správaní sa klienta

Nad rámec úprav, ktoré boli spomínané, sme pristúpili aj k preberaniu dodatočných informácií o žiadateľoch. Prioritne sa jedná o údaje o kreditnej minulosti žiadateľa. Niektorí posudzovaní žiadatelia mali v minulosti kontakt s HC alebo zápis v centrálnom registri úverov. Získavame nové atribúty, ktoré nesú informácie napríklad o tom, či klient bol v kontakte s HC v minulosti, či bola jeho žiadosť schválená a či kredit dokázal splatiť. *Tabuľka 7* poskytuje prehľad základných charakteristík pôvodných a transformovaných dát.

Tabuľka 7 Hlavné charakteristiky datasetov použitých v práci

Názov premennej	Počet pozorovaní	Počet atribútov*
Pôvodný dataset	307 511	121
Upravený dataset	300 433	341
Upravený dataset po selekcii atribútov	300 433	159

* počet atribútov zahŕňa aj sledovanú premennú

Zdroj: Vlastné spracovanie

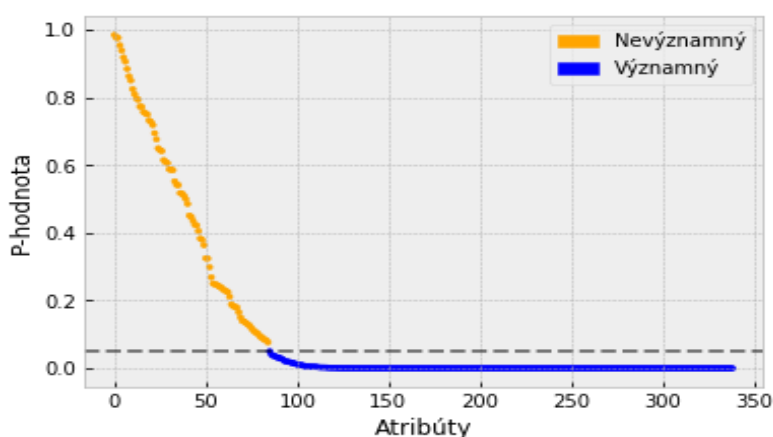
4.2.9 Výber štatisticky významných atribútov

Ako bolo v práci niekoľkokrát deklarované, trénovať modely budeme na dvoch rôzne upravených vstupných dátach. Druhá verzia úpravy si vyžaduje výber štatisticky významných a nekorelovaných charakteristík.

Vo fáze selekcie atribútov sme identifikovali štatisticky významné premenné na základe t-testu a testu chí-kvadrát. T-test bol prevedený pre numerické atribúty rozdelené do dvoch vzoriek, ktoré reprezentovali splatené a nesplatené úvery. T-testom musel byť preukázaný signifikantný rozdiel v nameraných hodnotách. Napríklad novovytvorený atribút pomeru príjmu k výške úveru nesúvisí s úverovou bonitou klienta, pretože neexistuje štatisticky významný rozdiel v hodnote atribútu medzi splatenými a zlyhanými úvermi na hladine významnosti 0,05 (p-hodnota = 0,084).

V prípade kategorických premenných sme sa pomocou testu chí-kvadrát opierali o frekvenciu výskytu splatených a nesplatených úverov v rámci jedného atribútu. Napríklad informácia o tom, či žiadateľ žije v regióne, v ktorom pracuje nerozlišuje medzi dlžníkmi (p -hodnota = 0,235). Obrázok 16 ukazuje výsledok testov hypotéz zľava doprava, pri ktorých sa zvyšuje hladina významnosti a znižuje sa p -hodnota. Zachovaných bolo 254 premenných (85 zahodených), ktoré sú štatisticky významné na odlíšenie splatených a nesplatených úverov. Tieto premenné sú reprezentované modrou farbou a ostatné nevýznamné premenné sú znázornené oranžovými bodmi.

Obrázok 16 Výsledok testovania hypotéz o významnosti atribútov



Zdroj: Vlastné spracovanie

4.2.10 Problém multikolinearity

Multikolinearita je stav, kedy medzi dvomi alebo viacerými nezávislými premennými nameriame vysokú závislosť. To znamená, že tieto premenné sú navzájom korelované. Ak prediktorové premenné v rovnakom regresnom modeli korelujú, nemôžu nezávisle predpovedať hodnotu závislej premennej. Inými slovami, vysvetľujú do veľkej miery rovnaké odchýlky v závislej premennej, čo znižuje ich samostatnú štatistickú významnosť. Kolinearita nemá až taký vplyv na chybovosť v predikciách ale prevažne na neadekvátne hodnoty koeficientov (váh) jednotlivých charakteristík. Tým spôsobuje problém interpretability modelu. „Model s korelovanými premennými často vedie k presnejšiemu odhadom ako model po odstránení kolinearity. Preto je prípustné použiť jeden model na predikcie a druhý na zvýšenie výpovednej hodnoty.“¹¹⁴

¹¹⁴ <https://medium.com/future-vision/collinearity-what-it-means-why-its-bad-and-how-does-it-affect-other-models-94e1db984168>

Aplikujeme zjednodušený koncept riešenia. Vytvorili sme korelačnú maticu a sledovali vzťahy medzi premennými, ktoré sú korelované nad 0,7.¹¹⁵ Keďže v našom modeli používame aj umelé nula jednotkové vysvetľujúce premenné a iné binárne premenné, pristupujeme k meraniu vzájomnosti cez Kendall tau-b koeficient. V našom upravenom datasete je niekoľko desiatkov atribútov, ktoré sú vzájomne silno korelované. Je viacero techník ako riešiť problém kolinearít. V práci volíme dva varianty. Jeden z nich je ponechanie súboru pozorovaní bez zásahu a aplikovanie regularizáciu Ridge, príp. Lasso, v lineárnom modeli. V druhom riešení sme vytvorili množinu premenných, ktoré boli medzi sebou silno korelované a ponechali si z nich len jednu premennú s najsilnejším vplyvom na vysvetľovanú premennú. Druhá menovaná metóda nie je v rámci tradičnej ekonometrie preferovaná. Pri úprave dát v strojovom učení je však pomerne používaná.¹¹⁶ Vo výsledku, odstraňuje 80 nezávislých premenných.

4.3 Fáza modelovania a hodnotenia

Všetky modely, ktoré využívame v práci na odhadovanie kreditnej kvality sú postavené na predpoklade lineárnej separácie zlyhaných úverov od splatených. Techniky, na ktorých sú postavené tieto modely boli vysvetlené v kapitole 3. *Metodika a metódy skúmania*. V ďalšom texte sa budeme venovať výlučne špecifikám, ktoré sú charakteristické pre aplikované modely.

4.3.1 Nastavenie hyperparametrov modelu

Niektoré modely si vyžadujú úpravu hyperparametrov, aby sa prispôbili analyzovaným dátam a zlepšil sa ich výkon. Ide primárne o regularizované modely, v ktorých optimalizujeme intenzitu regulácie. Pre zjednodušenie vyhľadávania ideálneho hyperparametru používame mriežkové vyhľadávanie s 3-násobnou krížovou validáciou (grid search with 3-fold cross validation). Štandardom je 5 alebo 10-násobná validácia, na to však vzhľadom na veľkosť súboru a dostupné technické vybavenie nemáme výpočtovú kapacitu. Interval, v ktorom boli hľadané optimálne hodnoty udáva *Tabuľka 8*.

¹¹⁵ Pôvodne sme plánovali použiť interval 0,8 – 1, no hneď niekoľko atribútov sa blížilo k spodnej hranici tohto intervalu, pričom za nimi už bola veľká medzera v hodnotách.

¹¹⁶ DORMANN, Carsten F. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. In *Ecography*, vol. 36, issue 1. January, 2013. p. 27-46

Tabuľka 8 Optimalizácia hyperparametrov individuálnych klasifikačných metód

Akronym	Metóda	Parameter	Interval
LR	Logistická regresia	-	-
L2-LR	Ridge reg. logistickej regresie	C *	[0.1, 0.01, 0.005]
L1-LR	Lasso reg. logistickej regresie	C	[0.1, 0.01, 0.005]
El-LR **	Elasticnet reg. logistickej regresie	pomer l1/l2	[0.6, 0.5, 0.4]
		C	[0.1, 0.01, 0.005]
SVM	Lineárny Support Vector Machine	-	-
Reg-SVM	Regularizovaný lineárny SVM	penalizácia	[l1, l2]
		C ***	[1, 0.1, 0.01]

* hyperparameter C je prevrátená hodnota regularizačného člena lambda. Čím nižšie C, tým silnejšia regularizácia.

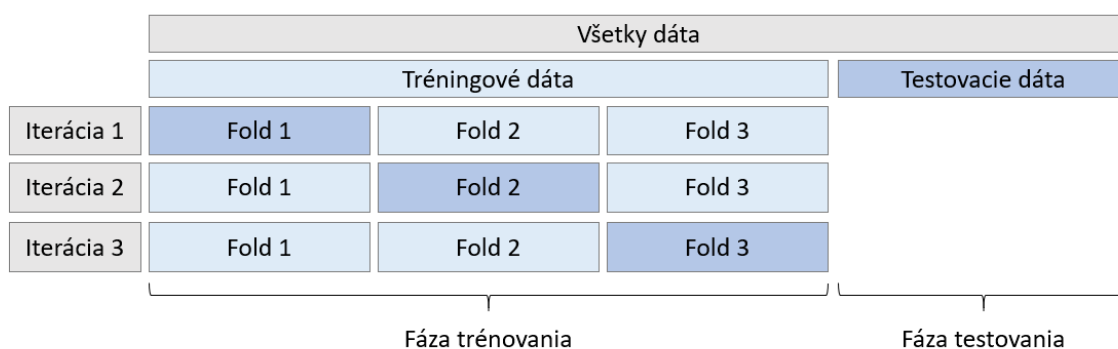
** ak je pomer l1/l2 rovný 0, využitá je penalizácia l2, pri hodnote 1 je použitá penalizácia l1.

*** čím nižšia hodnota C, tým väčšie narušenie rozpätia (margin) je povolené.

Zdroj: Vlastné spracovanie

Krížová validácia a s ňou spojené hľadanie najvhodnejšieho nastavenia hyperparametrov modelu prebieha trénovaním základného typu modelu s rôznym nastavením z daného intervalu. Tréningsové dáta sú náhodne rozdelené, v našom prípade, do troch rovnako veľkých vzoriek, pričom vždy jedna vzorka je ponechaná ako testovacia pre danú iteráciu a ostatné pozorovania vytvoria jeden tréningsový set (*Obrázok 17*). Každý model sa trénuje trikrát a vypočítava sa jeho priemerná úspešnosť na jednotlivých testovacích vzorkách. Trénuje sa taký počet modelov, aby odpovedal možným kombináciám rôznych nastavení hyperparametrov.

Obrázok 17 Príklad uplatnenej krížovej validácie



Zdroj: Vlastné spracovanie

Krížová validácia si vyžaduje definovanie vyhodnocovacej metriky ex ante, podľa ktorej sa určí najsprávnejšie nastavenie hyperparametrov. Pre logistickú regresiu ako metriku vyberáme AUROC, ktorá je štandardom pri posudzovaní výkonnosti binárneho klasifikátora. Lineárny SVM v *Pythone* nie je implementovaný, aby zvládal prevádzať každú predikciu na pravdepodobnostné vyjadrenie, a preto v tomto prípade používame F1 skóre, ktoré je svojou povahou podobné AUROC.

4.3.2 Metódy hodnotenia

Ako sme uviedli v sekcii 1.7.2 *Prehľad literatúry*, pri metrikách hodnotenia modelov kreditného skóringu panuje istá nejednoznačnosť. Väčšina štúdií sa spolieha na jednu metriku alebo metriky rovnakého typu. Štatistické merania výkonnosti sa vo všeobecnosti delia na tri typy. Tie, ktoré hodnotia diskriminačnú schopnosť kreditného skórovania (napr. AUROC). Tie, ktoré posudzujú presnosť pravdepodobnosti predpovedí (napr. Brier Score). A tie, ktoré hodnotia správnosť kategorických predpovedí (napr. klasifikačná tabuľka a chyby klasifikácie). Len málo štúdií kombinuje hodnotiace metódy z rôznych kategórií. Tento nedostatok sa snažíme v našej práci adresovať a pristúpiť ku komplexnejšiemu pohľadu na predikcie modelu. Každopádne, v celom obraze však chýba to najdôležitejšie, čo úverovú inštitúciu zaujíma, finančný rozmer. Len malé percento štúdií hodnotí model podľa vývoja zisku a strát. Naším cieľom je pokryť aj túto námietku a priniesť manažérsky pohľad.

Začnime najjednoduchšou technikou – presnosť (accuracy). Presnosť, skúma počet správnych predikcií. Model je presný, ak sa čo najviac jeho predpovedí zhoduje so skutočným stavom. Problém je dominancia jednej udalosti nad druhou. Inými slovami, jedna udalosť nastáva násobne častejšie ako iná. Vysokú presnosť vieme dosiahnuť predpovedaním výlučne dominantného výsledku. Podporovať proces schvaľovania úverov rozhodovacím modelom, ktorý preferuje stav – schváliť, by viedlo k nárastu nedobytných úverov a prepadu ziskovosti. Túto metriku preto ponúkame ako doplnkovú.

Pri klasifikačných úlohách sa preferuje „matica pomýlenia“ (confusion matrix) alebo nazývaná aj tabuľka klasifikácií. Na výpočet matice potrebujeme dostať predpovede modelu o pozorovaniach. Tie následne porovnáme so skutočným stavom. Každý riadok prezentuje skutočný stav a každý stĺpec predikciu modelu. Vzorový príklad matice pomýlenia ponúka *Obrázok 18*. Výhodou tejto tabuľky je, že sa z nej dajú vyvodiť ďalšie kritéria hodnotenia.

Obrázok 18 Tabuľka klasifikácií a jej derivácie

		Predpoveď modelu		
		Trieda 0	Trieda 1	
Skutočný stav	Trieda 0	True Negative (skutočne negatívny)	False Positive (nesprávne pozitívny)	(Precision) Pos. Pred. Value $PPV = TP/(TP+FP)$
	Trieda 1	False Negative (nesprávne negatívny)	True Positive (skutočne pozitívny)	Neg. Pred. Value $NPV = TN/(FN+TN)$
		(Recall / Sensitivity) True Pos. Rate $TPR = TP/(TP+FN)$	(Specificity) True Neg. Rate $TNR = TN/(FP+TN)$	

* trieda 0 označuje splatený úver. Pozitívna trieda 1 označuje problémový úver.

Zdroj: Vlastné spracovanie

Na základe tabuľky klasifikácií vidíme, či je model schopný rozlíšiť medzi sledovanými triedami. Ideálny model sa vyznačuje silnou ľavou diagonálou v tabuľke (znázornená tmavou farbou na *Obrázku 18*), kedy priradzuje pozorovania do správnych tried.

Z tejto tabuľky vieme vyvodiť ďalšie metriky, ktoré sú zaujímavé na sledovanie. Jednou z nich je miera skutočne pozitívnych predikcií (true positive rate – TPR), nazýva sa aj recall. Poukazuje na schopnosť modelu odhaľovať pozitívne prípady. V kontexte práce takto označujeme neplatičov. Ďalšou používanou jednotkou je miera skutočne negatívnych predikcií (true negative rate – TNR), nazýva sa aj specificity. Tá udáva dispozíciu modelu detegovať negatívne prípady. Pri úveroch ide o bezproblémovo splatené prípady. Často sa vyjadruje aj pozitívna predikčná hodnota (positive predictive value – PPV), nazývaná precision. Udáva aké percento z modelom označených pozitívnych prípadov je skutočne aj pozitívnych. V kontexte úverovania hovoríme o odmietnutých žiadateľoch, ktorí by prípadný úver reálne nesplátili. Precision sa snažíme maximalizovať, aby sme predchádzali nákladom obetovanej príležitosti.

Z tabuľky klasifikácií vyvodzujeme operačnú krivku ROC. Presnejšie sa budeme venovať výpočtu plochy pod operačnou krivkou AUROC alebo AUC. Model zaradzuje skúmané pozorovania do tried s určitou pravdepodobnosťou. Prirodzenou hranicou je 0,5. Túto hranicu však vieme zmeniť a tým ovplyvniť klasifikáciu do tried. Spravidla tak robíme ak je jedna skupina kvôli svojej početnosti vo vzorke preferovaná pred inou (naš prípad). Pri rôznych hraniciach dosahujeme rôzne hodnoty pre TPR a FPR. Na zostrojenie ROC krivky

musíme vypočítať TPR a FPR pre každú hranicu a zaznačiť hodnoty do grafu. Čím viac je krivka umiestnená vľavo hore, tým je model presnejší. AUROC je numerické vyjadrenie tvaru krivky a udáva plochu pod ňou. Nadobúda hodnoty od 0 do 1. Ideálny mechanizmus klasifikácie má AUROC rovný 1. Každý model s AUROC nad 0,8 je považovaný za veľmi kvalitný. AUROC je najpoužívanejšia metrika na hodnotenie účinnosti modelu binárnou závislou premennou, preto jej aj my budeme pripisovať veľkú váhu pri hodnotení modelov.

F-skóre pomáha zhrnúť výsledky tabuľky klasifikácií do jedného numerického ukazovateľa. F-skóre kombinuje precision a recall v jednej metrike. Vypočítava sa spôsobom uvedením v Rovnici 4.1.

$$F_{beta} = (1 + \beta^2) * \frac{precision * recall}{\beta^2 * precision + recall} \quad (4.1)$$

Čím väčšia β v F-beta skóre, tým väčší dôraz na recall. F1 predstavuje stav, pri ktorom sa kladie rovnaký dôraz na recall aj na precision. Pri skóre F2 je recall dvakrát tak dôležitý ako v prípade F1. F-skóre v intenciách schvaľovania úverov predstavuje stav, kedy vieme rozpoznať pozitívne prípady (budúci nesplatený úver) pri minimalizovaní zamietnutých žiadostí, ktoré by v skutočnosti boli splatené. Inými slovami, aby množinu zamietnutých žiadostí tvorili v čo najväčšej miere zlyhané úvery. Čím vyššie F2 skóre, tým nižšie náklady obetovanej príležitosti.

Lessmann¹¹⁷ navrhuje pri podobných odhadoch používať Brierovo skóre (Brier score). BS udáva pravdepodobnosť s akou model predpovedá, že žiadateľ úver nesplátí. Inak povedané, kvantifikuje mieru rizika obsiahnutú v žiadosti. Tento aspekt je mimoriadne dôležitý pri určovaní za aký úrok daný úver banka môže poskytnúť. BS meria priemernú kvadratickú chybu v pravdepodobnostnom priestore, teda rozdiel medzi pravdepodobnosťou zaradenia do určitej triedy a jej členstvom v danej triede.

$$brier\ score = (y_{pravdepodobnosť\ pre\ triedu\ 1} - y_{skutočná\ trieda})^2 \quad (4.2)$$

4.3.3 Hodnotenie kvality modelu

Po optimalizácii hyperparametrov na tréningových a validačných dátach sme pristúpili k hodnoteniu presnosti modelov na testovacích dátach. Pre testovacie dáta je špecifické to, že ide o údaje, ktoré náš model nevidel v tréningovej fáze. Hodnoty atribútov u týchto žiadateľov

¹¹⁷ LESSMANN, Stefan et al. Benchmarking state-of-the art classification algorithms for credit scoring: An update of research. In *European Journal of Operational Research*, vol. 247, issue 1. May 2015. p. 124-136.

neboli zohľadnené ani pri úprave dát, nezahrnuli sa do výpočtu strednej hodnoty doplnenej do prázdnych buniek, ani do procesu škálovania. Týmto krokmi zabezpečíme simulované žiadanie o úver novými klientmi. V ďalšom texte prezentujeme výsledky hodnotenia modelov.

Tabuľka 9 Výsledky modelov na upravenom datasete bez selekcie atribútov

Model	AUC	TPR	FPR	F2	BS	Presnosť
LR	.7655	.6960	.2997	.4312	.0679	.7000
L2-LR	.7656	.6940	.2993	.4302	.0680	.7001
L1-LR	.7656	.6929	.2993	.4296	.0679	.7000
El-LR	.7657	.6911	.3005	.4279	.0680	.6988
SVM	.7647	.6882	.2927	.4310	.0680	.7057
Reg-SVM	.7655	.7017	.3050	.4312	.0680	.6955

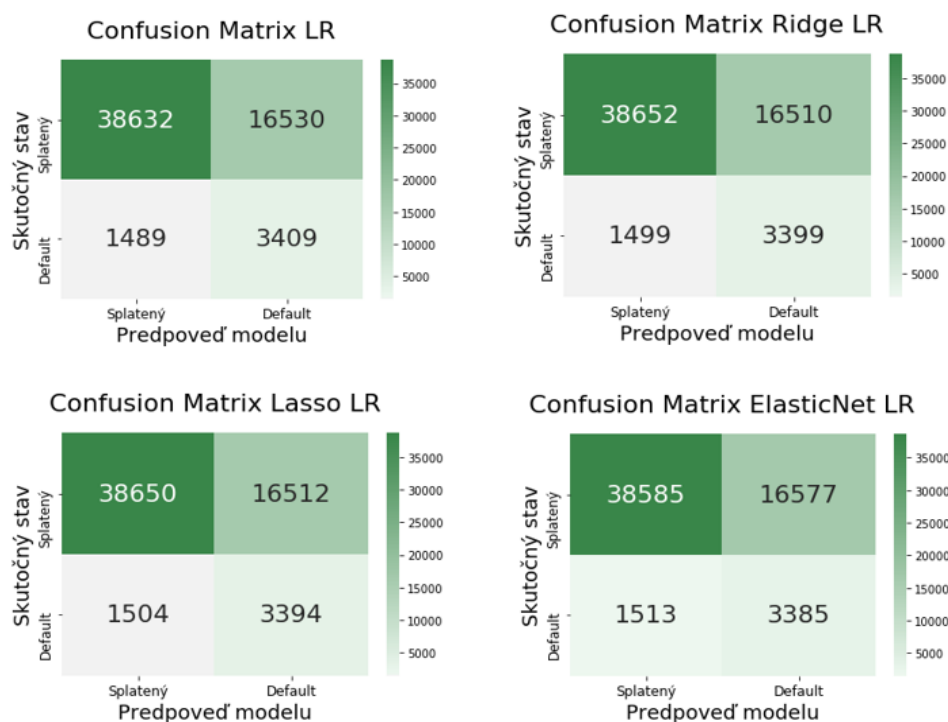
* tmavým sú zvýraznené najlepšie namerané hodnoty pre daný ukazovateľ

Zdroj: Vlastné spracovanie

Pri pohľade na výsledky testu nadobúdame pocit, že sme dosiahli na tomto súbore dát limit predikčných možností modelov s lineárnym prístupom. Hodnoty posudzovaných veličín sú veľmi podobné a líšia sa spravidla až na štvrtom desatinnom mieste. Výsledky prezentujeme v *Tabuľke 9*. Najlepšiu AUROC dosiahol El-LR, no v ostatných ukazovateľoch končil pomerne nízko. Zároveň sme pri El-LR narazili aj na časový problém v tréningovej fáze (*Tabuľka 13*), čím sa tento prístup stal najmenej praktický. So zreteľom na špecifiká bankovníctva nás zaujíma primárne TPR, F2 skóre a BS. V týchto hodnoteniach dosiahol najlepší výsledok regularizovaný SVM model s lineárnym jadrom. O niečo slabší výsledok registrujeme pre základný model logistickej regresie. Musíme pripustiť, že dosiahnuté F2 a FPR nie sú práve najideálnejšie ani pri jednom meraní. Konkrétne to znamená, že model by zamietal dosť veľa žiadostí o úver osobám, ktoré by v skutočnosti boli schopné daný úver splatiť. Jedným dychom musíme podotknúť, že proces prideľovania úverov nie je triviálny, a preto predpokladáme, že ani iný prístup by pri danej TPR (odhaľovaní zlyhaných úverov) neprinesol výrazne nižšiu úroveň nákladov obetovaných príležitostí.

Tabuľky klasifikácií na *Obrázku 19* a *Obrázku 20* prinášajú podrobný obraz o tom, ako úspešne boli jednotlivé modely pri rozdeľovaní žiadateľov do tried. Spomedzi logit modelov najvyšší počet nesplatených úverov dokázal zachytiť neregularizovaný model.

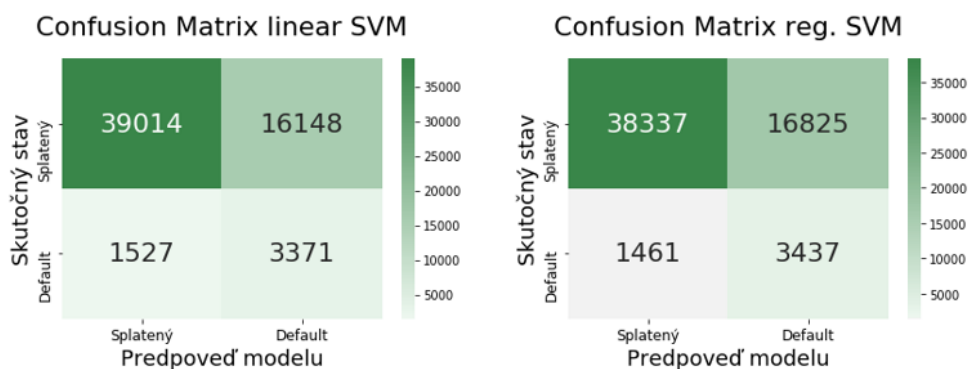
Obrázok 19 Tabuľka klasifikácií pre LR – dataset bez selekcie atribútov



Zdroj: Vlastné spracovanie

V počte detegovaných nesplatených úverov prekonal základný LR model regularizovaná metóda pomocných vektorov s lineárnym jadrom. Reg-SVM však zamietol spomedzi všetkých modelov najvyšší počet žiadateľov. Predikcie skúmaných modelov sú veľmi podobné a záleží na rozhodnutí manažmentu, či preferuje nižšie straty na portfóliu, alebo väčší podiel na trhu s úvermi.

Obrázok 20 Tabuľka klasifikácií pre SVM – dataset bez selekcie atribútov



Zdroj: Vlastné spracovanie

Zo štatistického hľadiska je výsledok experimentu pozitívny, čo potvrdzuje AUC blízke hodnote 0,8. Aj Brierovo skóre na úrovni 0,07 poukazuje, že modely sú relatívne dobré

v rozlišovaní mieri rizikovosti medzi klientami. Aby boli výsledky relevantné pre bankovú prax, potrebuje preukázať ešte zlepšenie na strane finančných ukazovateľov.

Tabuľka 10 Manažerský pohľad – dataset bez selekcie atribútov

Model	Celkový poskytnutý kapitál	Nesplatený kapitál	Podiel zlyhanej investície	Zamietnuté žiadosti
Home Credit úvery	32.40 mld.	2.69 mld.	8.19 %	-
LR	23.40 mld.	0.88 mld.	3.76 %	33.20 %
L2-LR	23.42 mld.	0.89 mld.	3.79 %	33.15 %
L1-LR	23.41 mld.	0.88 mld.	3.79 %	33.14 %
El-LR	23.38 mld.	0.90 mld.	3.85 %	33.23 %
SVM	23.73 mld.	0.91 mld.	3.83 %	32.50 %
Reg-SVM	23.24 mld.	0.87 mld.	3.74 %	33.74 %

Zdroj: Vlastné spracovanie

Tabuľka 10 zachytáva finančný rozmer navrhovaného prístupu v intenciách testovacieho súboru. Home Credit pri využití svojho modelu poskytol žiadateľom v testovacej vzorke úvery vo výške 32,84 mld., z čoho 2,69 mld. nebolo splatených v plnej miere.¹¹⁸ Miera zlyhaných úverov v peňažnom vyjadrení predstavuje 8,19 %. Ak by sme schvaľovali úvery pre tých istých klientov s využitím reg-SVM modelu, dosiahli by sme mieru zlyhaných finančných prostriedkov iba na úrovni 3,74 %, čo je výrazné zlepšenie oproti východiskovému stavu. Model by však v porovnaní s pôvodným bol o niečo prudentnejší a poskytol by úvery len vo výške 23,24 mld. Nedobytné položky by sa však nepresiahli 0,87 mld.

Rovnaký aparát modelov sme aplikovali aj na dataset so selektovanými atribútmi. Odstránené boli štatisticky nevýznamné a silno korelované premenné. Výsledky prezentujeme v *Tabuľke č. 11*.

¹¹⁸ Z poskytnutých dát nebolo možné určiť v akom momente životnosti úveru došlo k zlyhaniu, a či bola splatená aspoň časť poskytnutej sumy. Výpočet preto zjednodušujeme a za nesplatenú považujeme celú výšku úveru.

Tabuľka 11 Výsledky modelov na upravenom datasete po selekcii atribútov

Model	AUC	TPR	FPR	F2	BS	Presnosť
LR	.7651	.6956	.3003	.4306	.0679	.6994
L2-LR	.7652	.6958	.3000	.4301	.0679	.6997
L1-LR	.7653	.6960	.2993	.4309	.0679	.6996
El-LR	.7652	.6954	.3000	.4306	.0679	.6996
SVM	.7644	.6846	.2931	.4286	.0680	.7051
Reg-SVM	.7653	.6950	.2995	.4307	.0680	.7000

Zdroj: Vlastné spracovanie

Na základe výsledkov v *Tabuľke 11* konštatujeme, že odstránenie štatisticky nevýznamných alebo vzájomne korelovaných atribútov, nepomohlo zlepšiť celkovú predikčnú kapacitu modelovacích techník. Najlepší výsledok tentokrát dosiahla logistická regresia s L1 regularizáciou. V porovnaní s prechádzajúcou sériou testov na neselektovaných dátach sme však zaznamenali mierne zhoršenie.

Tabuľka 12 Manažersky pohľad – dataset po selekcii atribútov

Model	Celkový poskytnutý kapitál	Nesplatený kapitál	Podiel zlyhanej investície	Zamietnuté žiadosti
Home Credit úvery	32.40 mld.	2.69 mld.	8.19 %	-
L1-LR	23.41 mld.	0.89 mld.	3.80 %	33.23 %

Zdroj: Vlastné spracovanie

Z pohľadu finančných veličín boli výsledky jednotlivých modelov po selekcii atribútov takmer totožné, a preto prezentujeme iba výsledok L1-LR, ktorý z porovnania modelov vyšiel najlepšie (*Tabuľka 12*). V porovnaní s Reg-SVM trénovanom na datasete bez selekcie atribútov zaznamenávame jemne vyšší nesplatený kapitál (v absolútnej aj relatívnej miere). Naopak, L1-LR poskytla o niekoľko desiatok úverov viac. Ostáva na rozhodnutí manažmentu, či preferuje rozširovať sieť klientov za cenu vyšších strát na portfóliu, alebo zlepšiť percentuálnu návratnosť investície.

Posledný aspekt, ktorý nám ostáva zhodnotiť je praktickosť modelov. Prediktívne modelovanie je síce exaktná veda, no niekedy sa ani jej nevyhýba technika pokus-omyl. Preto je lepšie pracovať s modelom, ktorého trénovanie trvá kratšiu dobu. Prehľad rýchlosti tvorby

modelov poskytuje *Tabuľka 13*. Z tohto hľadiska je ideálna voľba logistická regresia bez regularizácie.

Tabuľka 13 Rýchlosť modelov na rôznych datasetoch

Model	Atribúty bez selekcie		Atribúty po selekcií	
	Tréning*	Test	Tréning*	Test
LR	4 m 7 s	306 ms	1 m 54 s	182 ms
L2-LR	22 m 17 s	293 ms	9 m 6 s	143 ms
L1-LR	36 m 31 s	297 ms	11 m 5 s	142 ms
El-LR	2 h 5 m 40 s	300 ms	38 m 6 s	150 ms
SVM**	6 m 14 s	478 ms	3 m 3 s	239 ms
Reg-SVM**	38 m 14 s	491 ms	19 m 39 s	204 ms

* tréning zahŕňa aj fázu hľadania optimálneho nastavenia hyperparametrov pomocou krížovej validácie. Platí len pre tie modely, na ktoré sa validácia vzťahuje.

** čas je uvedený bez kalibrácie, ktorou zisťujeme pravdepodobnosti zlyhania, keďže v SVM modely toto nie je implementované. Určenie pravdepodobnosti je potrebné pre PD modely alebo ocenenie úveru.

Zdroj: Vlastné spracovanie

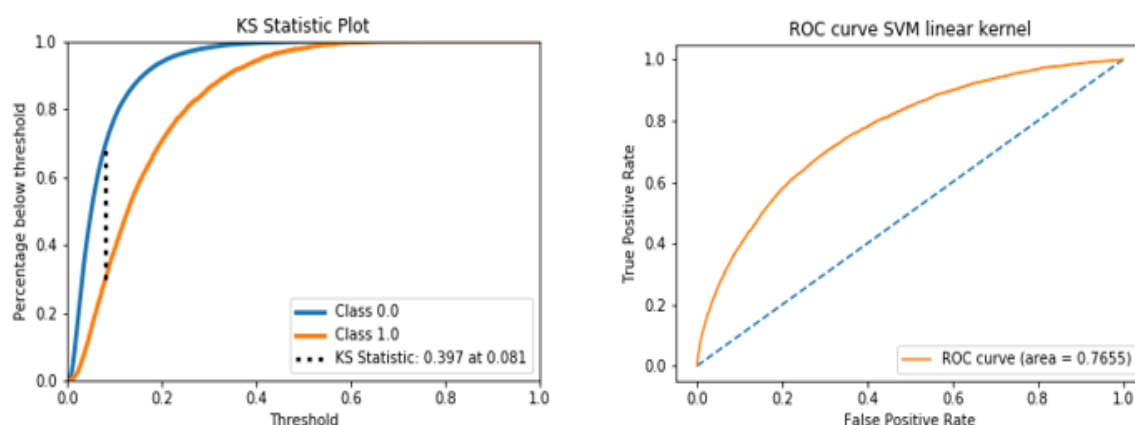
Prechod na zredukovaný súbor dát znížil čas trénovania na polovicu. Želané zlepšenie v presnosti predikcií však nepriniesol, práve naopak. Náš experiment týmto nepotvrdil teóriu o nutnosti redukovať počet premenných. Môžeme sa domnievať, že techniky strojového učenia sú uspokojené na to, aby z daných vstupov vždy vyťažili najrelevantnejšie závery. Každopádne, zhoršená úspešnosť na pozorovaniach po selekcii atribútov nie je výrazná. Ak by bola pre finančnú inštitúciu rozhodujúca výpočtová záťaž, môže k selekcii atribútov pristúpiť. Po odstránení multikolinearity predpokladáme, že úroveň interpretability modelov sa zvýšila. Túto tézu sme však neoverovali, a preto to nemôžeme kvalifikovane posúdiť.

4.3.4 Hodnotenie najlepšieho modelu

Kritériami na výber najlepšieho modelu boli výsledky v štyroch oblastiach, ktoré sú podľa nášho názoru najrelevantnejšie pre úverovú prax – TPR, F2, BS a miera zlyhaného kapitálu. Ako víťazný vzišiel lineárny SVM model s regularizačným parametrom $C=0,1$ trénovaný na neselektovaných atribútoch. Metóda pomocných vektorov na rozdiel od logistickej regresie nemá v sebe implementovanú funkciu kvantifikácie pravdepodobnosti s akou model rozhoduje o klasifikácii. K vyčísleniu pravdepodobnosti sme sa museli dopracovať cez krížovú validáciu predikcií a sledovaním ich úspešnosti. Pri súboroch,

v ktorých sú triedy nerovnomerne zastúpené je potrebné upraviť pravdepodobnostnú hranicu rozhodujúcu o zaradení pozorovaní. V našom prípade výrazne dominujú splatené úvery, a teda východisková hranica 0,5 určite nie je správnym nastavením. V tomto bode využívame Kolmogorov-Smirnov test (KS test), ktorý pomáha posúdiť rozdelenie na pozitívnu a negatívnu triedu. Najoptimálnejšie rozdelenie pravdepodobnosti dostaneme pri hranici 0,081. Daná hodnota sa spája s najlepším rozlíšením medzi triedami. Graf vľavo na *Obrázku 21* zobrazuje výsledok KS testu, v ktorom sú splatené úvery označené ako trieda 0 a nesplatené ako trieda 1. Bez ohľadu na nastavenie hranice, model má stále rozlišovaciu silu AUROC 0,7655 (*Obrázok 24* vpravo). Pravdepodobnostná hranica má vplyv iba na rozdelenie pozorovaní do tried.

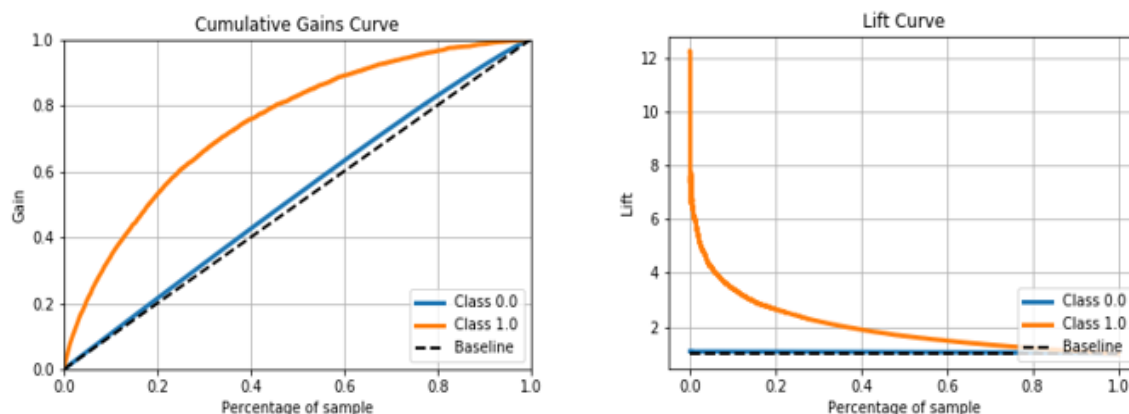
Obrázok 21 Kolmogorov-Smirnov test a krivka ROC



Zdroj: Vlastné spracovanie

V rámci nášho experimentu si musíme ešte zodpovedať otázku či a o koľko je náš model lepší ako náhodný generátor.

Obrázok 22 Kumulovaný zisk a krivka zdvihu



Zdroj: Vlastné spracovanie

Na grafe kumulovaného zisku (cumulative gains curve) vidíme ako je model schopný rozpoznávať jednotlivé triedy – *Obrázok 22* vľavo. Náhodný (baseline) model je znázornený prerušovanou čiernou čiarou. Modrá krivka, ktorá udáva silu rozoznávania splatených úver sa od náhodného modelu líši len mierne. Takmer celá predikčná sila modelu je definovaná jeho dispozíciou identifikovať budúce zlyhané úvery. Túto skutočnosť zachytáva oranžová krivka, ktorá sa podobá na ROC krivku súhrnného modelu.

Krivka zdvihu (lift curve) prezentuje koľkonásobne je použitý model lepší ako náhodný generátor – *Obrázok 22* vpravo. Modrá čiara (predikcia splatených úverov) opäť kopíruje náhodný model (prerušovaná čierna čiara). Priebeh oranžovej krivky (predikcia nesplatených úverov) nám dáva informáciu, že zvolený model je pre 40 % najjednoduchších pozorovaní 2-krát lepší ako náhodný. Pričom 5 % z pozorovaní zaradil model 4-krát presnejšie ako by to urobil náhodný výber.

Tieto výsledky sú jasným dôkazom, že proces klasifikácie žiadateľov o úver nie je triviálny. Predpokladáme, že mierne zlepšenie by sa dalo dosiahnuť zvýšením kvality dát. V získaných dátach bolo až 40 % atribútov významne znehodnotených. Rovnako by sa predikčná kapacita modelu mohla vylepšiť lepšou prácou s kompletným balíkom poskytnutých dát od HC. V intenciách našej práce, sme napríklad úplne opomenuli dáta z úverového registra (Credit Bureau), ktorý nám mohol poskytnúť informáciu o problémových žiadateľoch. Vzhľadom na rovnakú povahu vybraných modelov sa môžeme iba domnievať, či v dátach nie je ukrytý ešte nejaký iný, tentokrát nelineárny vzťah, ktorý by mohol byť komplexnejším modelovaním odhalený a využitý.

4.3.5 *Technické špecifiká experimentu*

Pre reprodukovateľnosť výsledkov uvádzame aj technické špecifiká zariadenia, na ktorom boli prevedené všetky experimenty. Úprava dát a tvorba modelov bola realizovaná za pomoci programovacieho jazyka *Python* verzie 3.6.8, v prostredí Jupyter notebook 6.0.3. Výpočty prebiehali na 64-bitovom Windows 10 zariadení s 2,1 GHz procesorom AMD FX-7500 Radeon a 8GB RAM. Použitých bolo niekoľko *Python* knižníc, dôležité boli najmä *pandas*, *numpy*, *scipy*, *matplotlib* a *sklearn*.

5 Diskusia

Systémy úverového hodnotenia sú významné nástroje manažmentu rizika v kreditných inštitúciách pri ochrane pred tzv. zlými úvermi. Ich úlohou je odhadovať individuálne predpoklady žiadateľov, ktoré vedú k neschopnosti riadne a na čas splatiť úver.

V časti 1.7.2 *Prehľad literatúry* sme sa venovali viacerým metódam tvorby kreditných modelov, ktoré sú prezentované ako alternatíva k odvetvovému štandardu vo forme logistickej regresie. Pokročilé a komplexné metódy strojového učenia v prezentovaných prácach často logistickú regresiu v hodnotení presnosti prekonávajú. Tento fakt, však na druhú stranu môže vzbudzovať legitímnu otázku, prečo ju tieto nové prístupy ešte nenahradili. Možnou odpoveďou je to, že presnosť je len jedným z kritérií, ktoré finančné inštitúcie zvažujú. Druhým sú náklady na implementáciu. V dnešnej dobe už cena za obstaranie IT vybavenia nie je až takou prekážkou ako v minulosti. Preto Finlay¹¹⁹ skôr poukazuje na iný aspekt. Komplexné modely podliehajú časovému znehodnoteniu viac ako jednoduchá logistická regresia. Model je nutné neustále prispôbovať zmenám v správaní sa klientov a ekonomickej realite, čo zvyšuje náklady na jeho prevádzku.

Hlavným cieľom tejto práce bolo predstaviť modelovanie kreditného rizika pomocou vybraných metód strojového učenia. Zvolili sme si koncept logistickej regresie a metódu pomocných vektorov, ktoré sme rozšírili o ich regularizované verzie a sledovali úspešnosť takýchto modelov. Ak chceme zodpovedať otázku, či sa nám podarilo naplniť tento cieľ, musíme výsledky našej práce porovnať s doterajším výskumom v tejto oblasti. Naš najlepší model dosiahol AUROC 0,766, čo hodnotíme ako uspokojivý výsledok. Modely, ktoré vykazujú AUROC nad 0,80 sú vo všeobecnosti považované ako veľmi kvalitné. Výsledkami našej práce sme sa tejto latke priblížili. Na tomto mieste je vhodné uviesť, aké AUROC hodnoty sú štandardom pre modelovanie kreditného rizika. Spomenuli sme, že najčastejšími východiskovými dátami sú Austria credit approval a German credit data. Louzada¹²⁰ analyzuje 189 prác publikovaných do roku 2015. Takmer 45% z nich pracovalo s týmito datasetmi. Vo výsledkoch uvádza, že úspešnosť na Austrian dataset sa najčastejšie nachádza v intervale 0,85-0,92 AUROC. Pre German dataset sa väčšina prác pohybovala v intervale

¹¹⁹ FINLAY, Steven. Multiple classifier architectures and their application to credit risk assessment. In *European Journal of Operational Research*, vol. 20, issue 2. April, 2011. p. 368–378.

¹²⁰ LOUZADA, Francisco – ARA, Anderson – FERNANDES, Guilherme B. Classification methods applied to credit scoring: A systematic review and overall comparison. In *Surveys in Operations Research and Management Science*, vol. 21, issue 2. December, 2016. p. 117-134.

0,73-0,82 AUROC. Ako konkrétny príklad uvádzame Zhanga¹²¹, ktorý skúma viacero modelovacích techník vo vzťahu k týmto datasetom. Hľadali sme štúdiu, v ktorej je použitá SVM s lineárnym jadrom a mohla by nám poslúžiť na porovnanie. V Austrian dataset Zhang so SVM dosiahol AUROC 0,87 a v German dataset 0,74. Oba tieto súbory dát predstavujú veľmi malú vzorku údajov. Spravidla do 1000 pozorovaní s 20 atribútmi. Väčšie datasety so sebou často prinášajú hluk (noise), resp. zhoršenú kvalitu a tým pádom aj nižšiu úspešnosť predikcií modelu. Kruppa¹²², ktorého sme už spomínali, pracuje s datasetom 64 524 pozorovaní. Pre neregularizovanú LR zaznamenal AUROC 0,748 a pre regularizovanú LR AUROC 0,779 s Brier Score 0,103 a 0,110. Tieto výsledky sú veľmi podobné záverom našej práce. Vzhľadom na dáta, ktoré sme použili, môže byť zaujímavé pozrieť sa na výsledky súťaže, ktorú vyhlásil Home Credit. Víťazný tím dosiahol AUROC 0,8057.¹²³ Riešitelia, ktorí sa umiestnili na najvyšších priečkach, používali spravidla nelineárne techniky a kombinované modely zložené z viacerých modelov (ensemble learning) a tie často dosahujú lepšie výsledky ako individuálne modely.

Nezodpovedanou otázkou preto ostáva, či by sme nedosiahli lepšie výsledky aplikáciou komplexnejších nelineárnych modelovacích techník. Zložitosť implementácie takýchto riešení však bola nateraz prekážkou. Retrospektívne je možno škoda, že sme sa nepokúsili z dostupných premenných vyťažiť nelineárne vzťahy aspoň vo forme polynomických atribútov.

Parciálnym cieľom bolo zmeniť pohľad na interpretáciu výsledkov. Doteraz sa väčšina prác zameriavala striktne na štatistické ukazovatele úspešnosti, ako napríklad AUROC. V tejto práci sa snažíme priniesť aj manažérsky pohľad na vec. Home Credit emitoval úvery v objeme 32,40 mld., z čoho 2,69 mld. bolo nesplatených. Podiel zlyhaných prostriedkov na celkovej expozícii predstavoval 8,19 %. Pri aplikácii nášho modelu na tej istej vzorke by boli výsledky o niečo iné. Emisia úverov by bola nižšia, konkrétne vo výške 23,24 mld., pričom nesplatené úvery by predstavovali len 0,87 mld. Percentuálne zastúpenie nedobytých prostriedkov by bolo 3,74 % z celkovej emisie.

¹²¹ ZHANG, Zhiwang – GAO, Guangxia – SHI, Yong. Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. In *European Journal of Operational Research* vol. 237, issue 1. August, 2014. p. 335-348

¹²² KRUPPA, Jochen et al. Consumer credit risk: Individual probability estimates using machine learning. In *Expert Systems with Applications*, vol. 40, issue 13. October, 2013. p. 5125–5131.

¹²³ <https://www.kaggle.com/c/home-credit-default-risk/leaderboard>

Záver

V práci sme priblížili vybrané techniky prediktívneho modelovania a poukázali na možné benefity, ktoré strojové učenie vie ponúknuť v oblasti kreditného skórovania. Zároveň sme sa snažili adresovať niektoré otázky, ktoré vystali z predchádzajúcich štúdií tejto problematiky. Analyzované dáta verne odpovedajú úverovej praxi, ako pôvodom, tak aj rozsahom, čo dosiahnutým výsledkom pridáva na relevantnosti a aktuálnosti.

Experimentálne výsledky preukázali, že prístup k modelovaniu kreditného rizika cez nástroje strojového učenia má reálne základy. Hodnota AUROC 0,766 je pri tak náročnej klasifikačnej úlohe skutočne dobrý výsledok. Pre banku je okrem klasifikácie na splatené a zlyhané úvery dôležité poznať aj pravdepodobnosť nesplatenia u jednotlivých žiadateľov. Pri hodnotení modelov sme použili aj metriku Brierové skóre, ktorá odráža istotu predikcií. Namerané BS bolo v priemere 0,068, čo je rovnako kvalitný výsledok aj keď s potenciálom na zlepšenie. Ďalšie parametre hodnotenia prezentujeme v *Tabuľke 9* a *Tabuľke 11*.

Z pohľadu banky, ako ekonomického subjektu orientovaného na dosahovanie zisku, je dôležitý aj finančný rozmer aplikácie modelu. V pôvodnom súbore úverov poskytnutých spoločnosťou Home Credit je zastúpenie nesplatených úverov vo výške 8,8 % a podiel zlyhaného kapitálu (kapitál poskytnutý žiadateľom, ktorí úver nesplátili) na úrovni 8,2 %. Po aplikácii odporúčaného modelu sme namerali podiel nedobytných úverov približne 3,8 %, čo v kapitálovom vyjadrení predstavovalo 3,7 % z poskytnutých prostriedkov. Tieto výsledky sme avšak dosiahli pri pomerne vysokej miere zamietnutých žiadostí. Z testovacej vzorky náš model neuspokojil cirka 1/3 žiadostí o úver. Informáciu o tom aký počet žiadateľov odmietal pôvodne používaný model spoločnosti Home Credit nemáme, a preto túto informáciu môžeme iba odhadovať zo všeobecne dostupných informácií. New York Federal Reserve Bank¹²⁴ uvádza, že podiel zamietnutých žiadostí o úver v roku 2018 predstavoval 19,9 %. V tomto kontexte je naša miera 33 % v zásade akceptovateľná. Domnievame sa, že nárast spôsobila povaha nebankového sektora, v ktorom sa úverov dožaduje primárne neprémiová klientela a tá celkovo dosahuje zlé kreditné skóre. Každá úverová inštitúcia sa však musí rozhodnúť sama za seba, či chce maximalizovať počet uspokojených žiadateľov a rozšíriť si svoju klientsku základňu, alebo minimalizovať nenávratné úvery a dosahovať vyššie zhodnotenie na investovanom kapitály.

¹²⁴ New York FED [elektronický zdroj] <https://www.newyorkfed.org/microeconomics/sce/credit-access#/experiences-credit-applications1>

Výzvy, ktoré sme si stanovili v kapitole 2. *Ciele práce* sme naplnili a odpovedali na otázky, s ktorými sme do výskumu vstupovali. Najdôležitejším bolo, že sme preukázali uspokojivé výsledky strojového učenia na dostatočne širokom produkčnom súbore dlžníkov. Druhým bodom bolo použitie nových štatistických metrík hodnotenia výsledkov modelu, ktoré by lepšie zodpovedali požiadavkám bankovej praxi. Uviedli sme dva nové spôsoby hodnotenia. F2 skóre, ktoré primárne sleduje chybnú klasifikáciu nesplatených úverov a prevzali Brierové skóre, ktoré hodnotí istotu s akou model triedi pozorovania. Ďalej sme proces hodnotenia modelu rozšírili o finančnú perspektívu, ktorá často chýba a na ktorú sa pozornosť upriamuje až v posledných rokoch.¹²⁵ Naša práca potvrdila konštatovania, ktoré prezentoval Giannone¹²⁶ preferujúci techniky hustého modelovania. Teda neredukovať v strojovom učení počet premenných aj keď sú samé o sebe štatisticky nevýznamné. V globále nám ponechanie takýchto atribútov pomohlo mierne vylepšiť predikčnú silu modelu.

V našej záverečnej práci sme sa výlučne opierali o štatistické metódy, ktorých predikcie sú ľahko vysvetliteľné a odôvodniteľné. Interpretovateľnosť výsledkov modelu je dôležitá požiadavka v tradičnom bankovom prostredí, v ktorom finančné domy podliehajú kontrole dohliadajúceho subjektu. Takáto potreba však neexistuje v nebankovom sektore, z ktorého dáta pochádzajú, a preto nechávame priestor budúcemu výskumu v tejto oblasti rozšíriť našu analýzu aj o komplikovanejšie nelineárne a neparametrické metódy.

Do budúca však ostávajú otvorené aj ďalšie otázky. Kreditné hodnotenie je dlhú dobu analyzované ako binárny klasifikačný problém. To však nerieši kľúčový cieľ banky, ktorým je vylepšenie ziskovosti na úverovom portfóliu. Uvažujeme nad tým, či by sa následný výskum nemal zaoberať modelovaním potenciálnej ziskovosti jednotlivých úverov. Pri stanovovaní ceny za úver by bolo pre banku prínosné, ak by takto zostavené modely boli schopné precízne kvantifikovať vnútornú rizikovosť klienta. V tejto oblasti navrhujeme silnejšiu orientáciu na kalibráciu pravdepodobnosti s akou model predikuje zlyhaný stav. Doterajší výskum prínosu samoučiacich systémov v oblasti kreditného rizika bol úzko orientovaný na prognózovanie budúceho stavu úveru. Zaujímavé by bolo sledovať prínos inteligentných softvérov pri modelovaní iných kreditných parametrov ako LGD, EAD, príp. CCF.

¹²⁵ KOZODOI, Nikita – LESSMANN, Stefan et al. A multi-objective approach for profit-driven feature selection in credit scoring. In *Decision Support Systems*, vol. 120. May, 2019. p. 106 – 117.

¹²⁶ GIANNONE, Domenico – LENZA, Michele – PRIMICERI, Giorgio E. Economic predictions with big data: The illusion of sparsity. In *Federal Reserve Bank of New York Staff Reports*, No. 847. April, 2018.

Zoznam použitej literatúry

Knižné zdroje:

- [1] BABOUČEK, Ivan a kol. *Regulace činnosti bank*. Praha: Bankovní institut vysoká škola, 2005, 317 s. ISBN 80-7265-071-8.
- [2] BAČA, Jozef. *Bankový dohľad*. Bratislava: Vydavateľstvo Ekonóm, 2001, 143 s. ISBN 80-225-1377-6.
- [3] DE HAAN, Jakob - OOSTERLOO, Sander - SCHOENMAKER, Dirk. *European financial markets and institutions*. Cambridge: Cambridge University Press, 2009, 410 p. ISBN 978-0-511-50686-4.
- [4] DE HAAN, Jakob - OOSTERLOO, Sander - SCHOENMAKER, Dirk. *Financial markets and institutions: european perspective*. 3rd ed. Cambridge: Cambridge University Press, 2015, 528 p. ISBN 978-1-107-53936-5.
- [5] FABOZZI, Frank J. *The handbook of financial instruments*. John Wiley & Sons, Inc. 2002, 864 p. ISBN 0-471-22092-2
- [6] GÉRON, Aurelien. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. 2st Edition, O'Reilly Media, Inc. 2019, 600 p. ISBN 978-1-492-03264-9.
- [7] HRVOLOVÁ, Božena - BADURA, Peter - MARKOVÁ, Jana. *Analýza finančných trhov*. 3. prepac. a dopl. vyd. Praha: Wolters Kluwer, 2015, 512 s. ISBN 978-80-7478-948-9.
- [8] HYRÁNEK, Eduard - BIKÁR, Miloš. *Manažment úverového financovania a úverových obchodov*. Vydavateľstvo EKONÓM, Bratislava, 2010, 224 s. ISBN 978-80-225-3042-2.
- [9] JURAFSKY, Daniel – MARTIN, James H.. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd Edition, Stanford University, 2014, 1024 p. ISBN 978-0130950697
- [10] POLOUČEK, Stanislav – KULHÁNEK, Lumír – MIRDALA, Rajmund et al. *Peniaze, banky, finančné trhy*. Bratislava: Iura Edition, 2010, 476 s. ISBN 978-80-8078-305-1.
- [11] REJNUŠ, Oldřich. *Finanční trhy: učebnice s programem na generování cvičných testů*. Praha: Grada Publishing, 2016, 384 s. ISBN 978-80-247-5871-8.
- [12] REVENDA, Zbyněk. *Centrální bankovníctví*. 3. aktualiz. vyd. Praha: Management Press, 2011, 558 s. ISBN 978-80-7261-230-7.

- [13] ROSE, Peter S. – MAREK, Karel. *Peněžní a kapitálové trhy*. Praha: Victoria Publishing, 1994, 1014 s. ISBN 80-85605-52-X.
- [14] SIVÁK, Rudolf - GERTLER, Ľubomíra - KOVÁČ, Urban. *Riziká a modely vo financiách a v bankovníctve*. Bratislava: Sprint dva, 2010, 346 s. ISBN 978-80-89393-44-2.
- [15] SIVÁK, Rudolf – GERTLER, Ľubomíra – KOVÁČ, Urban. *Riziká vo financiách a v bankovníctve*. Bratislava: Sprint dva, 2009, 345 s. Economics. ISBN 978-80-89393-03-9.
- [16] TKÁČOVÁ, Dana a kol. *Finančné trhy a bankovníctvo*. Bratislava. Wolters Kluwer, 2017, 310 s. ISBN 978-80-7552-528-4.
- [17] VANDERPLAS, Jake. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc. 2016, p548 p. ISBN 978-1491912058.

Časopisecké zdroje:

- [18] BAESESENS, Bart et al. Benchmarking state-of-the-art classification algorithms for credit scoring. In *Journal of the Operational Research Society*, Vol. 54. January, 2003. p. 627–635.
- [19] BECK, Thorsten – LAEVEN, Luc. Resolution of failed banks by deposit insurers: cross-country evidence. In: *Policy, Research working paper*. World Bank. May, 2006.
- [20] BELLONI, Alexandre – CHERNOZHUKOV, Victor – HANSEN, Christian B. Inference for high-dimensional sparse econometric models, In: *Advances in Economics and Econometrics*, 10th World Congress of the Econometric Society. December, 2011.
- [21] BERG, Tobias. On the rise of fintechs – credit scoring using digital footprints. In: *The Review of Financial Studies*. April, 2018.
- [22] CHANG, Yung-Chia et al. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. In: *Applied Soft Computing Journal*, Vol. 73. October, 2018. p. 914-920.
- [23] CHEN, Jiaqian – VERA, Giuseppe. Bank Balance Sheets and the Value of Lending. In: *IMF Working Paper*, No. 17/111. May, 2017. ISSN 1018-5941.
- [24] DORMANN, Carsten F. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. In *Ecography*, Vol. 36, Issue 1. January, 2013. p. 27-46
- [25] FINLAY, Steven. Credit scoring for profitability objectives. In: *European Journal of Operational Research*, Vol. 202. April, 2010. p. 528–537.

- [26] FINLAY, Steven. Multiple classifier architectures and their application to credit risk assessment. In: *European Journal of Operational Research*, Vol. 20, Issue 2. April, 2011. p. 368–378.
- [27] GAMBACORTA, Leonardo. How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. In: *BIS Working Papers* No 834. December, 2019.
- [28] GIANNONE, Domenico – LENZA, Michele – PRIMICERI, Giorgio E. Economic predictions with big data: The illusion of sparsity. In: *Federal Reserve Bank of New York Staff Reports*, No. 847. April, 2018.
- [29] GIETZEN, Thomas. Credit Scoring vs. Expert Judgement – a randomized controlled trial. In *Working papers on finance* No. 2017/09. Swiss Institute Of Banking And Finance. June, 2017.
- [30] KHANDANI, Amir E. et al. Consumer credit-risk models via machine learning algorithms. In: *Journal of Banking & Finance*, Vol. 34, issue 11. November, 2010. p 2767–2787.
- [31] KOZA, John R. et al. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: *Artificial Intelligence in Design '96*. Springer, Dordrecht. p. 151–170. ISBN 978-94-009-0279-4
- [32] KOZODOI, Nikita. – LESSMANN, Stefan et al. A multi-objective approach for profit-driven feature selection in credit scoring. In *Decision Support Systems*, vol. 120. May, 2019. p. 106 – 117.
- [33] KRUPPA, Jochen et al. Consumer credit risk: Individual probability estimates using machine learning. In: *Expert Systems with Applications*, Vol. 40, Issue 13. October, 2013. p. 5125–5131.
- [34] LESSMANN, Stefan et al. Benchmarking state-of-the art classification algorithms for credit scoring: An update of research. In: *European Journal of Operational Research*, Vol. 247, Issue 1. May, 2015. p. 124-136.
- [35] LOUZADA, Francisco – ARA, Anderson – FERNANDES, Guilherme B. Classification methods applied to credit scoring: A systematic review and overall comparison. In *Surveys in Operations Research and Management Science*, vol. 21, issue 2. December, 2016. p. 117-134.
- [36] MUNKHDALAI, Lkhagvadorj et al. An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. In *Sustainability MDPI Journals*, Vol. 11. January, 2019.

- [37] NIKOLIC, Nebojsa. The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. In: *Expert Systems with Applications*, Vol. 40, Issue 15. November, 2013. p. 5932-5944
- [38] REIM. Wiebke. et al. Risk management for product-service system operation. In: *International Journal of Operations & Production Management*. June, 2016. ISSN 0144-3577.
- [39] SCHÖLKOPF, Bernhard – LUO, Zhiyuan –VOVK, Vladimir. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Springer-Verlag Berlin Heidelberg, 2013. p. 4. ISBN 978-3-642-41136-6.
- [40] VAPNIK, Vladimir N. et al. Support vector clustering. In: *Journal of Machine Learning Research*, Vol. 2, 2001. p. 125–137.
- [41] ZHANG, Zhiwang – GAO, Guangxia – SHI, Yong. Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. In *European Journal of Operational Research*, vol. 237, issue 1. August, 2014. p. 335-348

Legislatíva, nariadenia a opatrenia:

- [42] Nariadenie Európskeho parlamentu a rady (EÚ) č. 549/2013 z 21. mája 2013 o európskom systéme národných a regionálnych účtov v Európskej únii.
- [43] Opatrenie NBS č. 4/2007 o vlastných zdrojoch
- [44] Zákon č. 129/2010 Z. z. o spotrebiteľských úveroch a o iných úveroch a pôžičkách pre spotrebiteľov a o zmene a doplnení niektorých zákonov
- [45] Zákon č. 314/2011 Z. z. ktorým sa mení a dopĺňa zákon č. 483/2001 Z. z. o bankách

Elektronické zdroje:

- [46] AVDJIEV, Stefan – GIUDICI, Paolo – SPELTA, Alessandro. *Measuring contagion risk in international banking*. [elektronický zdroj]. BIS Working Papers, No. 796. June, 2019. Dostupné na: <https://www.bis.org/publ/work796.pdf>
- [47] BOLTON, Christine. *Logistic regression and its application in credit scoring*. [elektronický zdroj]. University of Pretoria, 2019. p. 27. Dostupné na: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1024.2660&rep=rep1&type=pdf>
- [48] COSS, Soňa. Modely s binárne závislou premennou. [elektronický zdroj]. In: *Ekonomika Informatika*, vedecký časopis FHI EU v Bratislave a SSHI. Ročník 2015, náklad 15, číslo 2. Dostupné na: <http://ei.fhi.sk/index.php/EAI/article/view/18>

- [49] Financial Stability Board. [elektronický zdroj]. *Artificial intelligence and machine learning in financial services*. November, 2017. Dostupné na: <https://www.fsb.org/wp-content/uploads/P011117.pdf>
- [50] FUSEK, Roman. [elektronický zdroj]. Televízna stanica TA 3, Ekonomika, téma: Sú nebankovky "čistejšie"? Dostupné na: http://www.nbs.sk/_img/Documents/_Rozhovory/2018/Fusek_TA3_19_9.pdf
- [51] FUSTER, Andreas et al. *The Role of Technology in Mortgage Lending*. [elektronický zdroj]. Federal Reserve Bank of New York Staff Reports No. 863. Dostupné na: https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr836.pdf
- [52] NBS [elektronický zdroj]. Správa o činnosti Útvaru dohľadu nad finančným trhom Národnej banky Slovenska. Dostupný na: https://www.nbs.sk/_img/Documents/_Dohlad/Makropolitika/SC-UDF_2015_web.pdf
- [53] NBS [elektronický zdroj]. Štatistický bulletin 3. štvrťrok 2019. ISBN 1338-6328. Dostupné na: http://www.nbs.sk/_img/Documents/_Publikacie/BulletinMFS/2019/BulletinMFS_092019.pdf
- [54] NIELSEN, Michael [elektronický zdroj]. *Neural Network and Deep Learning*. Dostupné na: <http://neuralnetworksanddeeplearning.com/chap3.html>
- [55] ONUFEROVÁ, Marianna [elektronický zdroj]. Denník N. Dostupné na: <https://e.dennikn.sk/63329/zoznamy-neplaticov-ako-sa-dostanete-dnu-ako-von/>
- [56] PARALIČ, Ján [elektronický zdroj] *Pochopenie dát*. Technická univerzita v Košiciach. Dostupné na: <http://people.tuke.sk/jan.paralic/prezentacie/OZ/Pochopenie-dat.pdf>
- [57] ROUSE, Margaret [elektronický zdroj]. Dostupné na: <http://searchcompliance.techtarget.com/definition/risk-management>
- [58] SCHMIDT, Mark – LE ROUX, Nicolas – BACH, Francis [elektronický zdroj]. Minimizing Finite Sums with the Stochastic Average Gradient. In: *Mathematical Programming B*, Springer, vol. 162. p.83-112. Dostupné na: <https://hal.inria.fr/hal-00860051/document>
- [59] YU, Hwanjo – YANG, Jiong – HAN, Jiawei [elektronický zdroj]. Classifying Large Data Sets Using SVMs with Hierarchical Clusters. University of Illinois. Dostupné online: http://hanj.cs.illinois.edu/pdf/kdd03_scalesvm.pdf
- [60] [http://archive.ics.uci.edu/ml/datasets/statlog+\(australian+credit+approval\)](http://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval))
- [61] [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))
- [62] <http://www.ineko.sk/clanky/publikacie>
- [63] <https://www.kaggle.com/c/home-credit-default-risk/leaderboard>

- [64] <https://www.kaggle.com/c/home-credit-default-risk/overview>
- [65] <https://www.mckinsey.com/industries/financial-services/our-insights/global-banking-annual-review-2019-the-last-pit-stop-time-for-bold-late-cycle-moves>
- [66] <https://medium.com/future-vision/collinearity-what-it-means-why-its-bad-and-how-does-it-affect-other-models-94e1db984168>
- [67] <https://www.nbs.sk/sk/dohlad-nad-financnym-trhom/fintech>
- [68] <https://www.nyse.com/data/transactions-statistics-data-library>
- [69] <http://www.sbcbsk/o-spolocnosti/spolocny-register-bankovych-informacii-srbi/>
- [70] <http://www.tempest.sk/index.php/najhorucejsie-technologicke-trendy-roka-2020-s-ktorymi-sa-urcite-stretnete-aj-vy-5c1.html>

Prílohy

Príloha A: Opisná štatistika (údaje pred škálovaním vstupov na účely modelovania)

	Mean	Median	St.dev.	Max	Min	Skewness	Kurtosis
feature_001	0,01	0	0,12	1	0	8,03	62,42
feature_002	0,05	0	0,22	1	0	4,14	15,12
feature_003	0,04	0	0,2	1	0	4,7	20,12
feature_004	0,08	0	0,27	1	0	3,14	7,87
feature_005	0,23	0	0,42	1	0	1,28	-0,37
feature_006	0,18	0	0,38	1	0	1,67	0,77
feature_007	0,24	0	0,43	1	0	1,23	-0,5
feature_008	0,49	0	0,5	1	0	0,05	-2
feature_009	0,01	0	0,09	1	0	11,33	126,43
feature_010	0	0	0,01	1	0	158,22	25031,5
feature_011	0,71	1	0,45	1	0	-0,95	-1,1
feature_012	0	0	0,01	1	0	111,87	12513,25
feature_013	0,01	0	0,12	1	0	8,07	63,19
feature_014	0,09	0	0,28	1	0	2,9	6,4
feature_015	0	0	0,01	1	0	86,65	7505,95
feature_016	0,08	0	0,27	1	0	3,09	7,53
feature_017	0	0	0,06	1	0	16,21	260,64
feature_018	0	0	0	1	0	223,76	50068
feature_019	0	0	0,06	1	0	16,74	278,17
feature_020	0	0	0	1	0	387,58	150214
feature_021	0	0	0,06	1	0	17,75	312,93
feature_022	0	0	0,05	1	0	19,78	389,28
feature_023	0	0	0,03	1	0	30,88	951,81
feature_024	0,01	0	0,1	1	0	10,33	104,7
feature_025	0	0	0,02	1	0	63,27	4000,84
feature_026	0,01	0	0,09	1	0	11,35	126,72
feature_027	0	0	0,02	1	0	42,51	1804,87
feature_028	0	0	0,02	1	0	46,63	2172,09
feature_029	0	0	0,02	1	0	55,34	3060,69
feature_030	0,18	0	0,38	1	0	1,66	0,75
feature_031	0	-0,04	1	4,01	-3,67	0,02	0,18
feature_032	0	0,11	1	2,19	-3,35	-0,35	-0,35
feature_033	0	0,11	1	2,42	-5,04	-0,42	-0,12
feature_034	0	0,09	1	2,22	-3,19	-0,31	-0,34
feature_035	0	-0,13	1	3,74	-1,52	1,42	3,03
feature_036	0	0,07	1	1,72	-2,54	-0,3	-0,87
feature_037	0	0,09	1	2,58	-7,64	-0,54	0,55
feature_038	0	0,27	1	1,49	-6,86	-1,77	4,67
feature_039	0	0,36	1	1,22	-8,33	-2,37	7,92
feature_040	0	-0,18	1	4,53	-3,34	0,89	2,79
feature_041	0	0,1	1	2,58	-2,86	-0,33	-0,63
feature_042	0	0,27	1	1,79	-2,68	-0,79	-0,28
feature_043	0	0,02	1	2,21	-2,77	-0,26	-0,46

feature_044	0	0,03	1	4,16	-3,97	-0,59	3,48
feature_045	0	0,14	1	4,27	-3,47	-1,85	5,91
feature_046	0	0,04	0,99	0,14	-24,06	-23,93	574,88
feature_047	0	0,05	1	2,24	-34,17	-16,25	493,47
feature_048	0	-0,07	1	7,15	-2,4	1,58	7,77
feature_049	0	-0,46	1	3,64	-0,46	1,87	1,86
feature_050	0	0,11	1	4,69	-6,31	-0,7	4,07
feature_051	0	-0,06	1	3,22	-5,48	-1,64	8,54
feature_052	0	0,12	1	3,16	-6	-2,49	10,88
feature_053	0	0,08	1	4,98	-2,93	-0,65	3,68
feature_054	0	0,01	1	5,93	-5,11	-0,42	7,6
feature_055	0	0,01	1	4,25	-3,62	-0,19	3
feature_056	0	-0,28	1	13,62	-0,28	5,65	44,67
feature_057	0	-0,3	1	6,23	-0,75	2,53	6,59
feature_058	0	0,03	1	4,15	-3,85	-0,58	3,39
feature_059	0	0,14	1	4,16	-3,31	-1,76	5,23
feature_060	0	0,05	1	0,13	-21,61	-21,26	450,81
feature_061	0	0,06	1	2,22	-35,36	-17,31	551,05
feature_062	0	-0,08	1	7,25	-2,27	1,74	8,05
feature_063	-4,12	-4,61	1,1	0,01	-4,61	1,96	2,24
feature_064	-1,98	-1,91	0,44	0,01	-4,61	-0,77	3,6
feature_065	-1,71	-1,73	0,54	0,01	-4,61	-1,76	8,75
feature_066	-1,59	-1,52	0,51	0,01	-4,61	-2,51	10,6
feature_067	0	0,08	1	4,94	-2,8	-0,56	3,37
feature_068	0	-0,02	1	5,77	-5,03	-0,33	7,51
feature_069	0	0,02	1	4,22	-3,54	-0,19	2,98
feature_070	0	-0,27	1	14,21	-0,27	5,92	49,14
feature_071	0	-0,37	1	6,2	-0,52	2,71	7,19
feature_072	0	0,02	1	4,15	-3,95	-0,56	3,47
feature_073	0	0,14	1	4,27	-3,45	-1,84	5,84
feature_074	0	0,04	1	0,14	-23,47	-23,32	544,31
feature_075	0	0,05	1	2,25	-34,86	-16,55	514,99
feature_076	0	-0,07	1	7,14	-2,38	1,6	7,79
feature_077	0	-0,46	1	3,66	-0,46	1,89	1,96
feature_078	0	0,12	1	4,66	-6,26	-0,71	3,97
feature_079	0	-0,06	1	3,22	-5,46	-1,65	8,56
feature_080	0	0,13	1	3,15	-5,97	-2,49	10,78
feature_081	0	0,09	1	4,93	-2,92	-0,66	3,63
feature_082	0	0,01	1	5,88	-5,12	-0,46	7,63
feature_083	0	0	1	4,22	-3,62	-0,18	3,02
feature_084	0	-0,28	1	13,74	-0,28	5,71	45,5
feature_085	0	-0,31	1	6,21	-0,7	2,58	6,75
feature_086	0	0	1	4,17	-3,37	-0,14	2,57
feature_087	0	0,37	1	1,05	-2,29	-1,51	0,84
feature_088	0	-0,04	1	3,16	-3,47	0,24	-0,22
feature_089	0	-0,04	1	3,82	-4,41	0,29	0,16
feature_090	0	-0,02	1	2,11	-2,39	-0,39	-0,57
feature_091	0	0,01	1	3,02	-5,24	0,01	-0,12

feature_092	0	-0,26	1	4,76	-1,23	1,62	2,8
feature_093	0,09	0	0,29	1	0	2,83	6
feature_094	0,34	0	0,47	1	0	0,68	-1,54
feature_095	0,34	0	0,47	1	0	0,68	-1,53
feature_096	0,7	1	0,46	1	0	-0,85	-1,28
feature_097	0,7	1	0,46	1	0	-0,87	-1,24
feature_098	0,2	0	0,4	1	0	1,51	0,27
feature_099	0,09	0	0,28	1	0	2,93	6,57
feature_100	0,01	0	0,11	1	0	8,93	77,75
feature_101	0,13	0	0,34	1	0	2,18	2,76
feature_102	0	0	0,03	1	0	33,75	1137,35
feature_103	0	0	0,05	1	0	18,77	350,13
feature_104	0,01	0	0,08	1	0	13,03	167,87
feature_105	0,04	0	0,19	1	0	4,89	21,88
feature_106	0,81	1	0,39	1	0	-1,59	0,53
feature_107	0,23	0	0,42	1	0	1,28	-0,35
feature_108	0	0	0	1	0	245,12	60082,6
feature_109	0,18	0	0,38	1	0	1,66	0,75
feature_110	0,07	0	0,26	1	0	3,36	9,29
feature_111	0	0	0,01	1	0	137,02	18772,37
feature_112	0	0	0,01	1	0	125,74	15807,53
feature_113	0,52	1	0,5	1	0	-0,08	-1,99
feature_114	0,01	0	0,11	1	0	8,77	74,85
feature_115	0,71	1	0,45	1	0	-0,95	-1,09
feature_116	0,24	0	0,43	1	0	1,22	-0,5
feature_117	0,1	0	0,3	1	0	2,72	5,4
feature_118	0,64	1	0,48	1	0	-0,58	-1,66
feature_119	0,06	0	0,24	1	0	3,56	10,66
feature_120	0,15	0	0,35	1	0	2	1,98
feature_121	0,89	1	0,32	1	0	-2,45	3,99
feature_122	0,04	0	0,19	1	0	4,94	22,43
feature_123	0,01	0	0,09	1	0	10,74	113,38
feature_124	0,02	0	0,13	1	0	7,75	58
feature_125	0,05	0	0,21	1	0	4,21	15,71
feature_126	0,82	1	0,38	1	0	-1,66	0,75
feature_127	0,2	0	0,4	1	0	1,51	0,27
feature_128	1	1	0,04	1	0	-23,39	545,25
feature_129	0,28	0	0,45	1	0	0,98	-1,04
feature_130	0,06	0	0,23	1	0	3,84	12,78
feature_131	0,02	0	0,12	1	0	7,91	60,5
feature_132	0,02	0	0,14	1	0	6,94	46,13
feature_133	0,09	0	0,29	1	0	2,88	6,3
feature_134	0,06	0	0,24	1	0	3,67	11,51
feature_135	0	0	0,04	1	0	23,41	546,26
feature_136	0,04	0	0,19	1	0	4,91	22,14
feature_137	0	0	0,04	1	0	24,7	608,14
feature_138	0,18	0	0,38	1	0	1,66	0,76
feature_139	0,01	0	0,08	1	0	11,93	140,43

feature_140	0,07	0	0,25	1	0	3,46	9,99
feature_141	0,03	0	0,16	1	0	5,73	30,81
feature_142	0,01	0	0,09	1	0	10,62	110,74
feature_143	0	0	0,05	1	0	20,1	402,1
feature_144	0,1	0	0,31	1	0	2,58	4,64
feature_145	0	0	0,07	1	0	15,2	229,17
feature_146	0,02	0	0,15	1	0	6,52	40,49
feature_147	0	0	0,07	1	0	14,94	221,24
feature_148	0,22	0	0,41	1	0	1,35	-0,17
feature_149	0,52	1	0,5	1	0	-0,06	-2
feature_150	0,17	0	0,38	1	0	1,74	1,04
feature_151	0,08	0	0,27	1	0	3,08	7,51
feature_152	0,1	0	0,3	1	0	2,64	4,95
feature_153	0,74	1	0,44	1	0	-1,09	-0,8
feature_154	0,11	0	0,31	1	0	2,53	4,38
feature_155	0,75	1	0,43	1	0	-1,14	-0,69
feature_156	0,16	0	0,37	1	0	1,82	1,31
feature_157	0,17	0	0,37	1	0	1,8	1,25
feature_158	0,11	0	0,31	1	0	2,49	4,2
feature_159	0,16	0	0,37	1	0	1,81	1,28
feature_160	0,18	0	0,38	1	0	1,71	0,92
feature_161	0,17	0	0,37	1	0	1,77	1,13
feature_162	0	0	0,02	1	0	59,08	3488,47
feature_163	0	0	0,03	1	0	31,87	1013,43
feature_164	0	0	0,06	1	0	15,76	246,42
feature_165	0,01	0	0,08	1	0	12,02	142,42
feature_166	0,01	0	0,11	1	0	9,01	79,19
feature_167	0,02	0	0,14	1	0	7,01	47,19
feature_168	0,03	0	0,17	1	0	5,52	28,51
feature_169	0,05	0	0,22	1	0	4,15	15,26
feature_170	0,09	0	0,29	1	0	2,88	6,28
feature_171	0,12	0	0,33	1	0	2,3	3,27
feature_172	0,12	0	0,33	1	0	2,32	3,38
feature_173	0,11	0	0,31	1	0	2,47	4,12
feature_174	0,1	0	0,3	1	0	2,65	5,04
feature_175	0,09	0	0,29	1	0	2,87	6,23
feature_176	0,08	0	0,27	1	0	3,08	7,5
feature_177	0,07	0	0,25	1	0	3,49	10,21
feature_178	0,05	0	0,21	1	0	4,23	15,85
feature_179	0,03	0	0,17	1	0	5,6	29,32
feature_180	0,01	0	0,11	1	0	8,85	76,3
feature_181	0	0	0,06	1	0	15,95	252,45
feature_182	0	0	0,04	1	0	27,39	747,98
feature_183	0	0	0,02	1	0	45,18	2038,8
feature_184	0	0	0,01	1	0	87,75	7698,54
feature_185	0,01	0	0,09	1	0	11	119
feature_186	0,01	0	0,09	1	0	11,1	121,24
feature_187	0,02	0	0,14	1	0	6,98	46,77

feature_188	0,03	0	0,18	1	0	5,11	24,16
feature_189	0,22	0	0,41	1	0	1,35	-0,17
feature_190	0	0	0,03	1	0	34,48	1187,22
feature_191	0,02	0	0,15	1	0	6,56	41
feature_192	0	0	0,03	1	0	28,52	811,41
feature_193	0	0	0,06	1	0	17,92	319,1
feature_194	0	0	0,04	1	0	23,44	547,28
feature_195	0,03	0	0,18	1	0	5,15	24,52
feature_196	0	0	0,06	1	0	17,72	311,92
feature_197	0,01	0	0,1	1	0	10,02	98,43
feature_198	0	0	0,06	1	0	17,07	289,55
feature_199	0	0	0,02	1	0	52,71	2776,83
feature_200	0,01	0	0,09	1	0	10,5	108,25
feature_201	0	0	0,03	1	0	28,6	815,87
feature_202	0	0	0,01	1	0	67,96	4617,12
feature_203	0	0	0,04	1	0	25,78	662,64
feature_204	0,01	0	0,1	1	0	9,49	87,97
feature_205	0	0	0,05	1	0	18,62	344,76
feature_206	0	0	0,04	1	0	22,67	512,11
feature_207	0	0	0,02	1	0	52	2701,65
feature_208	0	0	0,07	1	0	15,25	230,64
feature_209	0	0	0,01	1	0	111,87	12513,25
feature_210	0,01	0	0,1	1	0	9,45	87,34
feature_211	0	0	0,04	1	0	22,99	526,75
feature_212	0,02	0	0,15	1	0	6,43	39,29
feature_213	0	0	0,03	1	0	32,36	1045,48
feature_214	0,04	0	0,19	1	0	4,95	22,48
feature_215	0,01	0	0,09	1	0	10,73	113,15
feature_216	0	0	0,03	1	0	31,13	967,29
feature_217	0,05	0	0,23	1	0	3,93	13,45
feature_218	0,01	0	0,09	1	0	11,37	127,36
feature_219	0,01	0	0,08	1	0	11,78	136,66
feature_220	0	0	0,04	1	0	28,37	802,63
feature_221	0	0	0,02	1	0	59,43	3529,56
feature_222	0,01	0	0,08	1	0	12,88	163,79
feature_223	0,03	0	0,17	1	0	5,61	29,45
feature_224	0,01	0	0,1	1	0	9,57	89,67
feature_225	0,01	0	0,08	1	0	12,45	152,97
feature_226	0,13	0	0,33	1	0	2,26	3,12
feature_227	0,01	0	0,07	1	0	13,82	189,09
feature_228	0	0	0,04	1	0	23,1	531,5
feature_229	0	0	0,03	1	0	29,76	883,87
feature_230	0,01	0	0,08	1	0	12,81	162,1
feature_231	0,01	0	0,11	1	0	9,24	83,32
feature_232	0	0	0,01	1	0	68,49	4689,34
feature_233	0	0	0,01	1	0	79,1	6254,13
feature_234	0	0	0,05	1	0	22,11	486,72
feature_235	0,03	0	0,16	1	0	6,02	34,24

feature_236	0	0	0,03	1	0	38,91	1512,36
feature_237	0,01	0	0,08	1	0	11,68	134,49
feature_238	0	0	0,06	1	0	15,97	253,11
feature_239	0,02	0	0,13	1	0	7,34	51,93
feature_240	0	0	0,07	1	0	15,17	228,08
feature_241	0,18	0	0,38	1	0	1,66	0,75
feature_242	0,14	0	0,35	1	0	2,07	2,29
feature_243	0,14	0	0,35	1	0	2,08	2,32
feature_244	0,16	0	0,37	1	0	1,83	1,33
feature_245	0,28	0	0,45	1	0	0,99	-1,01
feature_246	0,3	0	0,46	1	0	0,87	-1,24
feature_247	0,23	0	0,42	1	0	1,28	-0,37
feature_248	0,21	0	0,41	1	0	1,45	0,1
feature_249	0,16	0	0,37	1	0	1,82	1,29
feature_250	0,11	0	0,31	1	0	2,49	4,22
feature_251	0,07	0	0,25	1	0	3,44	9,84
feature_252	0,16	0	0,37	1	0	1,86	1,46
feature_253	0,1	0	0,3	1	0	2,71	5,36
feature_254	0,11	0	0,32	1	0	2,45	4,01
feature_255	0,09	0	0,29	1	0	2,81	5,88
feature_256	0,02	0	0,13	1	0	7,38	52,42
feature_257	0,16	0	0,37	1	0	1,86	1,45
feature_258	0,1	0	0,3	1	0	2,72	5,37
feature_259	0,11	0	0,31	1	0	2,47	4,09
feature_260	0,07	0	0,26	1	0	3,33	9,06
feature_261	0,01	0	0,1	1	0	9,65	91,15
feature_262	0,21	0	0,41	1	0	1,4	-0,05
feature_263	0,21	0	0,41	1	0	1,42	0,03
feature_264	0,95	1	0,22	1	0	-4,16	15,33
feature_265	0,2	0	0,4	1	0	1,51	0,28
feature_266	0,16	0	0,37	1	0	1,82	1,31
feature_267	0,11	0	0,31	1	0	2,47	4,08
feature_268	0,25	0	0,43	1	0	1,16	-0,66
feature_269	0,22	0	0,41	1	0	1,35	-0,18
feature_270	0,29	0	0,45	1	0	0,94	-1,12
feature_271	0,19	0	0,39	1	0	1,58	0,5
feature_272	0,15	0	0,36	1	0	1,94	1,76
feature_273	0,18	0	0,38	1	0	1,7	0,91
feature_274	0,07	0	0,25	1	0	3,46	9,96
feature_275	0,06	0	0,24	1	0	3,59	10,86
feature_276	0,01	0	0,12	1	0	8,01	62,11
feature_277	0,03	0	0,18	1	0	5,16	24,67
feature_278	0,17	0	0,37	1	0	1,77	1,13
feature_279	0	0,12	1	1,58	-3,69	-0,81	0,26
feature_280	0	0,01	1	2,55	-9,27	-2,26	17,94
feature_281	0	0,38	1	1,65	-3,36	-0,79	-0,33
feature_282	0	0	1	2,84	-3,4	0	-0,38
feature_283	0,3	0	0,46	1	0	0,87	-1,24

feature_284	0,3	0	0,46	1	0	0,87	-1,24
feature_285	0,3	0	0,46	1	0	0,87	-1,24
feature_286	0,3	0	0,46	1	0	0,85	-1,27
feature_287	0,3	0	0,46	1	0	0,85	-1,27
feature_288	0,3	0	0,46	1	0	0,85	-1,27
feature_289	0,31	0	0,46	1	0	0,8	-1,36
feature_290	0,31	0	0,46	1	0	0,8	-1,36
feature_291	0,31	0	0,46	1	0	0,8	-1,36
feature_292	0,31	0	0,46	1	0	0,8	-1,36
feature_293	0,32	0	0,47	1	0	0,77	-1,4
feature_294	0,32	0	0,47	1	0	0,77	-1,4
feature_295	0,32	0	0,47	1	0	0,77	-1,4
feature_296	0,33	0	0,47	1	0	0,71	-1,5
feature_297	0,33	0	0,47	1	0	0,71	-1,5
feature_298	0,33	0	0,47	1	0	0,71	-1,5
feature_299	0,34	0	0,47	1	0	0,68	-1,53
feature_300	0,4	0	0,49	1	0	0,39	-1,85
feature_301	0,4	0	0,49	1	0	0,39	-1,85
feature_302	0,4	0	0,49	1	0	0,39	-1,85
feature_303	0,41	0	0,49	1	0	0,35	-1,87
feature_304	0,41	0	0,49	1	0	0,35	-1,87
feature_305	0,41	0	0,49	1	0	0,35	-1,87
feature_306	0,44	0	0,5	1	0	0,26	-1,93
feature_307	0,45	0	0,5	1	0	0,22	-1,95
feature_308	0,45	0	0,5	1	0	0,22	-1,95
feature_309	0,45	0	0,5	1	0	0,22	-1,95
feature_310	0,46	0	0,5	1	0	0,14	-1,98
feature_311	0,46	0	0,5	1	0	0,14	-1,98
feature_312	0,46	0	0,5	1	0	0,14	-1,98
feature_313	0,49	0	0,5	1	0	0,04	-2
feature_314	0,49	0	0,5	1	0	0,04	-2
feature_315	0,49	0	0,5	1	0	0,04	-2
feature_316	0,49	0	0,5	1	0	0,04	-2
feature_317	0,49	0	0,5	1	0	0,02	-2
feature_318	0,49	0	0,5	1	0	0,02	-2
feature_319	0,49	0	0,5	1	0	0,02	-2
feature_320	0,5	0	0,5	1	0	0,02	-2
feature_321	0,5	0	0,5	1	0	0,02	-2
feature_322	0,5	0	0,5	1	0	0,02	-2
feature_323	0,5	0	0,5	1	0	0,02	-2
feature_324	0,5	1	0,5	1	0	0	-2
feature_325	0,5	1	0,5	1	0	0	-2
feature_326	0,5	1	0,5	1	0	0	-2
feature_327	0,51	1	0,5	1	0	-0,04	-2
feature_328	0,51	1	0,5	1	0	-0,04	-2
feature_329	0,51	1	0,5	1	0	-0,04	-2
feature_330	0,52	1	0,5	1	0	-0,06	-2
feature_331	0,52	1	0,5	1	0	-0,1	-1,99

feature_332	0,69	1	0,46	1	0	-0,8	-1,36
feature_333	0,8	1	0,4	1	0	-1,52	0,3
feature_334	0,87	1	0,34	1	0	-2,14	2,56
feature_335	0,87	1	0,34	1	0	-2,14	2,56
feature_336	0,87	1	0,34	1	0	-2,14	2,56
feature_337	0,87	1	0,34	1	0	-2,14	2,56
feature_338	0,87	1	0,34	1	0	-2,14	2,56
feature_339	0,87	1	0,34	1	0	-2,14	2,56

Príloha B: Popis premenných

Atribút	Jednotka	Popis
feature_001	(1 - nie, 0 - áno)	Trvalé bydlisko je kontaktnou adresou (región)
feature_002	(1 - nie, 0 - áno)	Trvalé bydlisko je adresou do práce (región)
feature_003	(1 - nie, 0 - áno)	Kontaktná adresa je pracovnou adresou (región)
feature_004	(1 - nie, 0 - áno)	Trvalé bydlisko je kontaktnou adresou (mesto)
feature_005	(1 - nie, 0 - áno)	Trvalé bydlisko je adresou do práce (mesto)
feature_006	(1 - nie, 0 - áno)	Kontaktná adresa je pracovnou adresou (mesto)
feature_007	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_008	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_009	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_010	(1 - áno, 0 - nie)	Klient poskytol dokument 2
feature_011	(1 - áno, 0 - nie)	Klient poskytol dokument 3
feature_012	(1 - áno, 0 - nie)	Klient poskytol dokument 4
feature_013	(1 - áno, 0 - nie)	Klient poskytol dokument 5
feature_014	(1 - áno, 0 - nie)	Klient poskytol dokument 6
feature_015	(1 - áno, 0 - nie)	Klient poskytol dokument 7
feature_016	(1 - áno, 0 - nie)	Klient poskytol dokument 8
feature_017	(1 - áno, 0 - nie)	Klient poskytol dokument 9
feature_018	(1 - áno, 0 - nie)	Klient poskytol dokument 10
feature_019	(1 - áno, 0 - nie)	Klient poskytol dokument 11
feature_020	(1 - áno, 0 - nie)	Klient poskytol dokument 12
feature_021	(1 - áno, 0 - nie)	Klient poskytol dokument 13
feature_022	(1 - áno, 0 - nie)	Klient poskytol dokument 14
feature_023	(1 - áno, 0 - nie)	Klient poskytol dokument 15
feature_024	(1 - áno, 0 - nie)	Klient poskytol dokument 16
feature_025	(1 - áno, 0 - nie)	Klient poskytol dokument 17
feature_026	(1 - áno, 0 - nie)	Klient poskytol dokument 18
feature_027	(1 - áno, 0 - nie)	Klient poskytol dokument 19
feature_028	(1 - áno, 0 - nie)	Klient poskytol dokument 20

feature_029	(1 - áno, 0 - nie)	Klient poskytol dokument 21
feature_030	(1 - áno, 0 - nie)	Anomália v informácií o odpracovaných dňoch
feature_031	Numerická	Príjem žiadateľa
feature_032	Numerická	Výška úveru
feature_033	Numerická	Výška anuitnej splátky
feature_034	Numerická	Cena statku, ktorý chce klient obstarat' (pre spotrebný úver)
feature_035	Numerická	Normalizovaná populácia regiónu, v ktorom žije klient (vyššia hodnota znamená, že klient žije v obývanejšom regióne)
feature_036	Numerická	Vek klienta v rokoch
feature_037	Numerická	Dĺžka odpracovaného času v rokoch
feature_038	Numerická	Doba od vytvorenia konta v Home Credit v rokoch
feature_039	Numerická	Dĺžka vlastníctva občianskeho preukazu v rokoch
feature_040	Numerická	Vek auta
feature_041	Numerická	Externý rating, zdroj 1
feature_042	Numerická	Externý rating, zdroj 2
feature_043	Numerická	Externý rating, zdroj 3
feature_044	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_045	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_046	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_047	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_048	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_049	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_050	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_051	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_052	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_053	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_054	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je

[illegible]

[illegible]

		priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_085	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_086	Numerická	Normalizované informácie o budove, v ktorej klient žije. Informuje aká je priemerná (prípona _AVG), modálna (prípona _MODE), mediánová (prípona _MEDI) veľkosť bytu, spoločný priestor, obytná plocha, vek budovy, počet výťahov, počet vchodov, stav budovy, počet poschodí atď.
feature_087	Numerická	Klient mal vyplnený údaj o telefóne
feature_088	Numerická	Príjem v pomere k výške úveru
feature_089	Numerická	Príjem v pomere k výške anuity
feature_090	Numerická	Výška úveru v pomere k výške anuity
feature_091	Numerická	Príjem v pomere k veľkosti rodiny (počet členov)
feature_092	Numerická	Odpracované dni ako pomer z dĺžky života
feature_093	(1 - áno, 0 - nie)	Úver má charakter revolvingový
feature_094	(1-muž, 0-žena)	Pohlavie
feature_095	(1 - áno, 0 - nie)	Vlastní auto
feature_096	(1 - áno, 0 - nie)	Vlastní nehnuteľnosť
feature_097	(1 - áno, 0 - nie)	Počet detí 0
feature_098	(1 - áno, 0 - nie)	Počet detí 1
feature_099	(1 - áno, 0 - nie)	Počet detí 2
feature_100	(1 - áno, 0 - nie)	Počet detí 3
feature_101	(1 - áno, 0 - nie)	Ručiteľ - člen rodiny
feature_102	(1 - áno, 0 - nie)	Ručiteľ - skupina osôb
feature_103	(1 - áno, 0 - nie)	Ručiteľ - iný A
feature_104	(1 - áno, 0 - nie)	Ručiteľ - iný B
feature_105	(1 - áno, 0 - nie)	Ručiteľ - druh/družka
feature_106	(1 - áno, 0 - nie)	Ručiteľ - nezaručený úver
feature_107	(1 - áno, 0 - nie)	Typ príjmu - komerčný
feature_108	(1 - áno, 0 - nie)	Typ príjmu - materská/rodičovská
feature_109	(1 - áno, 0 - nie)	Typ príjmu - dôchodca
feature_110	(1 - áno, 0 - nie)	Typ príjmu - štátny úradník
feature_111	(1 - áno, 0 - nie)	Typ príjmu - študent
feature_112	(1 - áno, 0 - nie)	Typ príjmu - nezamestnaný
feature_113	(1 - áno, 0 - nie)	Typ príjmu - zamestnaný
feature_114	(1 - áno, 0 - nie)	Vzdelanie - nižšia stredná škola
feature_115	(1 - áno, 0 - nie)	Vzdelanie - stredná škola
feature_116	(1 - áno, 0 - nie)	Vzdelanie - vysoká škola
feature_117	(1 - áno, 0 - nie)	Rodinný stav - civilné manželstvo
feature_118	(1 - áno, 0 - nie)	Rodinný stav - manželstvo
feature_119	(1 - áno, 0 - nie)	Rodinný stav - rozvedený
feature_120	(1 - áno, 0 - nie)	Rodinný stav - sám
feature_121	(1 - áno, 0 - nie)	Bývanie - dom
feature_122	(1 - áno, 0 - nie)	Bývanie - byt
feature_123	(1 - áno, 0 - nie)	Bývanie - kancelárske priestory
feature_124	(1 - áno, 0 - nie)	Bývanie - podnájom
feature_125	(1 - áno, 0 - nie)	Bývanie - s rodičmi

feature_126	(1 - áno, 0 - nie)	Služobný telefón
feature_127	(1 - áno, 0 - nie)	Služobné číslo
feature_128	(1 - áno, 0 - nie)	Súkromný mobil
feature_129	(1 - áno, 0 - nie)	Klient vlastní telefón
feature_130	(1 - áno, 0 - nie)	Emailová adresa
feature_131	(1 - áno, 0 - nie)	Typ zamestnania - upratovacie služby
feature_132	(1 - áno, 0 - nie)	Typ zamestnania - stravovacie služby
feature_133	(1 - áno, 0 - nie)	Typ zamestnania - služby
feature_134	(1 - áno, 0 - nie)	Typ zamestnania - doprava
feature_135	(1 - áno, 0 - nie)	Typ zamestnania - ľudské zdroje
feature_136	(1 - áno, 0 - nie)	Typ zamestnania - technické služby
feature_137	(1 - áno, 0 - nie)	Typ zamestnania - IT sektor
feature_138	(1 - áno, 0 - nie)	Typ zamestnania - manuálna práca
feature_139	(1 - áno, 0 - nie)	Typ zamestnania - nízko kvalifikovaná práca
feature_140	(1 - áno, 0 - nie)	Typ zamestnania - manažment
feature_141	(1 - áno, 0 - nie)	Typ zamestnania - zdravotníctvo
feature_142	(1 - áno, 0 - nie)	Typ zamestnania - súkromný sektor služby
feature_143	(1 - áno, 0 - nie)	Typ zamestnania - reality
feature_144	(1 - áno, 0 - nie)	Typ zamestnania - predaj
feature_145	(1 - áno, 0 - nie)	Typ zamestnania - sekretariát
feature_146	(1 - áno, 0 - nie)	Typ zamestnania - bezpečnostné služby
feature_147	(1 - áno, 0 - nie)	Typ zamestnania - obsluha
feature_148	(1 - áno, 0 - nie)	Počet členov v domácnosti 1
feature_149	(1 - áno, 0 - nie)	Počet členov v domácnosti 2
feature_150	(1 - áno, 0 - nie)	Počet členov v domácnosti 3
feature_151	(1 - áno, 0 - nie)	Počet členov v domácnosti 4
feature_152	(1 - áno, 0 - nie)	Rating regiónu, kde klient býva 1
feature_153	(1 - áno, 0 - nie)	Rating regiónu, kde klient býva 2
feature_154	(1 - áno, 0 - nie)	Rating mesta, kde klient býva 1
feature_155	(1 - áno, 0 - nie)	Rating mesta, kde klient býva 2
feature_156	(1 - áno, 0 - nie)	Žiadosť vyplnená v piatok
feature_157	(1 - áno, 0 - nie)	Žiadosť vyplnená v pondelok
feature_158	(1 - áno, 0 - nie)	Žiadosť vyplnená v nedeľu
feature_159	(1 - áno, 0 - nie)	Žiadosť vyplnená v štvrtok
feature_160	(1 - áno, 0 - nie)	Žiadosť vyplnená v utorok
feature_161	(1 - áno, 0 - nie)	Žiadosť vyplnená v stredu
feature_162	(1 - áno, 0 - nie)	Žiadosť vyplnená o 1 hodine
feature_163	(1 - áno, 0 - nie)	Žiadosť vyplnená o 2 hodine
feature_164	(1 - áno, 0 - nie)	Žiadosť vyplnená o 3 hodine
feature_165	(1 - áno, 0 - nie)	Žiadosť vyplnená o 4 hodine
feature_166	(1 - áno, 0 - nie)	Žiadosť vyplnená o 5 hodine
feature_167	(1 - áno, 0 - nie)	Žiadosť vyplnená o 6 hodine
feature_168	(1 - áno, 0 - nie)	Žiadosť vyplnená o 7 hodine
feature_169	(1 - áno, 0 - nie)	Žiadosť vyplnená o 8 hodine
feature_170	(1 - áno, 0 - nie)	Žiadosť vyplnená o 9 hodine
feature_171	(1 - áno, 0 - nie)	Žiadosť vyplnená o 10 hodine
feature_172	(1 - áno, 0 - nie)	Žiadosť vyplnená o 11 hodine
feature_173	(1 - áno, 0 - nie)	Žiadosť vyplnená o 12 hodine

feature_174	(1 - áno, 0 - nie)	Žiadosť vyplnená o 13 hodine
feature_175	(1 - áno, 0 - nie)	Žiadosť vyplnená o 14 hodine
feature_176	(1 - áno, 0 - nie)	Žiadosť vyplnená o 15 hodine
feature_177	(1 - áno, 0 - nie)	Žiadosť vyplnená o 16 hodine
feature_178	(1 - áno, 0 - nie)	Žiadosť vyplnená o 17 hodine
feature_179	(1 - áno, 0 - nie)	Žiadosť vyplnená o 18 hodine
feature_180	(1 - áno, 0 - nie)	Žiadosť vyplnená o 19 hodine
feature_181	(1 - áno, 0 - nie)	Žiadosť vyplnená o 20 hodine
feature_182	(1 - áno, 0 - nie)	Žiadosť vyplnená o 21 hodine
feature_183	(1 - áno, 0 - nie)	Žiadosť vyplnená o 22 hodine
feature_184	(1 - áno, 0 - nie)	Žiadosť vyplnená o 23 hodine
feature_185	(1 - áno, 0 - nie)	Odvetvie - poľnohospodárstvo
feature_186	(1 - áno, 0 - nie)	Odvetvie - bankovníctvo
feature_187	(1 - áno, 0 - nie)	Odvetvie - biznis 1
feature_188	(1 - áno, 0 - nie)	Odvetvie - biznis 2
feature_189	(1 - áno, 0 - nie)	Odvetvie - biznis 3
feature_190	(1 - áno, 0 - nie)	Odvetvie - čistiace služby
feature_191	(1 - áno, 0 - nie)	Odvetvie - stavebníctvo
feature_192	(1 - áno, 0 - nie)	Odvetvie - kultúra
feature_193	(1 - áno, 0 - nie)	Odvetvie - elektrina
feature_194	(1 - áno, 0 - nie)	Odvetvie - záchranné služby
feature_195	(1 - áno, 0 - nie)	Odvetvie - štátna správa
feature_196	(1 - áno, 0 - nie)	Odvetvie - hotelierstvo
feature_197	(1 - áno, 0 - nie)	Odvetvie - poľnohospodárstvo
feature_198	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 1
feature_199	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 10
feature_200	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 11
feature_201	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 12
feature_202	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 13
feature_203	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 2
feature_204	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 3
feature_205	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 4
feature_206	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 5
feature_207	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 6
feature_208	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 7
feature_209	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 8
feature_210	(1 - áno, 0 - nie)	Odvetvie - priemysel typ 9
feature_211	(1 - áno, 0 - nie)	Odvetvie - poisťovníctvo
feature_212	(1 - áno, 0 - nie)	Odvetvie - škôlka
feature_213	(1 - áno, 0 - nie)	Odvetvie - právo
feature_214	(1 - áno, 0 - nie)	Odvetvie - zdravotníctvo
feature_215	(1 - áno, 0 - nie)	Odvetvie - ozbrojené sily
feature_216	(1 - áno, 0 - nie)	Odvetvie - mobilné služby
feature_217	(1 - áno, 0 - nie)	Odvetvie - iný
feature_218	(1 - áno, 0 - nie)	Odvetvie - polícia
feature_219	(1 - áno, 0 - nie)	Odvetvie - pošta
feature_220	(1 - áno, 0 - nie)	Odvetvie - nehnuteľnosti
feature_221	(1 - áno, 0 - nie)	Odvetvie - náboženstvo

feature_222	(1 - áno, 0 - nie)	Odvetvie - stravovanie
feature_223	(1 - áno, 0 - nie)	Odvetvie - školstvo
feature_224	(1 - áno, 0 - nie)	Odvetvie - bezpečnosť
feature_225	(1 - áno, 0 - nie)	Odvetvie - bezpečnosť štátu
feature_226	(1 - áno, 0 - nie)	Odvetvie - samo živiť
feature_227	(1 - áno, 0 - nie)	Odvetvie - služby
feature_228	(1 - áno, 0 - nie)	Odvetvie - telekomunikácie
feature_229	(1 - áno, 0 - nie)	Odvetvie - obchod typ 1
feature_230	(1 - áno, 0 - nie)	Odvetvie - obchod typ 2
feature_231	(1 - áno, 0 - nie)	Odvetvie - obchod typ 3
feature_232	(1 - áno, 0 - nie)	Odvetvie - obchod typ 4
feature_233	(1 - áno, 0 - nie)	Odvetvie - obchod typ 5
feature_234	(1 - áno, 0 - nie)	Odvetvie - obchod typ 6
feature_235	(1 - áno, 0 - nie)	Odvetvie - obchod typ 7
feature_236	(1 - áno, 0 - nie)	Odvetvie - doprava typ 1
feature_237	(1 - áno, 0 - nie)	Odvetvie - doprava typ 2
feature_238	(1 - áno, 0 - nie)	Odvetvie - doprava typ 3
feature_239	(1 - áno, 0 - nie)	Odvetvie - doprava typ 4
feature_240	(1 - áno, 0 - nie)	Odvetvie - vysoké školstvo
feature_241	(1 - áno, 0 - nie)	Odvetvie - žiadne
feature_242	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau hodinu pred podaním žiadosti
feature_243	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau deň pred podaním žiadosti
feature_244	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau týždeň pred podaním žiadosti
feature_245	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau mesiac pred podaním žiadosti
feature_246	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau štvrt'rok pred podaním žiadosti
feature_247	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau pred menej ako 1 rokom
feature_248	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau pred menej ako 2 rokmi
feature_249	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau pred menej ako 3 rokmi
feature_250	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau pred menej ako 4 rokmi
feature_251	(1 - áno, 0 - nie)	Žiadateľ bol vyhľadávaný v Credit Bureau pred viac ako 4 rokmi
feature_252	(1 - áno, 0 - nie)	Okolie klienta - úver so splátkou 30 dní po splatnosti (1 osoba)
feature_253	(1 - áno, 0 - nie)	Okolie klienta - úver so splátkou 30 dní po splatnosti (2 osoby)
feature_254	(1 - áno, 0 - nie)	Okolie klienta - úver so splátkou 30 dní po splatnosti (viac ako 3 osoby)
feature_255	(1 - áno, 0 - nie)	Okolie klienta - úver s default pri stave 30 dní po splatnosti (1 osoba)
feature_256	(1 - áno, 0 - nie)	Okolie klienta - úver s default pri stave 30 dní po splatnosti (2 osoby)
feature_257	(1 - áno, 0 - nie)	Okolie klienta - úver so splátkou 60 dní po splatnosti (1 osoba)
feature_258	(1 - áno, 0 - nie)	Okolie klienta - úver so splátkou 60 dní po splatnosti (2 osoby)
feature_259	(1 - áno, 0 - nie)	Okolie klienta - úver so splátkou 60 dní po splatnosti (viac ako 3 osoby)
feature_260	(1 - áno, 0 - nie)	Okolie klienta - úver s default pri stave 60 dní po splatnosti (1 osoba)
feature_261	(1 - áno, 0 - nie)	Okolie klienta - úver s default pri stave 60 dní po splatnosti (2 osoby)
feature_262	(1 - áno, 0 - nie)	Budova - panely
feature_263	(1 - áno, 0 - nie)	Budova - tehla, kameň
feature_264	(1 - áno, 0 - nie)	Klient s minulosťou v Home Credit
feature_265	(1 - áno, 0 - nie)	Klient odmietol v minulosti úver 2 a viac krát
feature_266	(1 - áno, 0 - nie)	Klient odmietol v minulosti úver 1 krát
feature_267	(1 - áno, 0 - nie)	Klient s nevyužitým úverom v minulosti
feature_268	(1 - áno, 0 - nie)	Klient v minulosti schválený 1 krát
feature_269	(1 - áno, 0 - nie)	Klient v minulosti schválený 2 krát

feature_270	(1 - áno, 0 - nie)	Klient v minulosti schválený 3 až 4 krát
feature_271	(1 - áno, 0 - nie)	Klient v minulosti schválený viac ako 5 krát
feature_272	(1 - áno, 0 - nie)	Klient v minulosti zamietnutý 1 krát
feature_273	(1 - áno, 0 - nie)	Klient v minulosti zamietnutý 2 a viac krát
feature_274	(1 - áno, 0 - nie)	Default klienta - úver 1
feature_275	(1 - áno, 0 - nie)	Default klienta - úver 2
feature_276	(1 - áno, 0 - nie)	Default klienta - kreditná karta 1
feature_277	(1 - áno, 0 - nie)	Default klienta - kreditná karta 2
feature_278	(1 - áno, 0 - nie)	Default klienta v minulosti - akýkoľvek produkt HC
feature_279	Numerická	Predchádzajúci úver - Priemer dni na splátku
feature_280	Numerická	Predchádzajúci úver - Súčet dni zaplattenia pred splatnosťou
feature_281	Numerická	Predchádzajúci úver - Minimálna suma splátky zaplattená klientom
feature_282	Numerická	Predchádzajúci úver - Počet splatených splátok
feature_283	(1 - nie, 0 - áno)	Klient mal vyplnený atribút COMMONAREA_MEDI
feature_284	(1 - nie, 0 - áno)	Klient mal vyplnený atribút COMMONAREA_AVG
feature_285	(1 - nie, 0 - áno)	Klient mal vyplnený atribút COMMONAREA_MODE
feature_286	(1 - nie, 0 - áno)	Klient mal vyplnený atribút NONLIVINGAPARTMENTS_AVG
feature_287	(1 - nie, 0 - áno)	Klient mal vyplnený atribút NONLIVINGAPARTMENTS_MODE
feature_288	(1 - nie, 0 - áno)	Klient mal vyplnený atribút NONLIVINGAPARTMENTS_MEDI
feature_289	(1 - nie, 0 - áno)	Klient mal vyplnený atribút FONDKAPREMONT_MODE
feature_290	(1 - nie, 0 - áno)	Klient mal vyplnený atribút LIVINGAPARTMENTS_AVG
feature_291	(1 - nie, 0 - áno)	Klient mal vyplnený atribút LIVINGAPARTMENTS_MEDI
feature_292	(1 - nie, 0 - áno)	Klient mal vyplnený atribút LIVINGAPARTMENTS_MODE
feature_293	(1 - nie, 0 - áno)	Klient mal vyplnený atribút FLOORSMIN_AVG
feature_294	(1 - nie, 0 - áno)	Klient mal vyplnený atribút FLOORSMIN_MEDI
feature_295	(1 - nie, 0 - áno)	Klient mal vyplnený atribút FLOORSMIN_MODE
feature_296	(1 - nie, 0 - áno)	Klient mal vyplnený atribút YEARS_BUILD_AVG
feature_297	(1 - nie, 0 - áno)	Klient mal vyplnený atribút YEARS_BUILD_MEDI
feature_298	(1 - nie, 0 - áno)	Klient mal vyplnený atribút YEARS_BUILD_MODE
feature_299	(1 - nie, 0 - áno)	Klient mal vyplnený atribút OWN_CAR_AGE
feature_300	(1 - nie, 0 - áno)	Klient mal vyplnený atribút LANDAREA_AVG
feature_301	(1 - nie, 0 - áno)	Klient mal vyplnený atribút LANDAREA_MEDI
feature_302	(1 - nie, 0 - áno)	Klient mal vyplnený atribút LANDAREA_MODE
feature_303	(1 - nie, 0 - áno)	Klient mal vyplnený atribút BASEMENTAREA_MODE
feature_304	(1 - nie, 0 - áno)	Klient mal vyplnený atribút BASEMENTAREA_AVG
feature_305	(1 - nie, 0 - áno)	Klient mal vyplnený atribút BASEMENTAREA_MEDI
feature_306	(1 - nie, 0 - áno)	Klient mal vyplnený atribút EXT_SOURCE_1
feature_307	(1 - nie, 0 - áno)	Klient mal vyplnený atribút NONLIVINGAREA_AVG
feature_308	(1 - nie, 0 - áno)	Klient mal vyplnený atribút NONLIVINGAREA_MODE
feature_309	(1 - nie, 0 - áno)	Klient mal vyplnený atribút NONLIVINGAREA_MEDI
feature_310	(1 - nie, 0 - áno)	Klient mal vyplnený atribút ELEVATORS_MODE
feature_311	(1 - nie, 0 - áno)	Klient mal vyplnený atribút ELEVATORS_MEDI
feature_312	(1 - nie, 0 - áno)	Klient mal vyplnený atribút ELEVATORS_AVG
feature_313	(1 - nie, 0 - áno)	Klient mal vyplnený atribút WALLSMATERIAL_MODE
feature_314	(1 - nie, 0 - áno)	Klient mal vyplnený atribút APARTMENTS_MEDI
feature_315	(1 - nie, 0 - áno)	Klient mal vyplnený atribút APARTMENTS_MODE
feature_316	(1 - nie, 0 - áno)	Klient mal vyplnený atribút APARTMENTS_AVG
feature_317	(1 - nie, 0 - áno)	Klient mal vyplnený atribút ENTRANCES_MODE

feature_318	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	ENTRANCES_AVG
feature_319	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	ENTRANCES_MEDI
feature_320	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	LIVINGAREA_MEDI
feature_321	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	LIVINGAREA_MODE
feature_322	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	LIVINGAREA_AVG
feature_323	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	HOUSETYPE_MODE
feature_324	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	FLOORSMAX_MODE
feature_325	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	FLOORSMAX_MEDI
feature_326	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	FLOORSMAX_AVG
feature_327	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	YEARS_BEGINEXPLUATATION_MEDI
feature_328	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	YEARS_BEGINEXPLUATATION_MODE
feature_329	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	YEARS_BEGINEXPLUATATION_AVG
feature_330	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	TOTALAREA_MODE
feature_331	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	EMERGENCYSTATE_MODE
feature_332	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	OCCUPATION_TYPE
feature_333	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	EXT_SOURCE_3
feature_334	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	AMT_REQ_CREDIT_BUREAU_YEAR
feature_335	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	AMT_REQ_CREDIT_BUREAU_QRT
feature_336	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	AMT_REQ_CREDIT_BUREAU_HOUR
feature_337	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	AMT_REQ_CREDIT_BUREAU_DAY
feature_338	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	AMT_REQ_CREDIT_BUREAU_WEEK
feature_339	(1 - nie, 0 - áno)	Klient mal vyplnený atribút	AMT_REQ_CREDIT_BUREAU_MON
