EKONOMICKÁ UNIVERZITA V BRATISLAVE
NÁRODOHOSPODÁRSKA FAKULTA

Evidenčné číslo: 101006/I/2020/36097107838920196

POROVNANIE NEURÓNOVEJ SIETE A LOGISTICKEJ REGRESIE AKO METÓD NA PREDIKCIU ZLYHANIA FIRIEM

Diplomová práca

2019 **Bc. Tatiana Šišková**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
NÁRODOHOSPODÁRSKA FAKULTA

POROVNANIE NEURÓNOVEJ SIETE A LOGISTICKEJ REGRESIE AKO METÓD NA PREDIKCIU ZLYHANIA FIRIEM

Diplomová práca

Študijný program: Medzinárodné financie

Študijný odbor: Ekonómia a manažment

Školiace pracovisko: Katedra financií

Vedúci záverečnej práce: Marek Káčer Mgr., PhD.

Bratislava 2019 **Bc. Tatiana Šišková**

ABSTRAKT

ŠIŠKOVÁ, Tatiana: *Porovnanie neurónovej siete a logistickej regresie ako metód na predikciu zlyhania firiem.* – Ekonomická univerzita v Bratislave. Národohospodárska fakulta; Katedra financií. – Vedúci záverečnej práce: Marek Káčer Mgr., PhD. – Bratislava: NHF EU, 2020, s. 68.

Cieľom záverečnej práce je porovnať metódu neurónovej siete a logistickej regresie pri predikcii zlyhania spoločností na základe zadaného súboru dát, ktorý obsahuje informácie z účtovnej závierky a výkazov ziskov a strát.

Práca je rozdelená do troch nosných kapitol. Obsahuje 13 grafov, 13 tabuliek a 10 obrázkov.

Prvá kapitola je venovaná teoretickému základu bankrotových modelov. Opisuje ich predikčnú schopnosť a determinanty, ktoré vplývajú na možné zlyhanie firiem. Okrem toho je v nej stručný prierez názormi, ktoré sa venujú predikcii zlyhania od dvadsiateho storočia až po súčasnosť.

V ďalšej časti sme charakterizovali cieľ, metodiku práce a metódy skúmania. V tejto kapitole je tiež definovaný teoretický základ logistickej regresie a neurónových sietí.

Záverečná kapitola sa zaoberá samotnou aplikáciou modelov na upravený súbor dát. Detailne popisuje evaluáciu modelov, úpravu dát do vhodnej podoby a samotnú aplikáciou logistickej regresie a neurónovej siete. Kapitola tiež obsahuje zhrnutie výsledkov, ich porovnanie a interpretáciu.

Výsledkom riešenia danej problematiky je určenie, ktorá metóda predikcie zlyhania bola úspešnejšia, dokázala sa naučiť dáta lepšie a v závere dosiahnuť vyššiu výkonnosť pri predpovedi bankrotu.

Kľúčové slová:

neurónová sieť, logistická regresia, predikcia zlyhania, Z skóre, AUC index, ROC krivka, Python

ABSTRACT

SISKOVA, Tatiana: *Comparison of neural network and logistic regression approaches to corporate failure prediction*. – University of Economics in Bratislava. Faculty of National Economy; Department of Finance. – Thesis supervisor: Marek Káčer Mgr., PhD. – Bratislava: NHF UE, 2020, p. 68.

The aim of this thesis is to compare the method of neural network and logistic regression in predicting the failure of companies based on a given set of data that contain information from the financial statements and profit and loss statements.

The thesis is divided into three main chapters. Contains 10 pictures, 13 charts, 13 tables.

The first chapter is primarily devoted to the theoretical basis of bankruptcy models. It describes their predictive ability and determinants that affect the possible failure of companies. In addition, it provides a brief cross-section of views on prediction of failure from the twentieth century to the present.

The next chapter characterizes the goal, methodology of research and research methods. This chapter also defines the theoretical basis of logistic regression and neural networks.

Last chapter deals with the application of models to the modified data set. It describes in detail the evaluation of models, adjustment of data into a suitable form and the application of logistic regression and neural network. The chapter also contains a summary of the results, their comparison and interpretation.

The result of solving the problem is to determine which method of failure prediction was more successful, was able to learn the data better and in the end to achieve higher performance in predicting bankruptcy

Key words:

Neural networks, logistic regression, failure prediction, Z – score, AUC index, ROC curve, Python

OBSAH

Zoznam ilustrácií a tabuliek

Zoznam grafov

Zoznam obrázkov

Zoznam tabuliek

Úvod

V súčasnej zložitej a nejasnej situácii sú mnohé spoločnosti pod tlakom a vzhľadom k pozastavenej výrobe či službám nedokážu uniesť náklady, ktoré sú spojené s ich chodom. Bankrotom si každoročne prejde mnoho väčších, či menších spoločností a tento zložitý stav dokáže ovplyvniť vzťahy medzi obchodnými partnermi. Predpovedanie budúcej finančnej situácie podnikov či ich zhodnotenie je tak bežnou súčasťou pri vstupe do obchodných vzťahov. Okrem toho poznatok o budúcej finančnej situácie spoločnosti dokáže vysielať varovný signál pri blížiacich sa finančných ťažkostiach, ktoré častokrát ústia do bankrotu. V tomto smere je pre podniky a ich manažment kľúčové, aby dokázali predikovať svoju budúcu finančnú situáciu.

Modely, ktoré sú schopné predpovedať zlyhanie firiem sú dôležitými nástrojmi pre bankárov, investorov, veriteľov, ratingové agentúry a najmä pre samotné spoločnosti. Prvé štúdie zaoberajúce sa predikciou bankrotu spoločností boli publikované začiatkom 20. storočia, pričom predikcia mala slúžiť ako signál včasného varovania v prípade, že bola pozorovaná zhoršujúca sa situácia podniku spojená s rizikom, ktorá nakoniec vyústila do bankrotu.

V začiatkoch využívania predikčných metód sa používal len jeden ukazovateľ, ktorý dokázal najlepšie reflektovať súčasnú finančnú situáciu podniku (Beaver, 1966). O pár rokov neskôr, súbežne s rozvojom matematicko – štatistických metód, sme pozorovali vznikajúce predikčné bankrotové modely, ktoré boli vytvorené v kombinácii s finančnými pomerovými ukazovateľmi a ďalšími premennými, ktoré dokázali zohľadniť finančnú situáciu spoločnosti v jednom čísle. Na základe tohto čísla bola potom spoločnosť klasifikovaná ako podnik bez alebo s rizikom bankrotu v sledovanom časovom horizonte.

Altman (1968) predstavil prvý viacrozmerný model, ktorý dokázal odhadnúť zlyhanie firiem pred viac než päťdesiatimi rokmi. Vo svojej štúdii použil na konštrukciu modelu Z-skóre, lineárnu diskriminačnú analýzu. Od tej doby sa tento model často používal a bol akceptovaný výskumníkmi v oblasti financovania, bankovníctva a úverového rizika. Jedným z dôvodov, prečo je Altmanov model pomerne rozšírený môže byť aj dôvod, že bol jediným svojho druhu. Autor ho aktívne propagoval, treba však podotknúť, že bez dobrých výsledkov, ktoré model dosiahol by jeho obľuba určite nepretrvávala tak dlho. Model Z-skóre zahŕňa hlavné dimenzie finančného zdravia spoločností a stal sa prototypom ďalších mnohých modelov úverového rizika a zlyhania.

Nasledujúce štúdie sa zameriavali na špecifikovanie alternatívnych definícií výsledkov. Peel a Peel (1989) používali viac logitový prístup pri modelovaní finančnej tiesne oproti uprednostňovanému a obvyklému binárnemu výsledku. Peel a Wilson (1996) odhadli model s viacerými logitmi, ktorý definuje „*problémové akvizície*" ako dôležitý výsledok bankrotu.

Modely, ktoré sa snažia predpovedať zlyhanie využívajú prevažne finančné ukazovatele avšak v priebehu rokov sa objavilo aj niekoľko štúdií, ktoré pracovali práve s nefinančnými premennými (Grunert a kol., 2005; Altman a kol., 2009). Tieto nefinančné premenné možno ďalej rozdeliť do dvoch kategórií: premenné jednotlivých spoločností a makroekonomické premenné.

Niekoľko modelov na predikciu zlyhania bolo skonštruovaných taktiež pre slovenské spoločnosti, ktoré sa postupne rozvíjali na základe výstupov z trhovej ekonomiky v postkomunistickom období. Len niekoľko z nich však zahŕňalo nefinančné premenné medzi determinanty zlyhania firiem (Fidrmuc a Hainz, 2010; Wilson a kol., 2016; Altman a kol., 2017). Žiadna zo štúdií navyše netestovala modely uvedené v období mimo sledovanej periódy.

V začiatkoch 21. storočia sa tiež zdôraznili výhody zahrnutia premenných ako sú vek, druh podnikania, priemyselný sektor v kombinácii s finančnými ukazovateľmi (Grunert a kol., 2004).

Cieľom predloženej diplomovej práce je poskytnúť komplexný pohľad na komparáciu logistickej regresie a neurónových sietí ako metód na predikciu zlyhania firiem. Budeme pritom pracovať so vzorkou slovenských spoločností v období rokov od 2015 – 2017. Práca pozostáva z troch hlavných častí:

Prvá časť diplomovej práce je venovaná priblíženiu súčasného stavu problematiky predikcie zlyhania firiem. Ozrejmíme si historické ale aj aktuálne názory, predikčnú schopnosť modelov či jednotlivé determinanty zlyhania.

Ciele, metodika a metódy skúmania sú súčasťou druhej časti práce. V nej si zadefinujeme jednotlivé čiastkové ciele, ktoré nám v konečnom dôsledku pomôžu dosiahnuť náš hlavný zámer. Okrem toho si vysvetlíme vývoj a teoretické základy logistickej regresie a neurónových sietí, z ktorých budeme čerpať v záverečnej časti predkladanej práce.

V tretej, praktickej časti budú jednotlivé metódy konkrétne, logistická regresia a neurónové siete, testované na základe upraveného súboru dát. Naším cieľom bude zistiť,

ktorá z týchto metód dokáže zlyhanie firmy predikovať s väčšou presnosťou. Na vytváranie modelov nám poslúžil programovací jazyk Python, prostredníctvom webovej aplikácie Jupyter Notebook, ktorá bola vytvorená na interaktívne výpočty v niekoľkých programovacích jazykoch. Na základe dosiahnutých výstupov budeme ďalej výkonnosť modelov testovať a porovnávať pomocou ROC krivky či rôznych metrík odvodených z matice zámen. V závere kapitoly budeme prezentovať a porovnávať výsledky práce, ktoré dosiahneme jednotlivými krokmi a v neposlednom rade uvedieme zdroje, z ktorých sme pri písaní diplomovej práce čerpali.

1 Súčasný stav skúmanej problematiky

„Foresight is not about predicting the future, it's about minimizing surprise“

Karl Schroeder, canadian science fiction author.

Výskum v oblasti predikcie zlyhania spoločností bol v posledných desaťročiach veľmi populárnym nie len u výskumníkov z akademickej ale aj odbornej praxi. Problém zlyhania spoločností v moderných ekonomikách má významné hospodárske a sociálne dôsledky a keďže zaručený spôsob ani spoľahlivá metóda na predpovedanie udalosti zlyhania ešte neboli nájdené, môžeme konštatovať, že záujem o túto oblasť bude naďalej vysoký.

Beaver (1966) bol medzi prvými, ktorí sa pokúsili predpovedať zlyhanie spoločností a jeho štúdia sa v tejto oblasti považuje za míľnik. Vytvoril jednorozmernú analýzu pomerových ukazovateľov, ktoré signalizovali nadchádzajúce problémy dlhší čas pred samotným zlyhaním. Beaverov prístup bol nerovnomerný v tom, že každý pomer sa hodnotil z hľadiska toho, ako sa dá sám použiť na predpovedanie zlyhania bez ohľadu na ostatné pomery. Jeho analýza ukázala, že pomer cash flow k dlhu je najúčinnejšou premennou na predpovedanie bankrotu prostredníctvom poskytovania štatisticky významných signálov už päť rokov pred skutočným zlyhaním podniku. Altman (1968) sa pokúsil vylepšiť Beaverovu štúdiu použitím multivariačnej lineárnej diskriminačnej analýzy, o ktorej sa však preukázalo, že trpí určitými obmedzeniami. Altman taktiež zistil, že jeho päť pomerov (X1 – X5) prekonalo Beaverov pomer cash flow k celkovému dlhu. Altmanov model kombinuje sedem účtovných premenných (obežné aktíva, krátkodobé záväzky, dlhodobé aktíva, nerozdelený zisk, zisk pred úrokmi a daňami, dlhodobé záväzky, účtovnú hodnotu vlastného imania a čistý predaj), čím sa získa jedinečné Z-skóre, ktoré zoskupuje firmy do ohrozených, sivých a bezpečných zón.

V mnohých výskumoch sa vedci rozhodli pokračovať v rozširovaní Altmanovho modelu v nádeji, že dosiahnu vyššiu presnosť klasifikácie. Medzi príklady týchto pokusov patrí napríklad:

1) *pridelenie predchádzajúcich pravdepodobnostných členských tried (Deakin, 1972);*
2) *použitie vhodnejšieho kvadratického klasifikátora (Altman, 1977);*
3) *využitie modelov založených na cash flow (Gentry, 1985);*
4) *použitie štvrťročných informácií o finančných výkazoch (Baldwin, 1992);*
5) *používanie aktuálnych informácií o nákladoch (Aly a kol., 1992; Watson, 1986).*

Ani jeden z týchto pokusov však nepriniesol štatisticky významnejšie výsledky ako Altmanova doterajšia práca a okrem toho, vo väčšine prípadov praktické uplatňovanie týchto modelov prinieslo mnoho ťažkostí vzhľadom k ich zložitosti.

V roku 1980 použil Ohlson (1980) na predpovedanie zlyhania logistickú regresiu (alebo logitovú analýzu), čo je metóda, ktorá sa vyhýba argumentovaným obmedzeniam techniky multivariačnej lineárnej diskriminačnej analýzy. Logit, spolu s analýzou Probit sa tiež nazývajú aj modely podmienenej pravdepodobnosti, pretože poskytujú podmienenú pravdepodobnosť pozorovania, ktoré patrí do určitej triedy vzhľadom na hodnoty nezávislých premenných pre toto pozorovanie. Ohlsonove výsledky sa však nezlepšili ani vo výsledkoch diskriminačnej analýzy, čo naznačovalo, že boli potrebné ďalšie vylepšenia tejto techniky. Medzi rozšírenia Ohlsonovej štúdie patrí okrem iného:

1) *preskúmanie vplyvu priemyslu v predikčných modeloch zlyhania (Platt a Platt, 1980);*
2) *vývoj empirických modelov, ktoré by rozlišovali medzi neúspešnými firmami a firmami vo finančnej núdzi (Gilber a kol., 1980);*
3) *vývoj odvetvovo špecifických modelov (Platt a kol. 1994);*
4) *použitie viacúrovňovej logitovej analýzy (Johnsen a Melicher, 1994);*
5) *vývoj predikčných modelov pre sektor malých firiem (Keasy a Watson, 1987).*

Podmienené pravdepodobnostné modely však používateľovi nedokázali ponúknuť nič viac ako akékoľvek iné dovtedy používané techniky (Keasy a Watson, 1991).

Vedci, zaoberajúci sa modelmi zlyhania sa nevzdávali a naďalej využívali rôzne klasifikačné techniky v nádeji, že objavia „dokonalý“ model. Najpopulárnejšie z týchto techník sú: rekurzívne delenie, analýza prežitia či neurónové siete. Štúdia Laitinena a Kankaanpaa (1999) napríklad empiricky skúmala, či sa výsledky vyplývajúce z použitia týchto alternatívnych metód navzájom od seba líšia. Ich výsledky naznačili, že nebola nájdená žiadna bezchybná metóda, aj keď presnosť predikcie zlyhania sa menila v závislosti od použitej predikčnej metódy. Táto štúdia využívala alternatívnu techniku k problému predpovedania zlyhania spoločnosti a nazývala sa *viacrozmerné škálovanie*. Viacrozmerné škálovanie ponúka intuitívne znázornenie štatistických výsledkov, čo umožňuje ľahko interpretovať výsledky bez hlbokého porozumenia základných štatistických princípov. Z dôvodu častej nestability výsledkov vyplývajúcich zo štandardných štatistických techník musí používateľ doplniť štatistiku vlastným úsudkom. To je však možné len pri viacrozmernom škálovaní a nie v iných prístupoch, ktoré sú pre analytikov oveľa menej transparentné.

Výskum mnohých vedcov dokázal, že modely predpovedí tiesne spoločností sú v zásade nestabilné, pretože koeficienty modelu sa budú líšiť v závislosti od stavu hospodárstva (Mensah, 1984) a preto sa zdôrazňovala potreba derivácie modelu (Neophytou a Mar Molinero, 2001).

Hoci presnosť klasifikácie týchto štúdií je značne vysoká, všetky založili svoje výsledky na viacnásobnej diskriminačnej analýze, ktorá sa opiera na predpokladoch, ktoré sa často porušujú. Okrem toho boli takmer všetky spomínané štúdie zasadené do rozvinutých ekonomík a nezohľadňovali popri finančných aj nefinančné faktory (Oduro a Asiedu, 2017).

Káčer, Ochotnický a Alexy (2019) dávajú do pozornosti aj nové výskumné línie v súvislosti s predikciou zlyhania a poukazujú na zavedenie nefinančných premenných do modelov. Medzi nefinančné ukazovatele zaraďujú veľkosť spoločnosti, sektor, v ktorom podniká, jeho financovanie či formu vlastníctva.

Hlavným predmetom výskumu Altmana, Sabata a Wilsona (2010) bolo vytvorenie bankrotových modelov osobitne pre malé stredné podniky (MSP) vo Veľkej Británii, v ktorých skúmali okrem finančných aj nefinančné premenné. Altman a kol. (2017) tiež vykonali rozsiahle medzinárodné overenie výkonnosti účtovníckeho modelu, Altmanovho Z´´-skóre, ktoré bolo upravené o predikciu finančného zdravia súkromných, výrobných a nevýrobných firiem a neobsahovalo tržby z celkových aktív, ktoré sa javili byť citlivé na zmeny v priemyselnom sektore. Ich vzorka zahŕňala okrem 28 európskych krajín aj tri krajiny mimo Európy.

Významné míľniky v štúdiách, ktoré sa zaoberajú metódami na predikciu zlyhania pozorujeme aj na Slovensku, kde postkomunistická transformácia a predovšetkým proces privatizácie vytvorili podnikový sektor s rozmanitou škálou vo vlastníckych štruktúrach.

Výkonnosť portfólia bankových úverov na MSP na Slovensku skúmali Fidrmuc a Hainz (2010) a podľa ich výsledkov boli zisky pred zdanením, hotovosť, podiel bankových úverov na celkových aktívach a bankové účty významným determinantom zlyhania úveru spomedzi finančných premenných.

Wilson, Ochotnický a Káčer (2016) použili na predpovedanie zlyhania vplyv privatizačnej pasce, post-transformačnej recesie a zahraničného vlastníctva na výkonnosť podniku. Za týmto účelom zostavili niekoľko štandardných modelov s použitím finančných a nefinančných informácií.

1.1 Predikčné bankrotové modely

Ako sme spomínali, prvé štúdie zameriavajúce sa na predikciu bankrotu podnikov boli publikované na začiatku 20. storočia. Tieto štúdie slúžili ako signál včasného varovania pri zhoršujúcej sa situácii spoločností. Modely zlyhania firiem pritom využívali najmä úverové (aplikačný a behaviorálny skóring; minimalizácia rizika pri maximalizácii zisku) a finančné inštitúcie (dodržiavajú dohody o kapitálovej primeranosti) či investori a manažéri (musia rátať s rizikom protistrany).

V mnohých prípadoch spoločnosti využívajú na vyhodnotenie aktuálnej finančnej situácie najmä pomoc externých ratingov a expertného prístupu, ktorý je založený na krátkej histórii, poprípade si tvoria vlastné štatistické modely. Cieľom je pritom odhadnúť pravdepodobnosť zlyhania spoločnosti na základe minulých údajov a skúseností.

Vývoju týchto modelov a predikčných schopností sa budeme venovať v ďalších častiach tejto kapitoly.

1.1.1 Predikčná schopnosť modelov

Úlohou a cieľom použitia modelov na predikciu bankrotu je najmä dokázať, s čo najväčšou presnosťou, určiť nastávajúcu finančnú situáciu firiem. Na základe modelov sa následne dokážu podniky klasifikovať medzi prosperujúce a medzi podniky, ktorým hrozí riziko bankrotu. V rámci tejto analýzy môžu nastať štyri situácie:

1. *model dokázal správne identifikovať zbankrotovaný podnik;*
2. *model dokázal správne identifikovať podnik, ktorý sa do bankrotu nedostal;*
3. *model nedokázal správne klasifikovať podnik, ktorý sa do bankrotu nedostal, ale označil ho za zbankrotovaný – nastala teda chyba prvého rádu;*
4. *model nedokázal správne klasifikovať podnik, ktorý sa do bankrotu dostal, ale označil ho ako nezbankrotovaný – nastala teda chyba druhého rádu.*

Tabuľka 1: Klasifikácie výsledkov predikcie modelov

	Predikcia – bankrot	**Predikcia – nenastal bankrot**
Skutočnosť - bankrot	Správny výsledok (TP)	Chyba prvého rádu (FN)
Skutočnosť – nenastal bankrot	Chyba druhého rádu (FP)	Správny výsledok (TN)

Vzťahy uvedené v Tabuľke 1 nám hovoria, že predikčná schopnosť modelu dokáže vyjadriť koľko z celkového počtu podnikov bol model schopný správne klasifikovať ako podniky, ktoré zbankrotovali, alebo naopak, u nich k bankrotu nedošlo. Následne dokážeme tento vzťah znázorniť pomocou rovnice:

$$P = (TP+TN)/(TP+TN+FP+FN) \quad (1.1)$$

Zároveň platí, že čím je model presnejší a jeho schopnosť predikovať zlyhanie je vyššia, tým nastáva aj vyššia pravdepodobnosť, že model dokáže klasifikovať analyzovaný podnik správne. Model však nedokáže vždy správne podnik klasifikovať a tak vznikajú problémy s výskytom chýb I. a II. rádu. Načrtnutú chybovosť si vyjadríme pomocou nasledujúci vzťahov:

$$\text{Chyba I. rádu} = FN/(FN+TN) \quad (1.2)$$

$$\text{Chyba II. Rádu} = FP/(FP+TP) \quad (1.3)$$

Možno teda povedať, že čím vyššia chybovosť I. a II. rádu, tým je predikčná schopnosť modelu nižšia a to so sebou prináša určité náklady. Altman (1982) dokázal vo svojej publikácii náklady spojené s I. aj II. rádom identifikovať, pričom priradil k nákladom chýb I. rádu:

- nevykonávanie nápravných krokov zo strany manažmentu, nakoľko si neuvedomujú vážnosť situácie;
- možná strata investícií pre investorov, pretože nedisponujú dostatkom informácií o problémoch podniku;
- prípadné vystavenie sankciám pre audítorov a spoločnosti, pričom môže dôjsť aj ku strate goodwillu.

Altman (1982) radí k nákladom chýb II. Rádu:

- pri klasifikácii prosperujúceho podniku ako podniku, ktorý má finančné problémy, môže dôjsť ku *„samo naplňajúcemu sa proroctvu“;*
- audítori môžu vložiť väčšie množstvo nákladov na podrobnejšie skúmanie finančných problémov preto, aby sa vyhli bankrotu.

Keď porovnávame náklady spojené s chybami I. a II. rádu je očividné, že chyby I. rádu sú spojené s vyššími nákladmi ako chyby II. rádu (Bellovary a kol.,2007; Kuo, 2003; Gepp a Kumar, 2008).

Schopnosť modelov predikovať zlyhanie a s tým aj súvisiaca pravdepodobnosť vzniku chýb I. a II. rádu sú predovšetkým podmienené premennými a typom metód, ktoré boli pri tvorbe týchto modelov použité.

1.2 Determinanty zlyhania spoločností

Determinanty zlyhania firiem, môžeme vo všeobecnosti deliť do niekoľkých kategórií. V našej práci sa bližšie pozrieme na finančné a nefinančné premenné, pričom

obe spomínané kategórie majú vo vyspelých ekonomikách dlhú históriu pri predikovaní zlyhania spoločností.

Medzi finančné premenné môžeme radiť napríklad mieru likvidity, ziskovosti či pákový efekt. Medzi nefinančnými premennými, ktoré zväčša ponúkajú dodatočné informácie o finančných ukazovateľoch nájdeme napríklad veľkosť firmy, jej vek či odvetvie, v ktorom podniká. Vytváranie záverov z týchto determinantov však nie je také jednoduché ako by sa na prvý pohľad mohlo zdať. Napríklad tie finančné sa počítajú na základe údajov z minulých finančných výkazov jednotlivých podnikov, ktoré nemusia poskytovať presný obraz o súčasnej poprípade budúcej situácii. Dôležité je tiež poznamenať, že zatiaľ čo informácie zo súvahy poskytujú obraz o hospodárení podniku až na konci účtovného obdobia, jednotlivé finančné premenné podliehajú výkyvom počas sledovaného obdobia a tak nie je vždy pravidlom, že hodnoty vykázané na konci roka reflektujú skutočnosť.

V praktickej časti našej diplomovej práci budeme pracovať najmä s determinantmi, ktoré načrtol vo svojich štúdiách Altman (1968) a postupne predstavil Z-skóre I., Z-skóre II. a Z-skóre III.:

Z-skóre I. je vhodné na posudzovanie firiem obchodovaných na burze a tvoria ho nasledujúce ukazovatele:

$$X1: \frac{Obežný\ majetok - Súčet\ krátkodobých\ záväzkov}{Majetok} \quad (1.3)$$

Ukazovateľ X1 je tvorený pomerom prevádzkového kapitálu ku celkovým aktívam. Tento pomer skúma krátkodobú likviditu spoločnosti z hľadiska jej aktív. Zložka ukazovateľov prevádzkového kapitálu je v skutočnosti rozdiel medzi obežnými aktívami a krátkodobými záväzkami. Podľa Altmana je premenná X1 jednou z najdôležitejších v jeho modeli, pretože meria likviditu firmy. V súčasnom výskume sa pozoruje obrovský rozdiel medzi bankrotujúcimi a naopak, funkčnými firmami. Prvé z nich majú výrazne klesajúci trend, zatiaľ čo podniky bez bankrotu sa zdajú byť vo všeobecnosti lineárne k trendom prevádzkového kapitálu. Iní vedci uvádzajú, že Altman vo svojom modeli pridal likvidite až príliš veľký význam, pretože koncept pôžičiek sa v niekoľkých posledných desaťročiach zmenil, pričom spoločnosti prikladajú menšiu dôležitosť vysokej úrovni likvidity a využívajú tak väčší pákový efekt.

$$X2: \frac{Nerozdelený\ zisk}{Celkové\ aktíva} \quad (1.4)$$

Tento pomer poskytuje informácie o kumulatívnej ziskovosti spoločnosti, pričom obsahuje aj informácie o veku spoločnosti. Mladá spoločnosť sa vyznačuje nízkym

pomerom výnosov k celkovým aktívam, pretože je pre ňu ťažké vytvoriť nerozdelený zisk. Tento pomer predstavuje aj spôsob, ako vložiť do modelu počítadlo pákového efektu. Ak je pomer nerozdeleného zisku spoločnosti k celkovým aktívam pomerne vysoký, možno to interpretovať tak, že spoločnosť na financovanie svojich činností využíva vysoký objem nerozdeleného zisku a nie dlh. Posledným súborom informácií vychádzajúcich z tohto pomeru je rastový potenciál spoločnosti. Doba nerozdeleného zisku sa vzťahuje na sumu nerozdelených dividend. Toto množstvo peňazí sa zadržiava od akcionárov s cieľom financovať najmä nové investície v mene spoločnosti. Mnoho vedcov však kritizovalo tento pomer, pretože podľa nich relatívne mladá firma nie je schopná vytvárať nerozdelené zisky. Altman navyše tvrdí, že prvé tri roky existencie firmy sú najkritickejšie a mnohým firmám sa v tomto období nepodarí prežiť.

$$X3: \frac{EBIT}{Celkové\ aktíva} \tag{1.5}$$

Účelom tohto pomeru je vypočítať skutočnú produktivitu aktív spoločnosti bez ohľadu na daň alebo úroveň pákového efektu. Tento pomer môže navyše poskytnúť informácie o riadení spoločnosti a o tom, či ľudia riadiaci spoločnosť môžu kombinovať a efektívne využívať aktíva spoločnosti s cieľom dosiahnuť maximálny výsledok. Vedenie každej spoločnosti sa v podstate snaží maximalizovať EBIT spoločnosti danými aktívami. Opäť platí, že čím nižšia je hodnota, tým vyššia je očakávaná tendencia k zlyhaniu.

$$X4: \frac{Trhová\ hodnota\ vlastného\ imania}{Cudzí\ kapitál} \tag{1.6}$$

Pri výpočte trhovej hodnoty vlastného imania sa počet bežných aj prémiových akcií vynásobí trhovou cenou akcií. Účtovná hodnota celkového dlhu zahŕňa dlhodobé a krátkodobé pôžičky. Trend kapitalizácie podnikov, ktoré sú v konkurze ale aj mimo neho je klesajúci. Rozsah tohto úpadku v prípade bankrotujúcich firiem je však oveľa väčší ako v prípade tých, ktoré v konkurze nie sú. Tento prudký pokles kapitalizácie bankrotujúcich firiem môže byť indikátorom ich finančnej nestability. Výhodou tohto pomeru je, že využíva perspektívu trhu.

$$X5: \frac{Tržby}{Celkové\ aktíva} \tag{1.7}$$

Tento pomer sa používa veľmi často a odhaduje schopnosť firmy generovať tržby. Jeho význam je veľmi veľký, pretože interaguje so zložkami iných premenných modelu a v konečnom dôsledku môže poskytnúť informácie o schopnosti manažmentu vyrovnať sa s konkurenciou. Výpočet tejto premennej je pomerne jednoduchý. Podľa Altmana je však tento pomer najmenej významný na individuálnej úrovni v porovnaní s ostatnými štyrmi pomermi vzorky a nie je ani štatisticky významný.

Rovnica, ktorú predstavil vyzerala nasledujúco:

$$Z = 1,2 * X1 + 1,4 * X2 + 3,3 * X3 + 0,6 * X4 + 0,999 * X5 \quad (1.8)$$

Vyhodnocovala sa ako:

$Z > 2,99$ – finančná situácia je dobrá

$1,81 < Z < 2,99$ – firma je v sivej zóne

$Z < 1,81$ – finančná situácia je kritická

Altmanovo Z-skóre II. (1983) sa líšilo predovšetkým v zmene ukazovateľa *X4* = $\frac{Vlastné\ imanie}{Cudzí\ kapitál}$ a posudzovali sa ním firmy, ktoré neboli obchodované na burze a teda mali štatút súkromnej spoločnosti. Práve s týmto skóre budeme pracovať aj v praktickej časti, nakoľko súbor dát je vykázaný takýmto typom spoločností.

Rovnica mala tvar:

$$Z' = 0,717 * X1 + 0,847 * X2 + 3,107 * X3 + 0,420 * X4 + 0,998 * X5, \quad (1.9)$$

a interpretovala sa podľa daných kritérií:

$Z' > 2,9$ – finančná situácia je dobrá

$1,2 < Z' < 2,9$ – firma je v sivej zóne

$Z' < 1,2$ – finančná situácia je kritická

V rovnakom roku predstavil Altman svoje Z-skóre III., ktoré sa od predchádzajúcich dvoch líšilo v tom, že z neho autor vynechal ukazovateľ X5, ktorý sa ukázal byť najcitlivejší v medzi odvetvových špecifikách. Následná rovnica bola v tvare:

$$Z'' = 6,56 * X1 + 3,26 * X2 + 6,72 * X3 + 1,05 * X4, \quad (1.10)$$

pričom sa vyhodnocovala nasledovne:

$Z'' > 2,60$ – finančná situácia je dobrá

$1,10 < Z'' < 2,60$ – firma je v sivej zóne

$Z'' < 1,10$ – finančná situácia je kritická

1.2.1 Ďalšie potencionálne prediktory zlyhania

Existuje mnoho pohľadov a otázok prečo firmy zlyhávajú a zanikajú. Obdobie rastu sa strieda s obdobím poklesu, ktorý často vedie práve ku krachu. Túto skutočnosť môžeme pozorovať zo strany finančných premenných, keďže spoločnosť si nedokáže plniť svoje záväzky čo reflektuje finančnú udalosť. Taktiež však veľký vplyv na tieto spoločnosti môžu mať iné a to nefinančné premenné, ktoré môžeme rozlišovať podľa toho, že sú jedinečné pre každú spoločnosť (právna forma, vek, veľkosť firmy či sektor, v ktorom pôsobí) či makroekonomické premenné, ktoré sa odvíjajú od ekonomickej situácie v štáte, kde spoločnosť pôsobí (úverová a hospodárska situácia, ekonomický cyklus , inflácia)

1.2.2 Nefinančné premenné

V tejto podkapitole si uvedieme pár premenných, ktoré patri do kategórie nefinančných premenných a pôsobia nie len na ziskovosť ale taktiež aj na zlyhanie firiem:

- *Právna forma*: je pri predikcii zlyhania dôležitá vzhľadom na stupeň zodpovednosti spoločníkov a líši sa od právnej formy tej ktorej spoločnosti. V štúdii, ktorú rozpracovali Fidrmuc a Hainz (2010) sa ukázalo, že napríklad pri akciových spoločnostiach môžeme nájsť protichodné stimuly akcionárov a vedenia spoločnosti, ktoré nabáda manažment k vyšším rizikám. Na druhej strane pri rodinných firmách či spoločnostiach s ručením obmedzeným sú si vedenie a vlastníci bližší a problémy, ktoré by v konečnom dôsledku voviedli spoločnosť do krachu nie sú také výrazné.
- *Vek:* podľa Káčera, Ochotnického a Alexyho (2019) môžeme sledovať pri vplyve veku na neúspech podnikania dve protichodné tendencie. Na jednej strane môžeme povedať, že spoločnosť akumuluje skúsenosti a zisk v čase, a teda čím je staršia, tým viac skúseností a zisku hromadí, a preto klesá jej tendencia zlyhávať. Z iného hľadiska sa vyžaduje určitý čas kým spoločnosť začne zlyhávať nakoľko mladá spoločnosť má spravidla počiatočný kapitál, z ktorého dokáže fungovať aj keď nevytvára dostatočné príjmy. Ak skombinujeme tieto dve tendencie, miera zlyhania by mala byť v prvých rokoch pomerne nízka, čo zodpovedá času kedy je spoločnosť podporená štartovacím kapitálom. Po vyčerpaní tohto kapitálu je spoločnosť stále v procese získavania skúseností, a preto by toto obdobie malo byť spojené s najvyššou mierou zlyhania. Spoločnosti, ktoré tieto fázy ustáli sa stanú odolnými a miera neúspechov tak klesá.
- *Priemyselný sektor:* sektor, v ktorom daná spoločnosť pôsobí je jedným z kľúčových determinantov, nakoľko vývoj sektora odráža skutočnosť či daná spoločnosť bude alebo nebude generovať zisk vzhľadom k rastúcemu, poprípade znižujúcemu sa dopytu po jej službách či výrobkoch. Priemyselné ukazovatele použili vo svojej štúdii Fidrmuc a Hainz (2010); Wilson, Ochotnický a Káčer (2016) alebo Altman a kol. (2017). Altman a kol. (2017) tiež upozorňujú, že pomer obratu aktív (premenná X5) je obzvlášť citlivý na rozdiely medzi priemyselnými odvetviami.

1.2.3 Makroekonomické premenné

Medzi nefinančnými premennými nájdeme aj makroekonomické ukazovatele, ktoré v rámci ekonomiky ovplyvňujú vývoj danej krajiny a tiež subjekty, ktoré sa v nej nachádzajú. Vo veľkej miere tu môžeme badať vplyv inflácie, úrokovej miery, hrubého domáceho produktu či hospodárskych očakávaní.

Du Jardin (2008) vo svojej štúdii predkladá ďalší pohľad na rozdelenie premenných. Na základe zistení popísaných v jeho práci analyzoval 190 štúdií uskutočnených v oblasti modelovania predikcie spoločností. V nadväznosti na to vytvoril a následne usporiadal 5 typov premenných od najviac až po najmenej používané pri konštrukcii predikčných modeloch bankrotu.

Tabuľka 2: Premenné používané v modeloch (Du Jardin, 2008)

Premenná	Percentuálny podiel v 190 skúmaných štúdiách
Finančné pomerové ukazovatele	93%
Štatistické premenné (odchýlka, stredná hodnota a logaritmus vypočítané z pomerových ukazovateľov alebo finančných premenných)	28%
Vývojové premenné (vývoj v čase)	14%
Nefinančné premenné (charakteristika spoločnosti alebo jej okolia, ktoré sa nevzťahujú k finančnej situácii podniku)	13%
Trhové premenné (premenná alebo podiel vzťahujúci sa k cene akcií, ich výnosnosti a podobne)	6%
Finančné trhové premenné (údaje z finančných uzávierok a iných finančných dokumentov)	5%

1.3 Vývoj predikčných metód

Na to aké modely sa v jednotlivých štúdiách využívali malo za následok jednoznačne obdobie, do ktorého boli tieto štúdie zasadené. Pre jednotlivé obdobia sú tiež charakteristické aj metódy na základe ktorých boli predikčné modely zostrojené. Chronologicky teda môžeme modely rozdeliť vzhľadom na spôsob ich tvorby do troch hlavných období:

- Obdobie prvé (do roku 1966):

- ✓ Jednorozmerná diskriminačná analýza

- Obdobie druhé (1966 – 1990):
  - ✓ Viacrozmerná diskriminačná analýza
  - ✓ Logit a probit analýza
- Obdobie tretie (od 1990 – súčasnosť):n
  - ✓ Rozhodovacie stromy
  - ✓ Neurónové siete
  - ✓ Ďalšie metódy umelej inteligencie

V našej diplomovej práci využijeme na predikciu zlyhania metódy z druhého a tretieho obdobia, konkrétne logit a teda logistickú regresiu a neurónové siete. Predtým ako sa pustíme do modelovania si však v nasledujúcej kapitole priblížime tieto dve metódy na teoretickej úrovni.

2 Cieľ a metodika práce, metódy skúmania

Hlavným cieľom a účelom predkladanej diplomovej práce je komparácia dvoch metód, konkrétne logistickej regresie a neurónových sietí pri predikcii úpadku spoločností, ktoré podnikajú v podmienkach Slovenskej republiky. Základom pre zostrojenie modelov spomenutých vyššie sú pomerové determinanty X1 – X5 zostrojené pre účely Z-skóre II. Altmanom (1983), ktoré sme na základe jeho štúdií vypočítali z jednotlivých súvah a výkazov ziskov a strát. Našou snahou je poukázať na to, ktorý z testovaných modelov bude schopný dosiahnuť čo najvyššiu úspešnosť v predikcii zlyhania firiem a zároveň čo najmenšiu chybovosť. Naša práca by mala poskytnúť pohľad na komplexný postup od procesu spracovania dát až po validáciu modelu, poukázať na rozdiely v jednotlivých metódach pri predikcii zlyhania a vyvodiť záver, ktorý model bol pri predikcii úspešnejší.

V súlade so zadaným cieľom našej diplomovej práce, sme si stanovili nasledujúce parciálne ciele:

- poukázať na teoretické východiská jednotlivých modelov a objasniť informácie, z ktorých budeme následne čerpať v praktickej časti;
- pripraviť vierohodný súbor dát, na ktorý môžeme aplikovať metódu logistickej regresie a neurónových sietí;
- porovnať dosiahnuté výsledky, vyhodnotiť úspešnosť predikcie jednotlivých metód.

2.1 Metodika práce a metódy skúmania

V podkapitole metodiky práce a metód skúmania si priblížime teoretické základy logistickej regresie a neurónových sietí, teda metód, ktoré následne budeme aplikovať na súbor dát.

2.2 Logistická regresia

Logistická regresia bola navrhnutá v 60. rokoch minulého storočia ako alternatívny prístup k metóde najmenších štvorcov pre prípad, že vysvetľovaná (závislá) premenná je binárna. V minulosti sa väčšina úloh aplikácie logistickej regresie sústredila do oblasti medicíny a epidemiológie. Vysvetľovaná premenná v nich vystupovala buď ako prítomnosť alebo ako neprítomnosť choroby.

Logistická regresia sa od lineárnej líši v tom, že predikuje pravdepodobnosť či daná udalosť nastala alebo nie. Vypočítaná pravdepodobnosť je potom rovná 0 alebo 1. Aby sa vytvorila táto podmienka, používa logistická regresia takzvanú logitovú transformáciu,

ktorá nadväzuje na sigmoidálny vzťah medzi závislou premennou *y* a vektorom nezávisle premenných *x*. Pri veľmi nízkych hodnotách nezávislej premennej sa pravdepodobnosť y blíži k nule, zatiaľ čo pri vysokých hodnotách nezávislej premennej sa blíži ku číslu jedna. Rozdiel medzi logistickou a lineárnou regresiou spočíva v tom, že logistická regresia používa kategorickú, závislú premennú, zatiaľ čo lineárna regresia používa spojitú, závislú premennú. Centrálnu úlohu v tomto prípade zohráva logitová transformácia, ktorá vychádza z takzvaného pomeru šancí či nádejí. Na základe typu vysvetľujúcej premennej rozlišujeme:

a) *Binárnu logistickú regresiu* – ktorá sa týka binárnej závislej premennej, či znaku, ktorý nadobúda len dve možné hodnoty ako je napríklad prítomnosť a neprítomnosť, muž a žena. Vektor vysvetľujúcich nezávislých premenných či znakov môže obsahovať jednu alebo viac premenných či znakov a to spojitých, nazývaných prediktory, alebo kategorických, ktoré sa nazývajú faktory.
b) *Ordinálnu logistickú regresiu* – ktorá sa týka ordinálnej závislej premennej alebo znaku, ktorý môže nadobúdať tri a viac možných stavov prirodzeného charakteru, napríklad narastajúcej sily ako je silný nesúhlas, nesúhlas, súhlas, silný súhlas. Vektor vysvetľujúcich nezávislých premenných alebo znakov môže obsahovať prediktory no zároveň aj faktory.
c) *Nominálnu logistickú regresiu* – ktorá sa týka nominálnej závislej premennej, alebo znaku o viac ako troch úrovniach rôznych stavov, medzi ktorými je definovaná len odlišnosť. Vektor vysvetľujúci nezávislé premenné môže obsahovať prediktory a taktiež aj faktory.

2.2.1 Logistický regresný model

Pri logistickej regresii je dôležité vedieť či sa daná udalosť stala alebo nie. Sledujeme tu preto dichotomickú hodnotu závislej premennej, z ktorej sa odhaduje pravdepodobnosť či sa udalosť udiala alebo neudiala. Ak je predikovaná pravdepodobnosť väčšia ako 0.50, tak sa udiala a naopak, ak je pravdepodobnosť menšia ako 0.50, môžeme dedukovať, že udalosť sa neuskutočnila. Názov logistická regresia je odvodený od logitovej transformácie závislej premennej. Ak je táto transformácia použitá, logistická regresia a jej koeficienty môžu nadobúdať iný význam než v regresii metrickej premennej. Postup, ktorý vyčísľuje logistické koeficienty porovnáva pravdepodobnosť uskutočnenej udalosti $L_{(1)}$ voči pravdepodobnosti udalosti, ktorá sa neuskutočnila $L_{(0)}$ = 1 - $L_{(1)}$

a využíva pritom pravdepodobnostný model $L_{(1)}/L_{(0)}$, v ktorom je pravdepodobnosť $L_{(1)}$ vyjadrená logistickou funkciou:

$$L_{(1)} = \frac{1}{1+e^{c-z}}. \quad (2.1)$$

Pravdepodobnostný model, ktorý sa tiež nazýva pomer výhliadok (šancí), je možné vyjadriť nasledujúcim vzťahom:

$$\frac{L_{(1)}}{L_{(0)}} = \exp(a_0 + a_1x_1 + a_2x_2 + \ldots + a_px_p), \quad (2.2)$$

kde odhadované koeficienty a_0, a_1, a_2, ..., a_p predstavujú zmeny pomeru pravdepodobností, ktorý je lineárnou funkciou diskriminačnej funkcie o p pôvodných nezávisle premenných:

$$Z = a_0 + a_1x_1 + a_2x_2 + \ldots + a_px_p. \quad (2.3)$$

Po dosadení a zlogaritmovaní výrazu (2.1) a následnej úprave dostaneme:

$$C\text{-}Z = ln\left(\frac{L_{(1)}}{L_{(0)}}\right). \quad (2.4)$$

V súlade s klasifikačnými postupmi je $L_{(0)}=P(G=0|x)$ a $L_{(1)} = P(G=1|x)=1=P(G=0|x)$. Po úpravách rovnice (2.1) vyjde:

$$ln\left(\frac{L_{(1)}}{L_{(0)}}\right) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p, \quad (2.5)$$

kde $b_{(0)} = -C + a_0$, $b_i = a_i$ kde $i=1,\ldots,p$. Môžeme teda sledovať patrnú podobnosť s viacnásobnou lineárnou regresnou funkciou. Vyčíslenie pomeru pravdepodobnosti je bežne využívanou technikou práve na jej vysvetlenie. Túto skutočnosť môžeme sledovať aj pri tipovaní kde napríklad hovoríme, že daný tým má šancu na výhru 3:1, čo v praxi znamená, že favorizovaný tým má pravdepodobnosť víťazstva $\frac{3}{3+1} = \frac{3}{4} = 0.75$.

Platí teda pravdepodobnostný pomer $\frac{L_{(1)}}{L_{(0)}} = \frac{0,75}{1-0,75} = \frac{3}{1}$.

2.2.2 Odhady parametrov

Pre odhad parametrov logistických modelov sa štandardne používa metóda maximálnej vierohodnosti. Pre všeobecný model sa predpokladá multinomické rozdelenie a pre štandardný model s dvoma triedami binomické rozdelenie. Prítomnosť v prvej triede y=1 je pre G=1 a neprítomnosť v prvej triede y=0 je pre G=2 (prítomnosť v druhej triede). Východiskové dáta sú vo forme vektoru y, rozmeru n x 1 a matice X, ktorá má rozmer n x m. Pre i-tý objekt má y_i hodnotu buď 0 alebo 1 a $x_i{}^T$ je i-tý riadok matice X. Označme p(x,b) = p_1(x,b) a 1 – p(x,b) = p_2 (x,b). Následne za predpokladu binomického rozdelenia y môžeme zapísať logaritmus vierohodnosti funkcie v tvare:

$$Ln\,L(b) = \sum_{i=1}^{n}\{y_i\ ln\,p(x_i,b) + (1 - y_i)\,ln(1 - p(x_i,b)$$

$$= \sum_{i=1}^{n}\{y_i\,b^T x_i - ln(1 + exp(b^T x_i))\}. \qquad (2.6)$$

Kde $b^T = \{b_0, b_1\}$ a predpokladá sa, že prvý stĺpec matice X obsahuje len jednotky (absolútny člen). Aby sme dosiahli maximalizáciu ln L(b) využijeme fakt, že prvé derivácie sú v maxime rovné nule:

$$J = \frac{d \ln L(b)}{d\ b} = \sum_{i=1}^{n} x_i\ (y_i - p(x_i, b)) = 0. \qquad (2.7)$$

Ide o sústavu m+1 nelineárnych rovníc vzhľadom ku b.

2.2.3 Interpretácia regresných koeficientov

Základný predpoklad logistickej regresie hovorí, že logaritmus pravdepodobnostného pomeru $ln(L_{(1)}/L_{(0)})$ je lineárnou funkciou nezávislých premenných. Žiadne predpoklady o rozdelení nezávislých premenných x tu neexistujú. Medzi hlavné výhody tejto metódy patrí fakt, že pod x môžu byť rozoznávané ako diskrétne (faktory) tak aj spojité veličiny (prediktory). Logistický model zapisujeme v tvare (2.5) a nazývame ho viacnásobný logistický regresný model. Koeficienty b_i môžu byť interpretované ako regresné koeficienty, ktoré sú vyjadrené v logaritmoch, z čoho vyplýva, že ich bude nutné pretransformovať späť, pretože následne bude možné vysvetliť ich relatívny vplyv na pravdepodobnosť jednoduchšie. Pre logaritmus pravdepodobnostného modelu $ln(L_{(1)}/L_{(0)})$ sa používa termín logit, poprípade logitová transformácia pravdepodobnosti. Viacnásobný logistický regresný model je možné upraviť pri dosadení za $L_{(1)}$ = (1-$L_{(0)}$) do ekvivalentného tvaru:

$$L_{(0)} = \frac{1}{1+exp[-(b_0+b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)]}\,. \qquad (2.8)$$

Kladné znamienko koeficientu b_i zvyšuje pravdepodobnosť $L_{(1)}$ a záporné znamienko túto pravdepodobnosť naopak znižuje. V prípade, že b_i je kladné, funkcia exp je väčšia než 1 a pravdepodobnostný model $(L_{(1)}/L_{(0)})$ sa bude zvyšovať. Zvýšenie môžeme pozorovať vtedy keď sa predikovaná pravdepodobnosť odohranej udalosti $L_{(1)}$ zvýši a predikovaná pravdepodobnosť neodohranej udalosti $L_{(0)}$ sa zníži. Preto má model vyššiu predikovanú pravdepodobnosť odohranej udalosti $L_{(1)}$. Podobne ak je koeficient b_i záporný, je funkcia exp menšia než 1 a pravdepodobnosť sa zníži. Pre koeficient, ktorý sa rovná nule, smeruje funkcia exp(0) k hodnote 1, takže k žiadnej zmene pravdepodobnosti.

Môžeme pozorovať, že $L_{(1)}$ sa mení od 0 do 1, pričom pravdepodobnostný pomer $L_{(1)}/L_{(0)}$ sa mení od 0 po ∞. Ak je $L_{(1)}$ rovné 0.5 je pomer $L_{(1)}/L_{(0)}$ rovný 1. Na stupnici pravdepodobnostného pomeru zodpovedajú hodnoty od 0 do 1 hodnotám pravdepodobnosti $L_{(1)}$ od 0 do 0.5. Na druhej strane hodnoty pravdepodobnosti $L_{(1)}$ od 0.5 po 1 vedú k pomeru $L_{(1)}/L_{(0)}$ od 1 po ∞. Po zlogaritmovaní pomeru $L_{(1)}/L_{(0)}$ bude táto asymetria odstránená pretože pre $L_{(0)}$= 0 je ln $(L_{(1)}/L_{(0)})$ = ∞, pre $L_{(1)}$ = 0.5 je ln $(L_{(1)}/L_{(0)})$ = 0, pre $L_{(1)}$ = 1.0 je ln $(L_{(1)}/L_{(0)}) =$ - ∞.

2.2.4 Test významnosti regresných koeficientov

Metóda logistickej regresie umožňuje overiť hypotézu, že regresný koeficient v logistickom regresnom modeli sa líši od nuly. V tomto prípade nula naznačuje, že pravdepodobnostný model $L_{(1)}/L_{(0)}$ sa nemení a tým pádom pravdepodobnosť nie je ovplyvnená. Test významnosti každého regresného koeficientu je analogický podobný ako pri lineárnej regresii a preto môžeme k testovaniu štatistickej významnosti jednotlivých regresných koeficientov použiť Studentov t-test. Pre veľké výbery sa pri logistickej regresii používa Waldovo testovacie kritérium $W_{a,i}$=$(b_i/\mathrm{s}(b_i))^2$, ktoré vyčísľujú štatistickú významnosť nulovej hypotézy pre jednotlivé odhady regresných koeficientov rovnako ako pri viacnásobnej regresii.

Waldova štatistika $W_{a,i}$ má x^2 rozdelenie s 1 stupňom voľnosti a predstavuje tak štvorec pomerov odhadu regresného koeficientu a jeho smerodajnej odchýlky. $W_{a,i}$ má pre kategorizované premenné o 1 stupeň voľnosti menej než je počet kategórií. Waldova štatistika W_a má však jednu nežiaducu vlastnosť. Keď je absolútna hodnota regresného koeficientu b_i a odhad jeho smerodajnej odchýlky s(b_i) veľký, je výsledkom príliš malá hodnota testovacieho kritéria $W_{a,i}$, ktorá vedie k zlyhaniu zamietnutia nulovej hypotézy, že regresný koeficient je nulový. Preto, ak je regresný koeficient veľký, nemali by sme Waldovo kritérium používať. Namiesto toho by sme sa mali obrátiť na logistický model s touto premennou a bez tejto premennej a vyšetriť zmenu rovníc v logitovom kritériu (Meloun a kol., 2012).

2.3 **Neurónové siete**

Neurónová sieť predstavuje matematický model, ktorý sa využíva pri modelovaní správania určitých systémov, ktorých opis pomocou analytických metód častokrát nie je možný, poprípade neposkytuje dostatočné výsledky. Umelá neurónová sieť netvorí jednu

sieť, ale ich rozmanitú skupinu. Celková dosiahnutá funkčnosť je určená typológiou siete, charakteristikami jednotlivých neurónov a stratégiou učenia sa alebo tréningu (Schalkoff, 1997).

Umelé neurónové siete sa vyznačujú niekoľkými kľúčovými črtami, ako napríklad to, že ich vznik bol inšpirovaný biológiou (konkrétne centrálnou nervovou sústavou), majú schopnosť adaptácie či fakt, že využívajú výpočtový výkon.

Ľudský mozog pozostáva z približne 100 miliárd nervových buniek alebo neurónov, ktoré medzi sebou komunikujú prostredníctvom elektrických signálov v podobe krátkodobých impulzov v napätí bunkovej steny alebo membrány. Interneurónové spojenia sú sprostredkované elektrochemickými spojeniami nazývanými synapsie, ktoré sú umiestnené na vetvách bunky označovaných ako dendrity. Každý neurón prijíma niekoľko tisíc spojení od iných neurónov a tak doň prúdi veľké množstvo prichádzajúcich signálov, ktoré sa nakoniec dostávajú do tela bunky. Tu sú nejakým spôsobom integrované alebo zosumarizované a môžno povedať, že ak výsledný signál prekročí určitú prahovú hodnotu, neurón bude „strieľať“ alebo generovať napäťový impulz v reakcii. Toto sa potom prenáša na ďalšie neuróny prostredníctvom vetviaceho vlákna známeho ako axón.

Proces ako celok predstavuje školiaci algoritmus. Čo sa však stane, ak po tréningu neurónovej siete predstavíme danej sieti vzor, ktorý predtým nevidela? Ak sa sieť naučila základnú štruktúru problémovej domény, mala by správne klasifikovať novú a nepoznanú štruktúru. Ak však sieť takúto vlastnosť nemá, je pre ďalšie testovanie len málo praktická. Dobré zovšeobecnenie je preto jednou z kľúčových vlastností neurónových sietí (Gurney, 1997).

Neurónové siete sú jedným z mnohých modelov strojového učenia. Sú obľúbeným a výkonným nástrojom. Ich popularita sa nepripisuje iba vynikajúcim výsledkom, ale aj ich použiteľnosti a určitému zvýšeniu výpočtovej sily. Tradičné algoritmy strojového učenia ako rozhodovacie stromy a lineárna regresia vyžadujú veľmi starostlivú extrakciu prvkov a výber funkcií. Pre efektívne použitie je potrebné manuálne extrahovať z údajov informácie čo robí tento proces časovo náročným a vyžaduje si vysokú úroveň znalosti domény. Neurónové siete však vyžadujú pomerne nízku manipuláciu so vstupnými údajmi, pretože sú schopné extrahovať funkcie, ktoré považujú za užitočné, a interne ignorujú nadbytočné a zbytočné informácie.

Neurónové siete sa dokážu naučiť vytvárať reprezentáciu vstupných údajov a sú zložené z viacerých vrstiev. Tieto vrstvy umožňujú sieti vytvárať reprezentácie vstupných údajov na vysokej úrovni (abstraktnejšie) s každou pridanou vrstvou. Napríklad v prostredí

počítačového videnia sa prvá vrstva učí rozoznávať adges a krivky, druhá vrstva rozpoznáva geometrické tvary a posledná vrstva je schopná rozlíšiť, či je osoba muž alebo žena. Toto je nadmerné zjednodušenie skutočných vnútorných fungovaní siete, ale myšlienka zostáva rovnaká (Wang a Zheng, 2015).

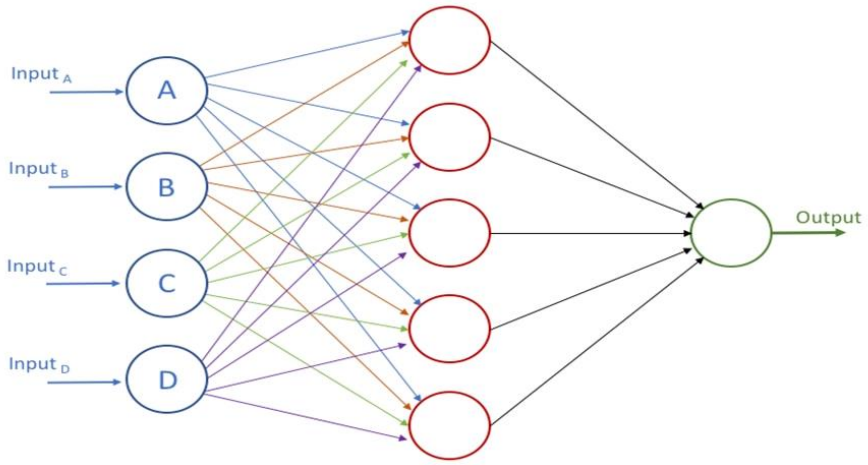
Neurónové siete zvyčajne využívame pri modelovaní zložitých vstupno/výstupných systémov, pri vyhľadávaní objektov s podobnými znakmi či ako aproximátory funkcií.

Ak sa na typickú neurónovú sieť pozrieme ako na model, môžeme konštatovať, že pozostáva z troch vrstiev:

- vstupnej vrstvy (angl. Input Layer);
- skrytej vrstvy (angl. Hidden Layer);
- výstupnej vrstvy (angl. Outpur Layer).

Jednotlivé vrstvy sú pritom poprepájané spojeniami, respektíve synapsiami, ktoré sú hodnotené váhami w_{ij}, w'_{ji}. Takto zadefinovaná neurónová sieť dokáže následne aproximovať funkciu f:F(x) →F(y). Odpoveď na otázku koľko neurónov sa musí v skrytej vrstve nachádzať, poprípade koľko skrytých vrstiev musí daná neurónová sieť obsahovať, aby sa funkčné zobrazenie dostatočne aproximovalo sa v mnohých prípadoch určuje experimentálne. Za dopredanú sieť môžeme označiť sieť, v ktorej jednotlivé synapsie smerujú iba medzi vrstvami a v každom prípade tak, že vychádzajú smerom od vstupných neurónov k výstupným a teda v nej nie sú žiadne synapsie, ktoré smerujú do predchádzajúcej vrstvy neurónov, poprípade v rámci jednej vrstvy. Okrem opísaných dopredaných neurónových sietí existujú taktiež aj rekurentné neurónové siete, kohenové siete, samo organizujúce sa mapy, RBF siete a mnohé ďalšie.

Obrázok 1: Jednoduchý model neurónovej siete (Tierney, 2018)

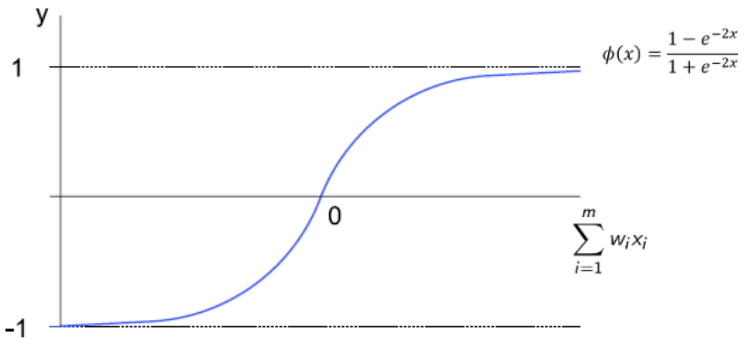


Pri dopredaných neurónových sieťach sa výsledná hodnota vektora u'_k, určuje ako u'_k=K($\sum_i w_{ik}\, u_j$(x)) analogicky pre u_j=k($\sum_i w_{ij}\, x_i$), kde je K takzvanou aktivačnou funkciou, najčastejšie zvaná sigmoida. Aktivačná funkcia je dôležitým prvkom v procese excitácie alebo inhibície umelého neurónu. Táto funkcia upravuje výstup na hodnoty, ktoré sa ďalej šíria v neurónovej sieti. Funkcie sa na základe článkov delia do dvoch základných tried, konkrétne lineárnych a nelineárnych, pričom najviac používanými sú nelineárne aktivačné funkcie. Voľba vhodnej aktivačnej funkcie v konečnom dôsledku rozhoduje o kvalite celej neurónovej siete, pričom o jej výbere rozhoduje úloha, ktorú daná neurónová sieť rieši. Pri klasifikačných úlohách je napríklad vhodná funkcia Sigmoid, pomocou ktorej sa dokážu pozorované vzorky rozdeliť do dvoch tried.

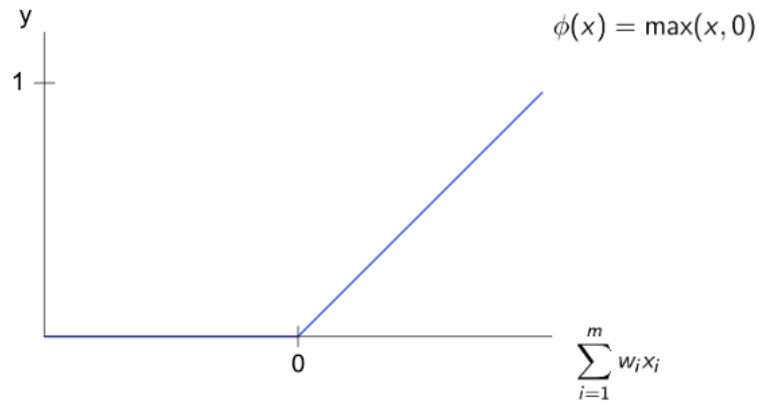
Ako príklad aktivačnej funkcie si môžeme uviesť aj ELU (angl. Exponential Linear Unit), ktorú budeme v konečnom dôsledku využívať pri modelovaní a v ktorej sa sa parameter α používa na reguláciu úrovne saturácie pre záporné vstupy a vopred stanovenú hodnotu. Funguje tak, že vráti 0, ak obdrží zápornú vstupnú hodnotu. Naopak, ak obdrží kladnú hodnotu, je táto hodnota výstupom funkcie. ELU nevráti záporný vstup na nulu, nakoľko tak pozitívne, ako aj záporné hodnoty sa môžu navzájom ovplyvňovať vo výslednej matici. Približná nulová stredná aktivácia jednotky nielen urýchľuje učenie s rýchlejšou konvergenciou, ale tiež zvyšuje odolnosť systému voči šumu. Aj keď má ELU vyššiu časovú zložitosť ako iné aktivačné funkcie v dôsledku exponenciálneho výpočtu, je častokrát preferovaná, nakoľko dokáže dosiahnuť lepší výkon domény (Clevert a kol., 2015). Taktiež, napríklad Sigmoid patrí medzi často využívané aktivačné funkcie. Medzi jej nevýhody však patrí fakt, že pre výstupné hodnoty blízko 0 alebo 1, je hodnota

derivácie veľmi malá a tým sa učenie spomaľuje. Rozličné príklady aktivačných funkcií si možno všimnúť nižšie:

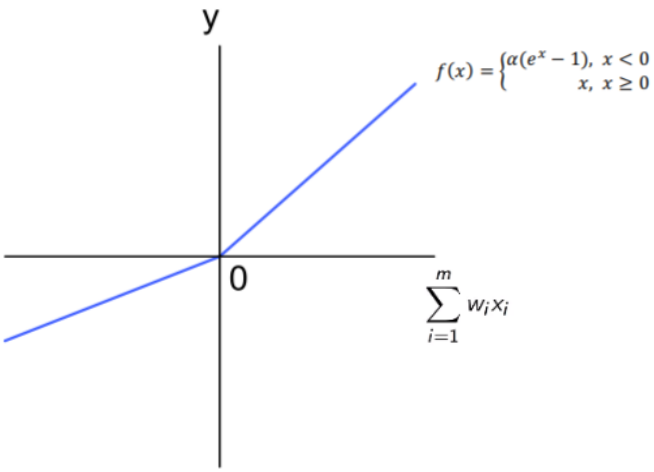
Graf 1: Hyperbolický tangens (Zhang a Woodland, 2015)



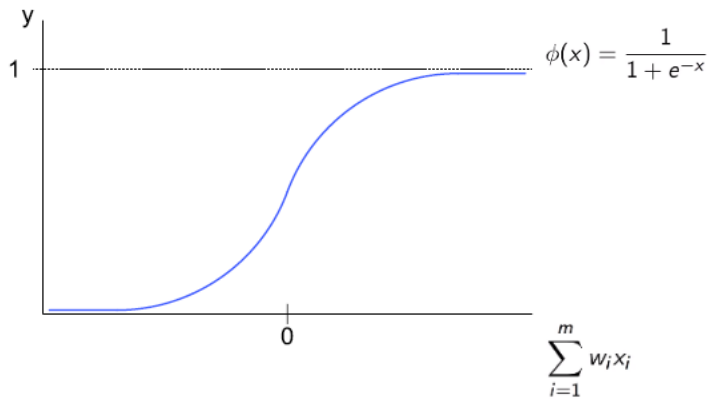
Graf 2: ReLU funkcia (Zhang a Woodland, 2015)



Graf 3: ELU funkcia (Clevert a kol., 2015)



Graf 4: Sigmoida (Zhang a Woodland, 2015)



2.3.1 Učenie neurónovej siete

To, čo danú modelovaciu techniku robí najzaujímavejšou je jej schopnosť učiť sa. Práve pri učení sa využíva funkcia C, ktorá vyhodnocuje kvalitu aproximácie dopredanej neurónovej siete, kde zároveň platí, že hodnota funkcie C je pre optimálne nastavenie váh neurónovej siete a pre dané vstupné dáta minimálna. Môžeme konštatovať, že funkcia C nám vyjadruje ako veľmi je momentálne nastavenie váh neurónovej siete vzdialené od optimálneho. Pre použitie pri N pozorovaniach vypočítame C ako:

$$C=\frac{1}{N}\sum_{1}^{N}(f(x_t)- y_t)^2. \qquad (2.9)$$

Pri supervízovom učení máme k dispozícii vstupnú množinu X a rovnako aj výstupnú množinu Y. Pomocou algoritmu spätného šírenia chyby (angl. Back Propagation Algorithm) postupne neurónovú sieť trénujeme (jednotlivé váhy medzi synapsiami), vhodným nastavením povolenej chyby a to tak, aby pri vstupných dátach, ktoré nepatria do tréningovej množiny X, neurónová sieť dokázala generovať čo najlepšie výsledky na dátach z toho istého systému, ktoré však neboli použité na trénovanie (množina K). Pri takomto druhu učenia, je dôležité zvoliť správnu typológiu neurónovej siete a taktiež aj požadovanú chybu pri učení, aby sa nestalo, že sa daná sieť preučí a spôsobí, že pri teste na vstupných hodnotách z množiny K dostaneme výsledky, ktoré sú neuspokojivé. Najčastejšie sa takáto forma učenia používa pri aproximácii funkcií.

Nesupervízované učenie je charakteristické tým, že dáta zo vstupnej množiny sú v neurónovej sieti reprezentované tak, že ich táto neurónová sieť kategorizuje do skupín. Mechanizmus, ktorý je skrytý za touto formou učenia sa nazýva súťaž. Za víťaza je označený neurón s najväčším výstupom a váhy smerujúce od neurónu k nemu sú inkrementované. Tento proces sa cyklicky opakuje a na jeho konci sú vstupné dáta z jednotlivých množín dát asociované s jednotlivými neurónmi. Táto forma učenia sa

pritom najčastejšie využíva pri vyhľadávaní a rozpoznávaní spoločných znakov, ako napríklad rozpoznávanie hlasu, či textu (Rojas, 2013).

2.3.2 *Výcvik neurónových sietí*

Prístupy k výcviku modelov strojového učenia sa dajú rozdeliť do štyroch rôznych kategórií v závislosti od použitých údajov a techník (Schmidhuber, 2015).

Neurónové siete môžu byť použité ktorýmkoľvek z týchto prístupov:

• *Dozorované učenie* si vyžaduje anotované údaje. Model trénujeme ukazovaním vstupov a porovnaním výstupu so základnou pravdou (štítky alebo anotácie). Potom modelu povieme, aké zlé boli jeho predpovede a podľa toho upravujeme jeho parametre.

• *Učenie bez dozoru* nevyžaduje žiadne anotované údaje. Model musí objaviť štruktúru údajov a musí byť schopný kategorizovať nové vstupy na základe nájdenej štruktúry. Napríklad učenie bez dozoru sa používa na zoskupovanie algoritmov.

• *Vzdelávanie pod dohľadom* je možné vnímať ako strednú cestu medzi spomínanými prístupmi. Jeho cieľom je zvýšiť množstvo dostupných údajov bez potreby ručného označovania, čo vyžadovalo súbor označených a neznačených údajov. Označená časť sa používa na trénovanie modelu pod dohľadom a na dosiahnutie východiskového bodu. Potom sa neoznačená časť použije na ďalšie školenie modelu. Napríklad vyškolený model sa môže použiť na označenie neoznačených vstupov, novo označené údaje sa potom môžu použiť na prípravu druhého modelu s väčším počtom dostupných údajov.

• *Výučba zosilnenia* je založená na pokusoch a omyloch. Nevyžaduje sa označený súbor údajov. Namiesto toho sa vyžaduje jasne definovaný cieľ a súbor opatrení, ktoré môže model vykonať. Model je odmenený za dobré činy (činy, ktoré ho približujú k cieľu) a trestaný za zlé činy. Veľkosť a frekvencia odmien sú stanovené ľubovoľne.

Výcvik neurónových sietí sa vykonáva pomocou algoritmu nazývaného gradientový zostup alebo jeho adaptácie. Je to optimalizačný algoritmus používaný na minimalizáciu funkcie. Táto funkcia sa nazýva stratová funkcia. Existuje veľa rôznych funkcií strát a všetky kvantifikujú chybu siete (nesprávne predpovede, akcie).

2.3.3 *Využitie neurónových sietí*

Použitie neurónových sietí je veľmi rozmanité no vo väčšine prípadov ich však využívame práve na:

- aproximáciu funkcií, regresnú analýzu, predikciu časových radov;
- klasifikáciu, rozpoznávanie spoločných znakov;

- spracovávanie dát;
- kvantitatívne obchodovanie, obchodovanie s vysokou frekvenciou.

Za nevýhody ich používania sa považujú najmä nároky na výpočtovú techniku a neschopnosť sa prispôsobiť všetkým zmenám, ktoré sa na trhu dejú (Moravčík, 2010).

3 Výsledky práce

Tretia časť našej diplomovej práce bude zameraná na praktickú aplikáciu logistickej regresie a neurónových sietí. Cieľom a úlohou bude na poskytnutom súbore dát, ktorý obsahuje položky z aktív, pasív a výkazu ziskov a strát firiem, otestovať sledované metódy a zistiť, ktorá z nich dokáže predikovať zlyhanie spoločnosti s väčšou presnosťou.

V prvom kroku opíšeme spôsob evaluácie jednotlivých modelov a teda, na základe akých parametrov budeme porovnávať ich výkonnosť. Ďalej si priblížime prípravu dát pred vstupom do neurónovej siete a logistickej regresie, následne sa zameriame na jednotlivé metódy a v záverečnom kroku vyhodnotíme ich predikčnú schopnosť a porovnáme výsledky z dosiahnutých výstupov.

Na vytváranie modelov nám poslúžil programovací jazyk Python, prostredníctvom webovej aplikácie Jupyter Notebook, ktorá bola vytvorená na interaktívne výpočty v niekoľkých programovacích jazykoch. Knižnice, ktoré v našom kóde použili boli: Pandas, NumPy, Matplotlib, Seaborn, Statsmodels a Keras.

3.1 Modely a ich evaluácia

Pre potreby porovnania výstupov z jednotlivých modelov a zisťovania ich výkonnosti sme sa rozhodli využiť ukazovatele, v rámci ktorých sme boli schopní s istotou rozhodnúť, ktorá metóda bola v našich podmienkach úspešnejšia pri predikcii zlyhania. Môžeme teda podotknúť, že aplikácia daných ukazovateľov nám pomohla vybrať najviac vyhovujúcu metódu. Vychádzali sme zo štúdií D. Powersa (2011) a rozhodli sme sa využiť metriky, ktoré sú bežne používané pri úlohách binárnej klasifikácie:

1. klasifikačná presnosť;
2. chybovosť;
3. senzitivita;
4. špecifickosť;
5. AUC – plocha pod ROC krivkou.

Pre definíciu metrík 1 – 4 použijeme poznatky získané z matice zámen (angl. Confusion Matrix) a pri poslednej menovanej metrike, AUC ploche, budeme vychádzať z ROC (angl. Receiver Operating Characteristic Curve) krivky. Skôr ako si však vysvetlíme výpočet a význam jednotlivých metrík zadefinujeme si pojem matica zámen.

3.1.1 *Matica zámen*

Okrem správne klasifikovaných objektov narazíme pri testovaní klasifikačných modelov aj ku objektom, ktoré sú klasifikované naopak, a teda nesprávne. Ak by sme chceli vysvetliť maticu zámen na základe našej situácie, mali by sme pri klasifikácii dve možnosti kam môže firma patriť; na jednej strane máme spoločnosti, ktoré skrachovali (označíme ich C_1) a na strane druhej spoločnosti, ktoré dané obdobie prežili bez zlyhania (budeme ich označovať ako C_0).

V prípade, že klasifikátor zaradí objekt správne, môžeme povedať, že objekt je, závisiac od jeho skutočnej príslušnosti do triedy, buď pravdivo negatívny (angl. True Negative - TN) alebo pravdivo pozitívny (angl. True Positive - TP). Podobne je to aj v prípade, ak klasifikátor zaradí objekt nesprávne, ten je v danom prípade buď falošne pozitívny (angl. False Positive - FP) alebo falošne negatívny (angl. False Negative - FN). Na základe toho ako klasifikátor objekty zaradí, môžeme povedať, že matica zámen pri binárnej klasifikácii ukazuje rozdelenie všetkých sledovaných objektov v testovacej množine do spomínaných štyroch skupín a to TP, TN, FP, FN.

Obrázok 2: Vzor matice zámen

		Predpovedané 0	Predpovedané 1
Skutočné	0	TN	FP
	1	FN	TP

Pri matici zámen nesmieme zabúdať na fakt, že sa mení vzhľadom k zvolenej prahovej hodnote (angl. Threshold). Určenie pravdepodobnosti $p(C_k|\chi)$ je súčasťou modelu a znamená, že objekt χ patrí do triedy C_k. V prípade, že prahová hodnota dosiahne hodnotu θ, objekt χ bude do triedy C_k zaradený práve vtedy, keď bude nadobúdať pravdepodobnosť $p(C_k|\chi) > \theta$, v opačnom prípade bude zamietnutý. Ak by sme však mali k dispozícii K tried a prahovú hodnotu by sme nastavili tak, že $\theta = 1/K$, žiadny z objektov by nezostal zamietnutý. Klasifikátor by zamietol všetky objekty len v prípade, ak by sme určili, že $\theta = 1$.

3.1.2 *Definícia metrík*

Ako sme načrtli v kapitole vyššie, nasledujúce riadky nám priblížia spomínané metriky, ktoré nám poslúžia k zmeraniu výkonnosti jednotlivých modelov:

Klasifikačná presnosť – určuje pomer správne zaradených objektov v pomere počtu všetkých objektov, vypočítame ju ako:

$$klasifikačná\ presnosť = \frac{TP+TN}{TP+TN+FP+FN}. \quad (3.1)$$

Chybovosť – na rozdiel od presnosti ukazuje opak a teda podiel chybne klasifikovaných subjektov zo všetkých pozorovaní:

$$chybovosť = \frac{FP+FN}{TP+TN+FP+FN}. \quad (3.2)$$

Senzitivita – sa tiež zvykne označovať ako TPR (angl. True Positive Rate). Senzitivita nám pomôže zistiť pomer medzi pravdivo pozitívnymi objektami a súčtu pravdivo pozitívnych a falošne negatívnych objektov:

$$senzitivita = \frac{TP}{TP+FN}. \quad (3.3)$$

3.1.3 ROC krivka

Skôr ako si zadefinujeme ROC krivku, uvedieme si ešte jednu veličinu, s ktorou budeme pracovať:

Špecifickosť – označujeme ju aj ako TNR (angl. True Negative Rate) a predstavuje pomer pravdivo negatívnych objektov k súčtu falošne pozitívnych a pravdivo negatívnych objektov:

$$špecifickosť: \frac{TN}{TN+FP}. \quad (3.4)$$

FPR (angl. False Positive Rate) ďalej označme ako:

$$\text{FPR} = 1 - \text{špecifickosť} = 1 - \frac{TN}{TN+FP} = \frac{FP}{TN+FP}. \quad (3.5)$$

Nech θ predstavuje prahovú hodnotu a Y výsledok testu príslušnosti objektu χ do triedy C_k. Ďalej si definujme t ako binárnu náhodnú premennú, ktorá určuje triedu objektu. Povedzme, že ak t = 1, tak objekt χ patrí do triedy C_k a v prípade, že t = 0, objekt χ do triedy C_k nepatrí. Výsledok testu Y potom bude:

$$\text{pozitívny, ak } Y > \theta$$
$$\text{negatívny, ak } Y \leq \theta .$$

Senzitivitu (TPR) a 1 – špecifickosť (FPR) si následne vyjadríme pomocou pravdepodobnosti, pričom obidve premenné sú závislé od prahovej hodnoty θ:

$$\text{TPR}(\theta) = p(Y > \theta \mid t = 1) \quad (3.6)$$

$$\text{FPR}(\theta) = p(Y > \theta \mid t = 0). \quad (3.7)$$

Z uvedeného vyššie môžeme dedukovať, že ROC krivka bude v podstate zobrazenie všetkých dvojíc (TPR (θ), FPR (θ)) pre rôzne prahové hodnoty θ Є R. Na osi x

sú potom zobrazované hodnoty FPR (θ) a na osi y, hodnoty TPR (θ), pričom $\chi \in [0,1]$. Graf ROC krivky sa vyobrazuje do štvorca, kde osi x aj y začínajú v bode 0 a končia v bode 1. Za výkonné sa považujú tie modely, ktorých ROC krivka sa nachádza blízko k ľavému hornému okraju. Presnejšie kritéria výkonnosti následne dokážeme definovať podľa indexu AUC.

Index AUC (angl. Area Under The Curve) predstavuje plochu pod krivkou ROC a jej hodnotu získame ako:

$$\text{AUC} = \int_0^1 ROC\ (\chi) d\ . \tag{3.8}$$

Aby sme dokázali zobraziť graf ROC krivky, použijeme empirický odhad, čo znamená, že použijeme hodnoty, ktoré sme získali na základe našich dát. Hodnota indexu AUC bude tým pádom empirická.

Možno dedukovať, že čím je hodnota indexu AUC bližšie k číslu 1 (alebo ku 100%, ak výsledok vyjadrujeme v percentách), tým je model užitočnejší. Veľmi nepravdepodobným javom je keď AUC = 1, vtedy hovoríme o takzvanom perfektnom modeli, ktorý sa nikdy nepomýli, ide však o veľmi extrémny prípad.

Tab. 1: Kategorizácia výkonnosti podľa indexu AUC (Bhaqwati a Sinha, 2015)

Hodnota indexu AUC	**Užitočnosť**
0,9 – 1	Perfektný
0,8 – 0,9	Dobrý
0,7 – 0,8	Slušný
0,6 – 0,7	Slabý
0,5 – 0,6	Neužitočný

3.2 Dataset

V našej diplomovej práci sme na aplikáciu jednotlivých metód na predikciu zlyhania použili vzorku 10 000 slovenských firiem v období rokov 2015 až 2017 čerpanú z databázy Finstat Premium, ktoré počas nášho pozorovania neboli obchodované na burze a teda im prislúchal štatút súkromnej spoločnosti. Na základe poznatkov Altmana (1983) ďalej vieme implementovať jeho Z-skóre II., ktoré sa vzťahuje práve na takéto typy spoločností. Poskytnutý dataset, alebo aj súbor dát celkovo obsahoval 568616 pozorovaní a 18 prediktorov, z čoho vyplýva, že mal veľkosť 568616 x 18. V nasledujúcich krokoch

opíšeme komplikácie pri práci s dátami a postup ich spracovania do podoby kedy boli pripravené na implementáciu logistickej regresie a neurónových sietí.

- **Nerovnomernosť datasetu**

Po načítaní datasetu do aplikácie Jupyter Notebook sme si všimli neproporčné výsledky zbankrotovaných firiem, ktorých bolo 864 a z celkového datasetu predstavovali 0.15% a nezbankrotovaných firiem, ktorých množstvo sa vyšplhalo na 567752 a v percentách predstavovali 99.85%. Klasifikácia na nevyvážených súboroch údajov vždy spôsobuje problémy, pretože štandardné algoritmy strojového učenia sú väčšinou zahltené mnoho početnými triedami a malé skupiny ignorujú (Wang a Yao, 2009). Väčšina algoritmov strojového učenia funguje najlepšie, keď je počet vzoriek v každej triede približne rovnaký. Dôvodom je skutočnosť, že väčšina algoritmov je navrhnutá tak, aby maximalizovala presnosť a znížila chyby (Boyle, 2019).

V našej diplomovej práci sme sa rozhodli vyriešiť disproporciu zbankrotovaných a nezbankrotovaných spoločností v poslednom kroku, nakoľko sme sa museli najskôr vysporiadať s viacerými problémami opísanými nižšie, ktoré nám poskytnutý súbor dát dodatočne okresali.

- **Chýbajúce hodnoty**

Okrem nepomeru zlyhaných a nezlyhaných firiem, sme sa museli zaoberať množstvom chýbajúcich dát v datasete.

Luengo, J. (2011) predstavil tri problémy spojené s chýbajúcimi hodnotami a to:

- strata efektívnosti;
- komplikácie pri spracovaní a analýze údajov;
- skreslenosť vyplývajúca z rozdielov medzi chýbajúcimi a úplnými údajmi.

Hodnoty jednotlivých položiek z aktív, pasív a výkazu ziskov a strát spoločnosti nevykazovali rovnomerne v každom roku. Pre lepšie priblíženie, sme sa rozhodli vizualizovať množstvo chýbajúcich dát v jednotlivých determinantoch X1 – X5 pomocou grafov:

- *Determinant X1:*

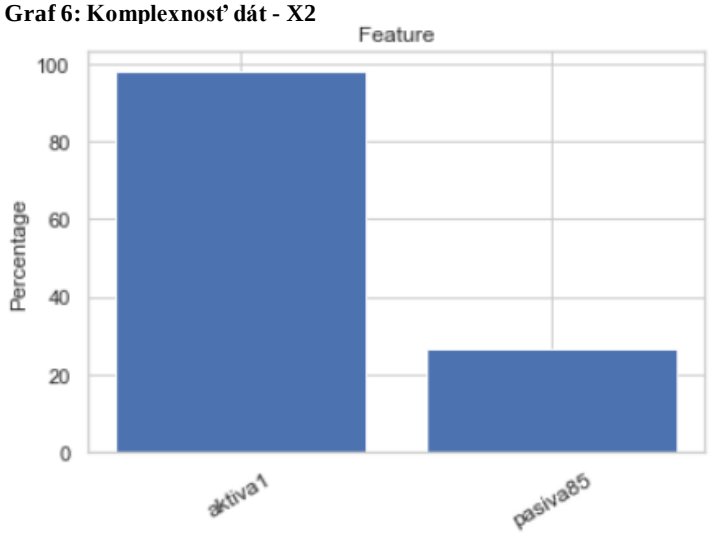
Graf 5: Komplexnosť dát - X1

Tabuľka 3: Komplexnosť dát - X1

Názov položky	Komplexnosť dát
Aktíva 1 / Majetok	98,32%
Aktíva 30 / Obežný majetok	92,12%
Pasíva 106 / Súčet krátkodobých záväzkov	95,44%

- *Determinant X2:*

Graf 6: Komplexnosť dát - X2

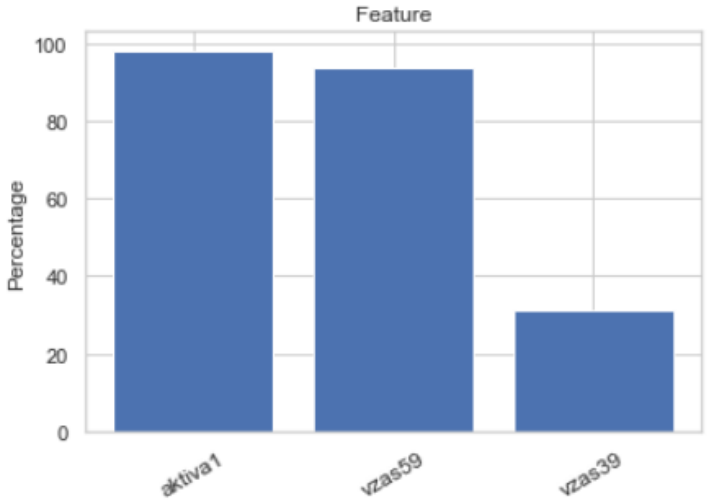


Tabuľka 4: Komplexnosť dát - X2

Názov položky	Komplexnosť dát
Aktíva 1 / Majetok	98,32%
Pasíva 85 / Nerozdelený zisk z minulých rokov	26,46%

- *Determinant X3:*

Graf 7: Komplexnosť dát - X3

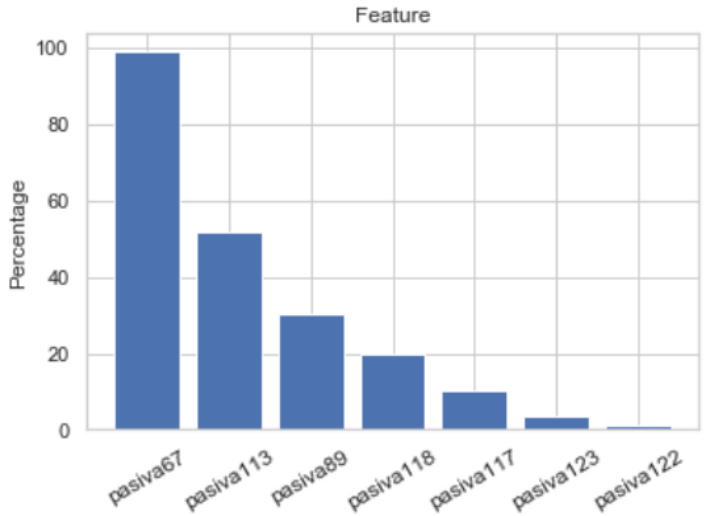


Tabuľka 5: Komplexnosť dát - X3

Názov položky	Komplexnosť dát
Aktíva 1 / Majetok	98,32%
VZaS 59 / Výsledok hospodárenia za účtovné obdobie pred zdanením	93,88%
VZaS 39 / Nákladové úroky	31,3%

- *Determinant X4:*

Graf 8: Komplexnosť dát - X4

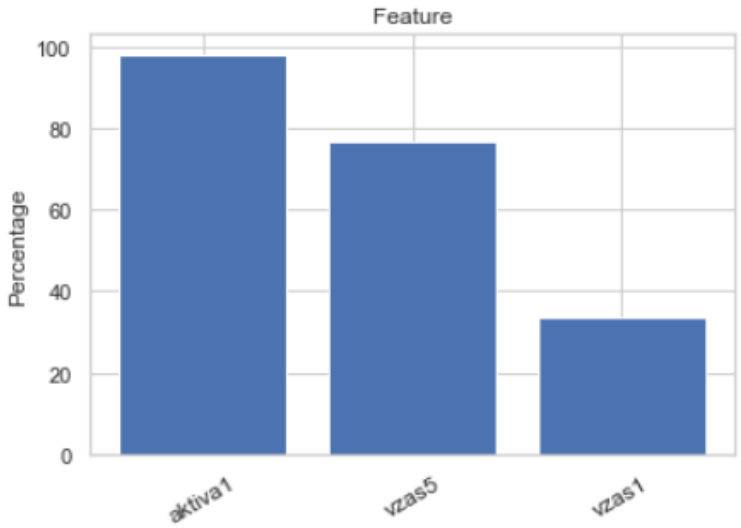


Tabuľka 6: Komplexnosť dát - X4

Názov položky	Komplexnosť dát
Pasíva 67 / Vlastné imanie	98,95%
Pasíva 113 / Záväzky voči zamestnancom	51,92%
Pasíva 89 / Súčet rezerv	30,55%
Pasíva 118 / Bankové úvery	19,82%
Pasíva 117 / Krátkodobé finančné výpomoci	10,37%
Pasíva 123 / Krátkodobé výdavky budúcich období	3,42%
Pasíva 122 / Dlhodobé výdavky budúcich období	1,02%

- *Determinant X5:*

Graf 9: Komplexnosť dát - X5



Tabuľka 7: Komplexnosť dát - X5

Názov položky	Komplexnosť dát
Aktíva 1 / Majetok	98,32%
VZaS 5 / Tržby z predaja vlastných výrobkov a služieb	76,62%
VZaS 1 / Tržby z predaja tovaru	33,37%

Ako môžeme z pozorovania vyššie dedukovať, najviac chýbajúcich hodnôt evidujeme v prípade ukazovateľa *dlhodobých výdavkov budúcich období* a to až 98.98% a pri *krátkodobých výdavkoch budúcich období* vo výške 96.58%. Keďže sú však tieto pozorovania súčasťou determinantu X4, ktorý vystupuje v Altmanovom Z–skóre II., rozhodli sme sa ich zachovať a chýbajúce dáta nahradiť.

- **Nahradenie chýbajúcich dát**

Aby sme mohli pokračovať v modelovaní a testovaní jednotlivých metód, bolo potrebné sa za účelom hodnoverného modelu rozhodnúť, akými hodnotami a na základe akej metodiky budeme chýbajúce dáta nahrádzať. Podľa Kaisera (2011) existujú tri hlavné stratégie riešenia chýbajúcich údajov:

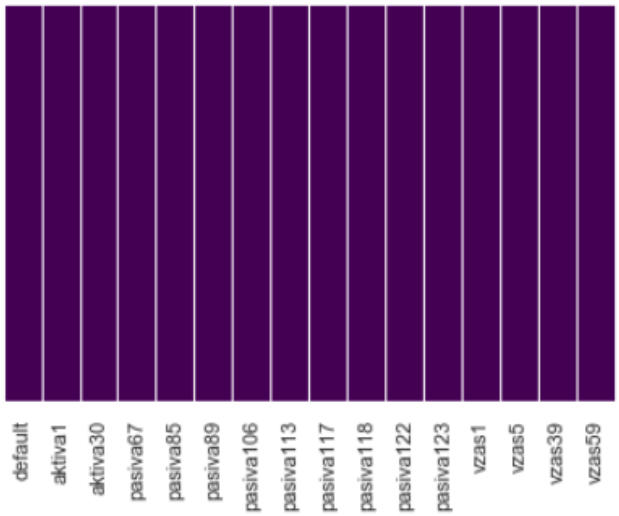
Najjednoduchším riešením problému nahradenia chýbajúcich hodnôt je redukcia súboru údajov a eliminácia všetkých vzoriek s chýbajúcimi hodnotami. Ďalším riešením je považovať chýbajúce hodnoty za takzvané špeciálne hodnoty a v závere, problém s chýbajúcimi hodnotami rieši rôznymi metódami substitúcie chýbajúcich hodnôt – ako

napríklad ich nahradením hodnotami priemeru, mediánu alebo nulou. Metóda substitúcie chýbajúcich dát hodnotou priemeru nahrádza každú chýbajúcu hodnotu priemerom atribútu priemer sa pritom vypočíta na základe všetkých známych hodnôt vzorky. Táto metóda však môže byť ovplyvnená prítomnosťou odľahlých hodnôt a práve preto sa zdá byť prirodzené použiť namiesto priemeru strednú hodnotu, aby sa zabezpečila odolnosť vzorky. V tomto prípade sú chýbajúce údaje nahradené mediánom všetkých známych hodnôt tejto vzorky, do ktorej patrí ukazovateľ s chýbajúcou hodnotou.

Na základe danej metodiky sme sa rozhodli na našom súbore dát postupne otestovať obe metódy náhrady chýbajúcich dát. Z výsledku sme sa dozvedeli, že ak nahradíme chýbajúce hodnoty v jednotlivých vzorkách hodnotou priemeru, dosiahneme o 2% lepší výsledok ako po dosadení mediánu. Preto sme sa rozhodli vo vzorkách, kde je komplexnosť dát vyššia ako 80%, nahradiť chýbajúce dáta práve priemerom. Dodávame, že nahrádzanie mediánom alebo priemerom má zmysel však len vtedy, ak je chýbajúcich hodnôt vo vzorke menší počet (10 – 20%). Chýbajúce dáta ostatných, menej komplexných vzoriek sme teda substituovali nulou.

Po daných krokoch a aplikácii náhrad bol náš dataset kompletný a nevykazoval žiadne chýbajúce pozorovania:

Graf 10: Komplexnosť dát vzorky



- **Odstránenie nulového alebo záporného deliteľa**

Ďalším a nemenej významným krokom bolo v rámci nášho súboru dát zistiť, či sa v ňom nachádzajú pomery, ktoré v deliteli vykazujú nulu alebo záporné číslo. Keďže X1 – X5 sú pomerové ukazovatele, daný test bol nevyhnutný pre dosiahnutie neskresleného datasetu. Na základe vykonaného príkazu sme dostali výsledky zobrazené v tabuľke:

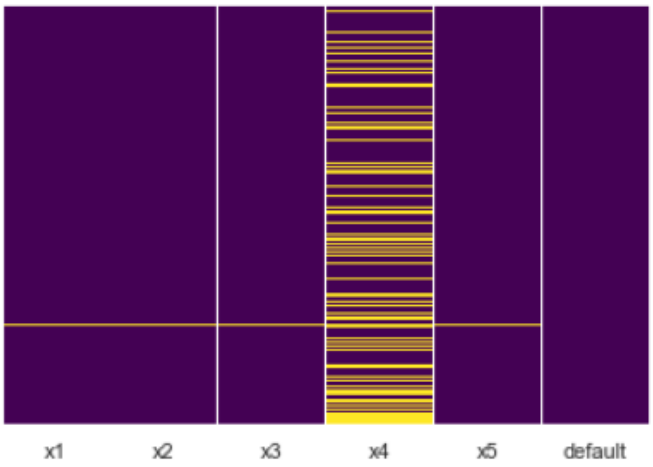
Tabuľka 8: Proporcia nulového alebo záporného deliteľa

Pomerový ukazovateľ	Proporcia nulového/záporného deliteľa
X1	0.5%
X2	0.5%
X3	0.5%
X4	36.77%
X5	0.5%

Tabuľka 8 nám teda naznačuje, že najväčšie percento deliteľa v podobe nuly alebo záporného čísla bolo detekované v ukazovateli X4 a to v až v hodnote 36.77%.

Graf 11 znázorňuje množstvo zachytených deliteľov obsahujúcich nulu poprípade záporné číslo žltou farbou, naopak, fialové prostredie vyjadruje situáciu, kedy v deliteli bolo párne číslo.

Graf 11: Detekcia ukazovateľov s nulovým/záporným deliteľom



Problémom delenia nulou je, že jeho výsledkom ostáva nekonečno, pri delení záporným číslom je výsledok v mínusovej hodnote a tým pádom pre náš súbor dát tieto pomerové ukazovatele neprinášajú relevantné vstupy, preto sme sa rozhodli ich z nášho datasetu odstrániť. Po vykonaní daného kroku bolo pre nás kľúčové zistiť, koľko spoločností s vykázaným bankrotom nám v súbore ostalo:

Tabuľka 9: Množstvo pozorovaní pred a po odstránení deliteľov obsahujúcich nulu a záporné číslo

Množstvo pozorovaní pred odstránením deliteľov obsahujúcich nulu a záporné číslo	568616
Množstvo pozorovaní po odstránení deliteľov obsahujúcich nulu a záporné číslo	358782
Množstvo firiem s pozitívnym záznamom o zlyhaní po odstránení deliteľov obsahujúcich nulu a záporné číslo	651

Z tabuľky vyššie sme teda usúdili, že po aplikácii príkazu, ktorý odstránil z nášho súboru dát všetky prvky obsahujúce v deliteli číslo nula alebo záporné číslo, sme mali k dispozícii 651 spoločností, ktoré za sledované obdobie podľahli zlyhaniu.

- **Undersampling**

Jeden z posledných krokov spracovania dát sa zaoberal proporciou pomeru zlyhaných a nezlyhaných firiem. Ako sme mali možnosť vidieť v postupe vyššie, po odstránení pozorovaní sme sa dostali pri zlyhaných firmách dostali na číslo 651 a pri nezbankrotovaných firmách na číslo 358131. Do ďalšieho modelovania však nezahrnieme všetky pozorovania, ktoré sú nám momentálne dostupné.

Pri nasledujúcom postupe sme sa nechali inšpirovať analýzou Rogera Steina (2007) ktorá tvrdí, že ak je počet zlyhaných pozorovaní vo vzorke veľmi malý, môže to viesť k vysokej variabilite výsledkov, ktoré sa týkajú klasifikačnej presnosti. Dramatický pokles tejto premenlivosti nastáva pri zvyšujúcom sa podiele zlyhaných pozorovaní v celkovej vzorke. Nakoľko problém opisovaný v analýze Rogera Steina má mnoho podobných čŕt ako naše skúmanie, rozhodli sme sa postupovať spôsobom, že do odhadu modelu zahrnieme rovnaký počet zlyhaných a nezlyhaných pozorovaní. Ak by sme aplikovali postup opísaný vyššie na náš dataset, znamenalo by to, že ak máme počet zlyhaných pozorovaní 651, do ďalšieho modelovania bude vstupovať rovnaké množstvo nezlyhaných pozorovaní. Zároveň týmto krokom nastavujeme cut – off hranicu na úrovni 0.50 čím docielime nenáročnú interpretovateľnosť výsledku nášho modelu.

Na základe metodiky Rogera Steina sme sa rozhodli aplikovať metódu pod vzorkovania (angl. Undersampling), ktorá dokáže odstrániť vzorky zo súboru údajov, ktoré patria do väčšinovej triedy (v našom prípade nezbankrotované spoločnosti), aby sa lepšie vyvážilo rozdelenie. Tento postup sa líši od nadmerného vzorkovania (angl. Oversampling), ktoré zahŕňa pridávanie vzoriek do menšinových tried (v našom prípade zlyhané spoločnosti), v snahe znížiť skreslenie v rozdelení vzorky. Pri metóde pod

vzorkovania, môžeme pracovať s troma verziami tejto techniky a to konkrétne s NearMiss-1, NearMiss-2 a NearMiss-3:

NearMiss-1 vyberá vzorky z väčšinovej triedy, ktoré majú najmenšiu priemernú vzdialenosť od troch najbližších vzoriek z menšinovej triedy.

NearMiss-2 vyberá vzorky z väčšinovej triedy, ktoré majú najmenšiu priemernú vzdialenosť od troch najvzdialenejších vzoriek z menšinovej triedy.

NearMiss-3 zahŕňa výber daného počtu vzoriek väčšinovej triedy pre každý príklad v menšinovej triede, ktorá je najbližšia (Browniee, 2020).

Jednotlivé metódy sme následne testovali na našom súbore dát, pričom najlepšie výsledky sme dosiahli pomocou prvej metódy NearMiss - 1. V tabuľke nižšie sme uviedli celkový počet zlyhaných a nezlyhaných podnikov pred a po úprave:

Tabuľka 10: Počet pozorovaní pred a po aplikácii Undersamplingu

Y = 0 (nezlyhané podniky)	pred úpravou	567782
	po úprave	651
Y=1 (zlyhané podniky)	pred úpravou	864
	po úprave	651

- **Odstránenie vzdialených pozorovaní**

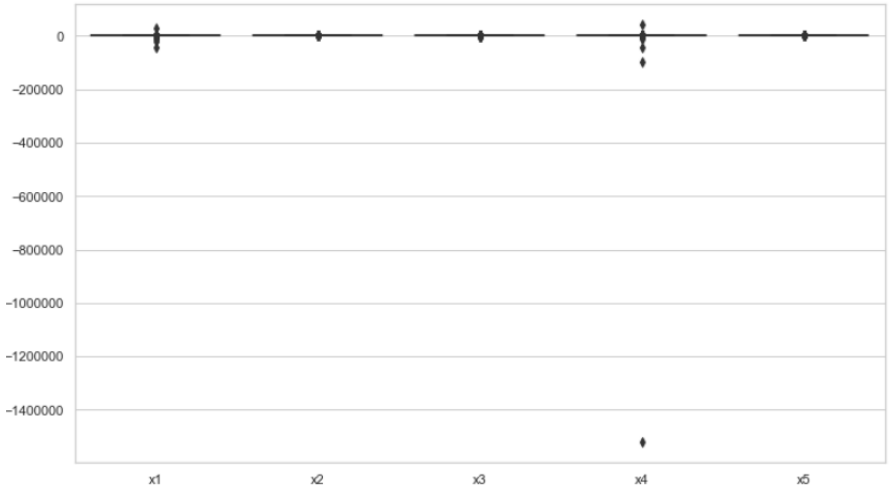
Rôzne vedecké disciplíny sa pravidelne stretávajú s odľahlými hodnotami (angl. Outliers) (Aguinis, 2013). Odľahlé hodnoty sa bežne definujú ako pozorovania, ktoré sa líšia od väčšiny ostatných prípadov vo vzorke (Barnett, 1978; Barnett a Lewis, 1994 a Lorenz, 1987). Takáto odchýlka dát môže byť spôsobená napríklad: chybou merania alebo kódovania, odberom vzorky z nesprávnej populácie, výnimočnými okolnosťami alebo nesprávnym prispôsobením štatistického modelu. V našom prípade ide o abnormálne vzorky spoločností, ktoré za sledované obdobie vykázali v rámci definujúcich pomerov X1 – X5 príliš vysoké alebo príliš nízke hodnoty, ktoré reflektujú ich postavenie na trhu. Keďže sme sa chceli vyhnúť vplyvu skreslenia našich dát, rozhodli sme sa na základe štúdií Barnett a Lewis (1994) tieto dáta odstrániť aby sme získali čo najprimeranejší odhad parametrov spoločností.

Avšak v literatúre sme sa s tým, čo predstavujú odľahlé hodnoty, stretli s oveľa viac rozdielnymi názormi ako s tým, či ich odstrániť alebo nie. Jednoduché východiskové body (napr. štandardné odchýlky, ktoré sú od priemeru vzdialené 3 alebo viac bodov) sú dobrým východiskom, hoci niektorí vedci uprednostňujú vizuálnu kontrolu údajov. Iní (napríklad

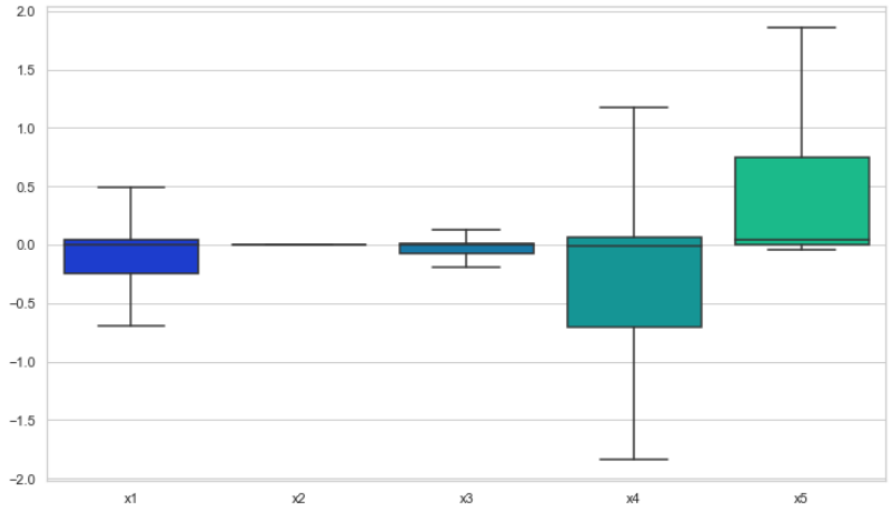
Lorenz, 1987) tvrdia, že detekcia odľahlých údajov je iba špeciálnym prípadom skúmania údajov pre vplyvné dátové body. Jednoduché pravidlá, ako napríklad Z-skóre, predstavujú nezložitú a efektívnu metódu, na základe ktorej sme aj v našej práci odľahlé pozorovania detekovali a následne odstránili.

Van Selst & Jolicœur (1994) predstavili teóriu Z-skóre, v ktorom je možné vzdialenosť medzi skóre a priemerom vyjadriť na základe jednotky štandardnej odchýlky. Napríklad skóre, ktoré je vzdialené jednu štandardnou odchýlkou nad strednou hodnotou, má skóre Z-1. Ak Z–skóre daného pozorovania prekročí ľubovoľnú, vopred určenú hodnotu (napríklad 2.5 alebo 3.0 - v našom prípade 3), považuje sa za extrémne, respektíve odľahlé pozorovanie. Na grafe 12 sú zobrazené odľahlé pozorovania nášho datasetu. Ako môžeme vidieť, tieto pozorovania sú také výrazné, že úplne deformujú krabicové grafy (angl. Box-plot). Grafy tohto typu posudzujú dáta na základe kvartilov (hodnôt, ktoré delia súbor dát na štyri časti, pričom každá z nich obsahuje práve 25% jednotiek) a značia sa nasledovne: x_{25}= dolný kvartil; x_{50}= druhý kvartil alebo aj medián a x_{75}= horný kvartil. V prípade, že hodnota nespadá do daných kvartilov, je vyznačená mimo zobrazenia krabicového grafu, tie sme pre lepšiu vizualizáciu zobrazili do Grafu 12.

Graf 12: Odľahlé pozorovania bez boxplotov



Graf 13: Box-ploty bez odľahlých pozorovaní



Odstránením odľahlých hodnôt sme prišli o 0.77% pozorovaní, čo v číselnom vyjadrení predstavovalo 10 hodnôt a náš súbor dát sa tak zmenšil z pôvodných 1302 pozorovaní na 1292.

- **Rozdelenie datasetu na závislú a nezávislé premenné a následne do tréningovej a testovacej množiny**

Posledným krokom pred samotnou aplikáciou modelov bolo rozdelenie výsledného datasetu na závislú premennú (v našom prípade binárna premenná bankrotu) a nezávislé premenné (ukazovatele X1 – X5).

Okrem rozdelenia vzorky na jednotlivé premenné, je potrebné v rámci nášho súboru vybrať tréningovú množinu, na ktorej bude prebiehať fáza učenia jednotlivých metód, čiže adaptácia váh. Aby sme vedeli otestovať naučenú predikčnú schopnosť siete a logistickej regresie, potrebujeme druhú, testovaciu množinu. Okrem toho je dôležité sa držať pravidla, že model by sa nemal testovať na dátach, na ktorých sa sieť učila (s výnimkou krížovej validácie) a teda, množiny by mali byť disjunktné.

Na základe poznatkov Baesens a kol. (2003) and Kim a kol. (2004) sme sa rozhodli rozdeliť súbor dát v pomere 70%, ktoré prestavovali tréningovú vzorku a 30%, zastupujúce testovaciu vzorku. V konečnom dôsledku však pomer, v akom rozdelíme do množín závisí od veľkosti datasetu, s ktorým pracujeme. Čím viac dostupných dát má model, z ktorých sa dokáže učiť, tým existuje väčšia pravdepodobnosť, že sa dokáže naučiť správne váhy a v konečnom dôsledku správne kategorizovať. Na záver je potrebné zdôrazniť práve dôležitosť testovacej množiny. Ak by sme ju totiž nevytvorili a vložili by sme všetky dáta do tréningovej množiny, zvýšili by sme síce možnosť lepšieho naučenia váh, no

v konečnom dôsledku by sme nevedeli vyhodnotiť, ako dobre sa model tieto váhy naučil. Rozdelenie údajov sme sa rozhodli vykonať náhodne, aby sa zabránilo systematickým rozdielom medzi tréningovou a testovacou vzorkou, čo mohlo vyústiť do problémov s reprezentatívnosťou vzoriek.

Po vykonaní daných krokov sme pozorovali náhodné pomery v tabuľke 11.

Tabuľka 11: Náhodné rozdelenie dát v súbore do jednotlivých častí

Y = 0 (nezlyhané podniky)	Tréningová časť	51.00%
	Testovacia časť	48.97%
Y=1 (zlyhané podniky)	Tréningová časť	49.00%
	Testovacia časť	51.03%

3.3 Lineárna regresia

Po skončení procesu predspracovania dát sme sa mohli pustiť do aplikácie a vyhodnocovania jednotlivých modelov. Rozhodli sme sa finálny súbor dát najskôr testovať prostredníctvom lineárnej regresie, ktorú sme implementovali v knižnici sklearn.linear_model (Yale, 2014).

Pred samotnou aplikáciou modelu sme sa rozhodli vizualizovať jednotlivé odhadnuté parametre s príslušnou významnosťou. Koeficienty logistickej regresie sú odhadované pomocou metódy maximálnej vierohodnosti, čo znamená, že parametre udávajú jednotlivým nezávisle premenným váhu.

Interpretovať odhadnuté parametre prostredníctvom logistickej regresie je o niečo zložitejšie ako v porovnaní s klasickým lineárnym pravdepodobnostným modelom. Kvôli nelineárnym vzťahom medzi nezávislými premennými a pravdepodobnosťou je potrebné interpretovať výsledky nasledovne: ak sa zvýši hodnota nezávislej premennej o jednotku, tak sa v závislosti od toho, či je koeficient kladné alebo záporné číslo, zvyšuje alebo znižuje šanca, respektíve pravdepodobnosť zlyhania. Hodnotu koeficientu teda môžeme označiť ako marginálny vplyv vysvetľujúcej premennej na logaritmus podielu šancí. Podielom šancí je znázornený pomer pravdepodobnosti voľby každej alternatívy.

Z-hodnota reprezentuje informáciu, ako ďaleko sa tento parameter nachádza od nuly. V našom prípade je hladina významnosti α na hranici 0.5, preto môžeme povedať, že parametre ktoré majú hodnotu $P>|z| > 0.05$ sú pre náš model nevýznamné. Z danej definície teda možno zhodnotiť, že ukazovatele X3 a X4 nie sú pre odhad pravdepodobnosti signifikatné.

Naopak, veľký vplyv na šancu zlyhania spoločností môžeme odvodiť z parametrov X1, X2 a X5.

Obrázok 3 zobrazuje základné štatistiky koeficientov v modeli logistickej regresie.

Obrázok 3: LOGIT - Klasický výstup

Results: Logit

Model:	Logit	Pseudo R-squared:	0.293
Dependent Variable:	default	AIC:	1276.5796
Date:	2020-05-17 18:08	BIC:	1302.3993
No. Observations:	1292	Log-Likelihood:	-633.29
Df Model:	4	LL-Null:	-895.51
Df Residuals:	1287	LLR p-value:	3.4724e-112
Converged:	1.0000	Scale:	1.0000
No. Iterations:	13.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
x1	-0.9889	0.2002	-4.9390	0.0000	-1.3813	-0.5964
x2	1.1289	0.2628	4.2949	0.0000	0.6137	1.6440
x3	0.0623	0.0490	1.2732	0.2029	-0.0336	0.1583
x4	-0.0000	0.0000	-0.1915	0.8482	-0.0000	0.0000
x5	2.3302	0.2848	8.1821	0.0000	1.7720	2.8884

Aby sme dokázali odhadnúť čo najlepší výsledok v rámci učenia logistickej regresie, bolo potrebné zaoberať sa regularizáciou, ktorá zabraňuje preučeniu parametrov a zavádza tak lepšiu generalizáciu do klasifikátora. Preučenie spôsobuje, že klasifikátor by mal mať síce dobrú úspešnosť na tréningovej časti dát ale na testovacej časti by už nebol taký úspešný.

Ako prvým bodom v rámci regularizácie sme sa zaoberali parametrom C, ten reprezentuje kladné číslo s pohyblivou rádovou čiarkou (štandardne 1.0), ktoré definuje relatívnu silu regularizácie. Menšie hodnoty pritom znamenajú silnejšiu regularizáciu (NG, 2004). Na základe našich dát sme parameter definovali ako C = 1.0, nakoľko táto hodnota dokázala prispieť k najlepšiemu odhadu logistickej regresie.

Okrem tohto parametra si priblížime najčastejšie využívané typy regularizácií, konkrétne L1 a L2 (Tang a kol., 2014).

• *L2 regularizácia (angl. Ridge regression):*

$$\Lambda(h) = \lambda \cdot ||(a_1, a_2, \ldots, a_n)||^2 = \lambda \cdot (a_1^2 + a_2^2 + \ldots + a_n^2) \qquad (3.7)$$

Regularizácia L2 v podstate "tlačí" váhy nepotrebných atribútov na nulu, pričom väčšie váhy „tlačí" viac. Takže čím dôležitejší atribút, tým väčšiu váhu si môže dovoliť mať.

• *L1 regularizácia (angl. Lasso):*

$$\Lambda(h) = \lambda \cdot (|a_1| + |a_2| + \ldots + |a_n|) \qquad (3.8)$$

Regularizácia L1 podobne "tlačí" nepotrebné atribúty na nulu, pričom ale všetky váhy sú tlačené rovnako. To nám vie vynulovať nepotrebné atribúty a zároveň znížiť výpočtové nároky, čo znamená, že nemusíme pri výpočte rátať s tými ukazovateľmi, ktoré majú nulový koeficient.

Môžeme teda povedať, že aj keď pôvodný účel regularizácie je zabránenie overfittingu dá sa využiť pre selekciu atribútov. Lasso aj Ridge Regression teda dokážu meniť koeficienty pri korelovaných atribútoch na hodnoty blízke nule, Lasso dokonca aj na nulu (čím dané atribúty úplne vypadnú z modelu) (NG, 2004). Pri selekcii stačí potom vybrať tie atribúty, ktoré majú najväčšie koeficienty, prípadne ich majú nenulové.

V rámci nášho testovania sme postupne využili obe popísané regularizácie v nádeji, že prispejú k lepšiemu odhadu zlyhania firiem aj keď v podmienkach našich dát, nakoľko je ich len niečo cez 1000, overfitting nehrozí. Najlepší výstup modelu, na základe predpovede, sme pozorovali v situácii, keď boli obe spomínané regularizácie vynechané a príkaz bol zadefinovaný ako *l1_ratio=None*.

Po aplikácii týchto krokov sme sa mohli dopracovať k finálnym štatistikám modelu, ktorý dosahoval najlepšie výsledky na základe logistickej regresie, uvádzajúc, že model dokázal správne odhadnúť či firma na základe dát spadá medzi zbankrotované alebo nezbankrotované spoločnosti s presnosťou 87.63%, pričom chybný odhad sa vyskytol v 12.37%. Nižšie uvádzame maticu zámen, z ktorej ďalej odvodíme metriky na porovnávanie výsledkov modelov:

Obrázok 4: Matica zámen v podmienkach logistickej regresie

		Predpovedané 0	Predpovedané 1
Skutočné	0	TN = 157	FP = 7
	1	FN = 41	TP = 183

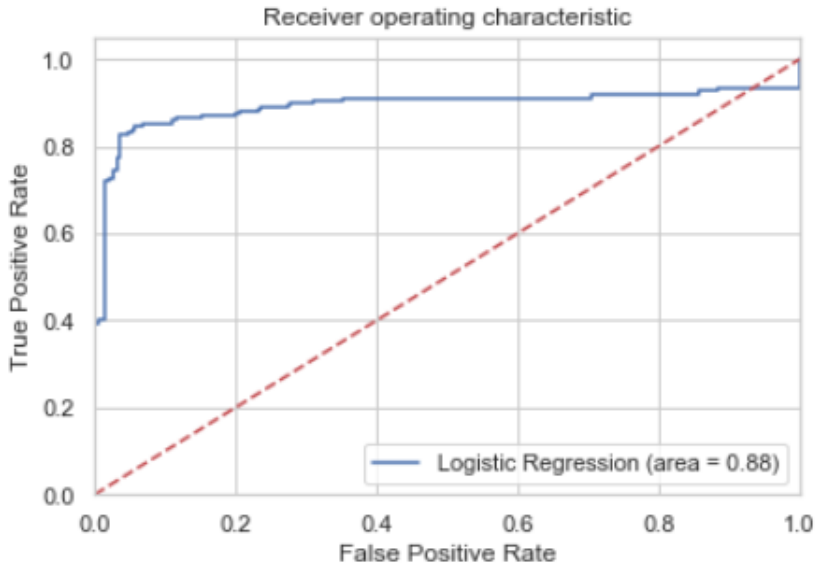
Na základe matice zámen sme boli ďalej schopní definovať metriky, ktoré sme si v úvode praktickej časti uviedli ako kľúčové a ktoré nám pomohli pri porovnávaní výkonnosti jednotlivých modelov.

Tab. 2 Klasifikačný report pre lineárnu regresiu

Klasifikačná presnosť	Chybovosť	Senzitivita	Špecifickosť	Hodnota AUC
87,6%	12,4%	81,7%	95,73	88%

Výsledok indexu AUC, teda plochy pod krivkou ROC nám vyšiel v hodnote 88% a podľa teoretického podkladu teda môžeme dedukovať, že jeho hodnota reflektuje užitočnosť modelu akú „dobrú“. Čím vyššiu hodnotu totiž index AUC dosahuje, tým je model presnejší. Okrem toho nižšie prikladáme aj ROC krivku prislúchajúcu k danému modelu. Dôkladnejšiu analýzu si uvedieme v kapitole 3.5 – Porovnanie výsledkov.

Obrázok 5: ROC krivka pre model logistickej regresie



3.4 Neurónové siete

Aby sme dokázali porovnať dosiahnuté výstupy pomocou logistickej regresie, bolo potrebné na náš súbor dát aplikovať metódu neurónových sietí.

V začiatkoch aplikácie modelu bolo nevyhnutné definovať štandardizáciu hodnôt, poprípade ich normalizáciu do rozsahu < n;m >, ktorá sa nazýva aj min-max normalizácia:

- *štandardizácia metód* – dokáže transformovať dáta na dáta s nulovým priemerom a zároveň s jednotkovou štandardnou odchýlkou. Daná štandardizácia dokáže dopomôcť k lepšej konvergencii modelu, ktorý využíva lineárne aktivačné funkcie.
- *normalizácia hodnôt do rozsahu < n;m > (Min-max normalizácia)* – sa výborné hodí ku nelineárnym aktivačným funkciám a dopomáha k tomu, aby sme neprišli o znalosti v dátach, ktorých hodnota by bola vyššia ako obor hodnôt aktivačnej funkcie (dáta, ktoré by prevyšovali limit by boli zovšeobecnené) a tým prispieva k lepšej konvergencii modelu. Je teda pochopiteľné, že hodnoty pre m a n sú volené na základe oboru hodnôt aktivačnej funkcie sigmoid.

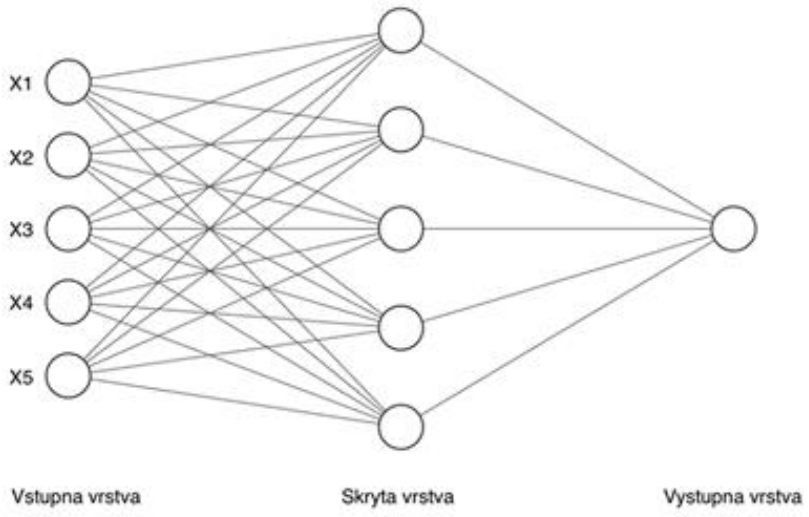
Považujeme za nutné poznamenať, že v práci sme nepoužili obe metódy súčasne. Na základe implementácie sa ukázalo, že štandardizácia hodnôt dosahovala empiricky lepšie výsledky.

Nižšie sme postupne uviedli jednotlivé variabilné zložky nášho modelu, ktoré dopomohli k jeho najlepšej výkonnosti.

V danom modeli sme sa rozhodli použiť jednu skrytú vrstvu, ktorá obsahovala rovnaký počet neurónov ako vstupná vrstva. Častokrát sa pri modelovaní využíva situácia kedy vstupná vrstva obsahuje dvojnásobok neurónov. V podmienkach zadaných dát sme ale v prostredí, kedy obe vrstvy obsahovali rovnaký počet neurónov dosiahli najvýkonnejší model.

Naším cieľom je úloha binárnej klasifikácie, čo znamená, že v danom modeli sa nachádza len jeden výstupný neurón, ktorý rozhodujeme o čom či daná firma podľahla alebo nepodľahla bankrotu. Výstupný neurón teda určuje pravdepodobnosť každej spoločnosti.

Obrázok 6: Model neurónovej siete v daných podmienkach



Nasledujúcim krokom pri modelovaní výstupu bolo definovať aktivačné funkcie (viac v kapitole 2.3) pre skrytú aj výstupnú vrstvu, ktoré pomáhajú budovať predikčný model. Po niekoľkých pokusoch nám kombinácia ELU – v skrytej vrstve a Sigmoid vo výstupnej vrstve priniesli najlepší možný výsledok.

Obrázok 7: Architektúra neurónovej siete

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 5)                 30
_________________________________________________________________
dense_2 (Dense)              (None, 5)                 30
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 6
=================================================================
Total params: 66
Trainable params: 66
Non-trainable params: 0
_________________________________________________________________
```

V záujme učenia neurónových sietí sa využívajú rôzne algoritmy založené na gradientnom zostupe. Jedným z najznámejších algoritmov určeným na učenie neurónových sietí je stochastický gradientný zostup (angl. Stochastic Gradient Descent / SGD). Tento algoritmus garantuje nájdenie lokálneho minima funkcie, ale môže k tomuto lokálnemu minimu konvergovať veľmi pomaly. Pre zlepšenie učenia neurónových sietí, bolo preto vytvorených niekoľko algoritmov, ako napríklad ADAM, u ktorého bolo dokázané, že k lokálnemu minimu konverguje rýchlejšie ako ostatné algoritmy (Kingma a Ba, 2017).

ADAM predstavuje adaptívny algoritmus založený na momentovej metóde. Tá sa od SGD líši tým, že si okrem lokálneho gradientu pamätá aj predchádzajúce kroky. Jeden krok je potom lineárnou kombináciou gradientu v aktuálnom bode a predchádzajúceho kroku. Adaptívne metódy sa hodia práve na riedke dáta, pretože viac upravujú menej frekventované parametre. Na základe jeho pozitívnych výsledkov, sme sa rozhodli ho zahrnúť do modelovania.

Jednou z posledných otázok k finálnemu výsledku bolo nastavenie množstva epôch. Tie určujú ako dlho bude trvať doba učenia a teda koľkokrát sa tréningová množina prejde vo fáze učenia (ak predpokladáme, že učenie nebude ukončené predčasne). Pri čísle 20 epôch sa nám náš model nedokázal dostatočne natrénovať naopak, pri množstve 100 epôch sme dosahovali veľmi dobré výsledky. Vo finále sme sa však rozhodli pre množstvo 60 epôch, nakoľko táto hodnota dosahovalo empiricky najlepšie výsledky.

Veľkosť dávky (angl. Batch Size) vyjadruje veľkosť sekvencie vstupov, ktorá naraz postúpi do modelu a z ukazovateľa rýchlosti učenia (angl. Learning rate) vieme vyčítať to, ako rýchlo bude algoritmus učenia postupovať a teda o akú veľkú časť bude posúvať váhy.

V tabuľke 12 uvádzame detaily konfigurácie, ktoré sme zadefinovali v záujme dosiahnutia čo najpresnejšej predikcie v našom modeli.

Tabuľka 12: Konfigurácia predikčného modelu

Atribút	Hodnota
Počet neurónov vo vstupnej vrstve	5
Počet neurónov v skrytej vrstve	5
Počet neurónov vo výstupnej vrstve	1
Počet epôch	60
Veľkosť dávky	4
Algoritmus učenia	ADAM
Rýchlosť učenia	0.003
Aktivačná funkcia skrytej vrstvy	ELU
Aktivačná funkcia výstupnej vrstvy	Sigmoid

Po časti konfigurácie predikčného modelu sme boli pripravení spustiť jeho učenie a zistiť s akou pravdepodobnosťou bol schopný odhadnúť zlyhanie jednotlivých firiem. Model nám vo finále poskytol výsledok, ktorý poukazoval na to, že neurónovej sieti sa na základe poskytnutých dát podarilo odhadnúť zlyhanie spoločností s presnosťou 89.43%. Nižšie uvádzame maticu zámen:

Obrázok 8: Matica zámen pre Neurónové siete

		Predpovedané 0	Predpovedané 1
Skutočné	0	TN = 168	FP = 11
	1	FN = 30	TP = 179

Podobne ako pri modelovaní výstupu po aplikácii logistickej regresii, dokážeme na základe danej matice zámen vypočítať metriky vyhodnocujúce výkonnosť modelov.

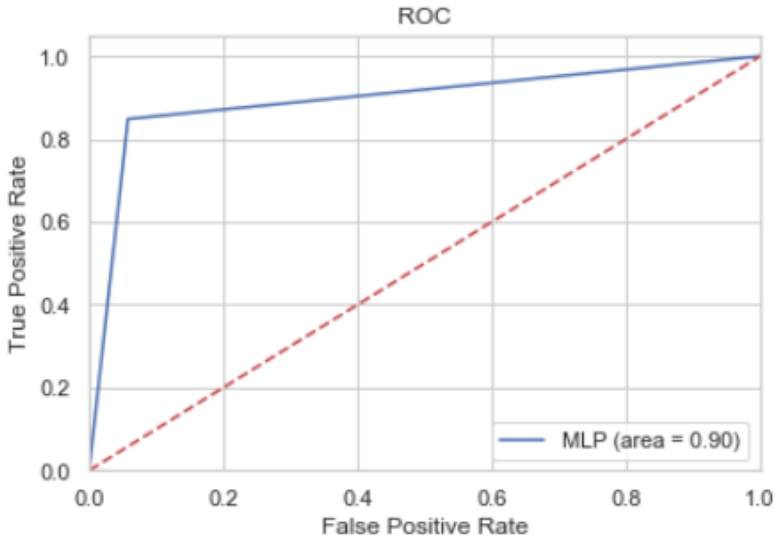
Tabuľka 13: Klasifikačný report pre Neurónové siete

Klasifikačná presnosť	Chybovosť	Senzitivita	Špecifickosť	Hodnota AUC
89,4%	10,6%	85,7%	93,86%	90%

V prípade neurónových sietí môžeme pozorovať o niečo lepší výsledok AUC indexu, ktorý za daných okolností dosiahol hodnotu 90% a teda môžeme zhodnotiť, že

výkonnosť modelu je výborná. Okrem klasifikačného reportu sme uviedli vizualizáciu ROC krivky v podmienkach neurónových sietí.

Obrázok 9: ROC krivka v podmienkach Neurónovej siete



3.5 Porovnanie výsledkov

V záverečnej kapitole našej diplomovej práci sme sa venovali porovnaniu výsledkov, ktoré sme dosiahli po vykonaní všetkých predošlých úkonov v modeli logistickej regresie či neurónových sietí. V danej časti zhrnieme výsledky aplikácie našich metrík a taktiež zhodnotíme výkonnosť jednotlivých modelov na základe ROC krivky, respektíve AUC indexu. Podotýkame, že v práci sme sa venovali výsledkom, ktoré dosiahli naše najvýkonnejšie modely.

Špecifickosť a senzitivita patria medzi štatistiky týkajúce sa klasifikačnej presnosti, avšak obe závisia od nastavenej cut – off hranice, ktorá bola v našej práci na hodnote 0.5. Možno teda dedukovať, že všetky pozorovania s predikovanou pravdepodobnosťou vyššou ako 0.5 budú zaradené do kategórie Y = 1 a teda zaradené vzorky budú patriť medzi zlyhané spoločnosti. V prípade, že bude pozorovaná predikovaná pravdepodobnosť nižšia ako 0.5, bude pozorovanie patriť do kategórie Y = 0, ktoré reprezentuje nezlyhané spoločnosti. Na zopakovanie, špecifickosť udáva percentuálny podiel pozorovaní, ktoré v skutočnosti boli nezlyhané a aj náš model ich správne zaradil medzi nezlyhané. Naopak, senzitivita uvádza percentuálny podiel pozorovaní, ktoré v skutočnosti boli zlyhané a podobne, aj náš model ich označil ako zlyhané. Keďže cieľom našej práce je určiť, ktorý model dokáže lepšie predikovať zlyhané spoločnosti, je pre nás metrika senzitivity smerodajnejšia.

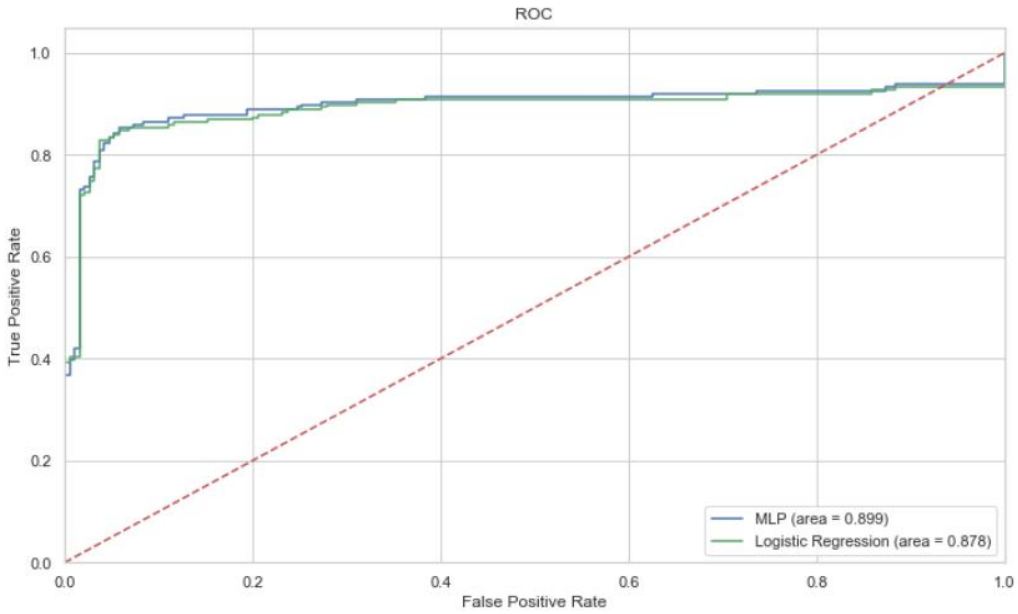
V modeli logistickej regresie dosiahla špecifickosť hodnotu 95.73% a ukazovateľ senzitivity hodnotu 81.7%. Model neurónových sietí dosiahol o niečo horší výstup v ukazovateli špecifickosť a to konkrétne 93.86%. Senzitivite sa ale v podmienkach neurónovej siete darilo o niečo lepšie a dokázala dosiahnuť hodnotu 85.7%. Ako sme spomínali vyššie, v podmienkach našej práce a nášho účelu je ukazovateľ senzitivity podstatnejší, preto môžeme zhodnotiť, že neurónovej sieti sa v danej klasifikácii darilo lepšie.

Okrem toho, z pohľadu potencionálnych dôsledkov je prijateľnejšie, ak dokáže model označiť spoločnosť, ktorá smeruje do úpadku, aj keď v skutočnosti sa považuje za zdravú, ako keby ju model označil za bezproblémovú, no v skutočnosti by sa borila s mnohými komplikáciami a problémami. V súvislosti s daným bodom, sme sa teda pozreli aj na túto problematiku. Z porovnania matíc zámen sme zistili, že model logistickej regresie dokázal odhadnúť 7 spoločností a model neurónových sietí 11 spoločností, ktoré smerujú do úpadku aj keď sú tieto v skutočnosti považované za zdravé. Naopak model logistickej regresie odhadol 41 spoločností a model neurónových sietí 30 spoločností ktoré vyhodnotili ako zdravé, aj keď sa v skutočnosti označovali za zlyhané. Opäť teda môžeme zhodnotiť, že modelu neurónových sietí sa darilo v odhade lepšie.

Okrem vyššie spomenutých metrík sme sa tiež zamerali na klasifikačnú presnosť, ktorá nám v rámci výsledku logistickej regresie skončila na čísle 87.6% a neurónových sietí 89.4%. V podmienkach chybovosti, ktorá je teda opakom klasifikačnej presnosti a teda zobrazuje podiel chybných odhadov dosiahla logistická regresia hodnotu 12.4% a neurónová sieť 10.6%, v tomto prípade je však nižšie číslo vo výsledku pozitívnejšie, nakoľko z neho vyplýva množstvo chybných odhadov modelu.

Poslednou menovanou metrikou bola ROC krivka, na základe ktorej sme vedeli interpretovať AUC index v jednotlivých modeloch a teda skonštatovať ich klasifikačnú presnosť. Výkonnosť modelu logistickej regresie, ako môžeme tiež dedukovať z obrázka 10, bola na hodnote 87.8%. Model neurónových sietí si viedol o pár percent lepšie a dosiahol výkonnosť na hodnote 89.9%. Pre lepšie interpretovateľné výsledky sme sa rozhodli percentá zaokrúhliť a môžeme teda zhodnotiť, že v podmienkach záveru AUC indexu si model logistickej regresie viedol na základe daných dát o niečo horšie a dosiahol „dobré“ výsledky na hodnote 88%. O niečo lepšie vedel predikovať zlyhanie firiem model neurónových sietí a to na čísle 90%, ktoré dokážeme na základe AUC indexu zhodnotiť ako „výbornú“ výkonnosť.

Obrázok 10: ROC krivka v podmienkach neurónovej siete a logistickej regresie



Ako vyplýva z porovnania vyššie, hodnoty metrík umelej neurónovej siete vykazujú o niečo lepšie výsledky v podmienkach našich hyperparametrických nastavení ako klasická štatistická metóda, konkrétne klasifikácia pomocou logistickej regresie. V závere teda môžeme konštatovať, že neurónová sieť ako všeobecný aproximátor bola schopná sa naučiť váhy dostatočne dobre, avšak nie výrazne lepšie ako logistická regresia.

4 Záver

Neistota je vzhľadom na súčasné turbulentné podmienky veľmi častým javom. Či už ju pociťujeme v súkromnom alebo profesijnom živote. Zamestnanci sa boja o svoje miesta a zamestnávatelia o stratu dlho budovaných firiem. Častokrát si nechávajú vypracovávať podrobné analýzy svojich podnikateľských modelov či rozhodnutí a snažia sa popri maximalizácii zisku podporovať svoju firmu v raste.

Cieľom predkladanej záverečnej práce bola práve konštrukcia predikčných modelov v podmienkach logistickej regresie a neurónových sietí založená na informáciách z účtovných výkazov spoločností pôsobiacich na Slovensku. Zaujímalo nás akú klasifikačnú presnosť dokážu vykázať nami aplikované metódy na základe verejne dostupných informácií a na ich výstupe zhodnotiť relevanciu a úspešnosť predikcie. Samotným odhadom však predchádzalo niekoľko činností, ktoré sme sa snažili čitateľovi priblížiť pomocou stručného, no výstižného prehľadu najdôležitejších pojmov a súvislostí.

V prvej časti práce sme sa snažili objasniť teoretické podklady a priblížiť si skúmanú problematiku bankrotových modelov a popísali sme jednotlivé determinanty, ktoré majú vplyv na zlyhanie spoločností. Okrem toho sme zadefinovali problematiku v súčasnosti a zachytili rôzne názory v priebehu dvadsiateho a súčasného storočia.

Kapitola cieľov a metodiky nám dopomohla k definícií zámeru a našich motivácií pre ďalšie pokračovanie v diplomovej práci a súčasne nám poskytla teoretický základ, z ktorého sme vychádzali pri aplikácii a opise jednotlivých modelov. Daná časť práce nás viedla k detailnému návrhu riešenia, ktorého aspekty sme sa následne snažili koncipovať v súlade so zaužívanými postupmi a aktuálnymi trendami v skúmanej problematike. Venovali sme sa klasifikácii logistickej regresie, jej odhadom či interpretácii a testom významnosti regresných koeficientov. Pri neurónových sieťach nás zaujímal najmä spôsob akým sa dokážu učiť, ako ich možno cvičiť a tiež ich využitie.

Výsledky práce boli poslednou kapitolou nášho skúmania, na základe ktorej sme sa postupne dopracovali ku pravdepodobnosti zlyhania spoločností z nášho súboru dát v podmienkach logistickej regresie a neurónových sietí. V rámci dosiahnutia relevantných výsledkov, sme však museli podstúpiť niekoľko dodatočných krokov. Aby sme vedeli porovnať dosiahnuté výsledky, bolo potrebné si zadefinovať určitý spôsob evaluácie modelov cez rôzne metriky ako klasifikačná presnosť, chybovosť, senzitivita a špecifickosť, ktoré sme vedeli vyčítať z matice zámen, alebo cez AUC index, ktorý sa dá interpretovať cez ROC krivku. Pred samotnou aplikáciou bolo potrebné náš súbor dát

upraviť do takej miery, ktorá nebude deformovať dosiahnuté výsledky. Cez jednotlivé kroky ako napríklad odstraňovanie odľahlých pozorovaní či podvzorkovanie sme dospeli k finálnej podobe našich dát, na ktoré sme následne aplikovali metódu lineárnej regresie a neurónových sietí a tak sme zisťovali, ktorý z týchto prístupov bude vhodnejší a presnejší v rámci predikcie zlyhania spoločností na našom súbore dát.

V závere sme porovnali dosiahnuté výsledky na základe predvolených metrík a zistili sme, že model neurónovej siete bol schopný sa naučiť dáta lepšie a teda dokázal odhadnúť zlyhanie firiem s o niečo vyššou presnosťou. Štandardný štatistický prístup v podobe logistickej regresie však zaostával len vo veľmi nízkom percente, preto považujeme oba odhady za veľmi presné a s relevantnými výsledkami.

5 Zdroje

AGUINIS, H. – GOTTFREDSON, R. K. – JOO, H. -Gottfredson, R. K., & Joo, H., *Best-practice Recommendations For Defining, Identifying, And Handling Outliers*. In : Organizational Research Methods, 2013, vyd. 16, s. 270–301.

ALTMAN, D.; *Statistics in practice*. In : Statistics in medicine, 1982, vyd. 17, č. 23, s. 59-71.

ALTMAN, E. I.; *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*. In :The Journal of Finance, 1968, roč. 23, č.4, s. 589 – 609.

ALTMAN, E. I. – HALDEMAN, R. G. – NARAYANAN, P.; *ZETATM Analysis A New Model To Identify Bankruptcy Risk Of Corporations.* In : Journal of Banking & Finance, Elsevier, 1977, vyd. 1, č. 1, s. 29-54.

ALTMAN, E. I.*; Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy*. New York: John Wiley & Sons, 1983. S. 368. ISBN 978-0-471-08707-6.

ALTMAN, E. I. – SABATO, G. – WILSON, N.; *The Value of Non-financial In1formation in Small and Medium-sized Enterprise Risk Management*. In : Journal of Credit Risk, 2010. s. 95 – 127.

ALTMAN, E. I. – IWANICZ-DROZDOWSKA, M. – LAITINEN, E. K. – SUVAS, A.; *Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model.* In : Journal of International Financial Management & Accounting, 2017, vyd. 28, č. 2, s. 131 – 171.

ALY, I. M. – BARLOW, H. A. – JONES, R. W.; *The Usefulness of SFAS No. 82 (Current Cost) Information in Discriminating Business Failure: An Empirical Study*. In : Journal of Accounting, Auditing & Finance, 1992, vyd. 7 , č.2, s. 217-229.

BALDWIN, J – GLEZEN, W. G.; *Bankruptcy Prediction Using Quarterly Financial Statement Data.* In : Journal of Accounting, Auditing & Finance, 1992, vyd. 7, č. 3, s. 269-285.

BARNETT, V., *The Study Of Outliers: Ppurpose And Model*. In: Journal of the Royal Statistical Society, Series C (Applied Statistics), 1978, vyd. 27, s. 242–250.

BARNETT, V. – LEWIS, T., *Outliers in Statistical Data*. In : John Wiley & Sons, 1994, vyd. 3.

BEAVER, W. H.; *Financial Ratios as Predictors of Failure. Journal of Accounting Research,* In : Supplement to the Journal of Accounting Research. 1966, roč. 4, s. 71 – 111.

BELLOVARY, J.L. – GIACOMINO J.L. – AKERS, M. D.; *A Review of Bankruptcy Prediction Studies: 1930 to Present.* In : Journal of Financial Education, 2007, vyd. 33, s. 3-41.

BHAQWATI, CH. P. – SINHA, G. R., *Reliable Computer-aided Diagnosis System using Region based Segmentation of Mammographic Breast Cancer Images* [online]. Tenerife: WSEAS Press, 2015, vyd. 10, č.12, s. 131 – 142. [cit.2020-2-1]. Dostupné na: https://www.researchgate.net/publication/299844562_Reliable_Computer-aided_Diagnosis_System_using_Region_based_Segmentation_of_Mammographic_Breast_Cancer_Images

BOYLE, T., Dealing with Imbalanced Data [online]. Publikované: 03.02.2019. [cit.2020-15-2]. Dostupné na: https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

BROWNIEE, J., *Undersampling Algorithms for Imbalanced Classification* [online]. Publikované: 20.01.2020. [cit.2020-03-03]. Dostupné na: https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

CLEVERT, D.A. – UNTERHINER, T. – HOCHREITER, S.; *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. Cornell University, 2015.

DEAKIN, E. B.; *A Discriminant Analysis of Predictors of Business Failure.* In : Journal of Accounting Research, 1972, vyd. 10, č. 1, s. 167 – 179.

DU JARDIN, P.; *Bankruptcy Prediction And Neural Networks: The Contribution Of Variable Selection Methods.* In: Proceedings of the Second European Symposium on Time Series Prediction 2008. s. 271-284.

FIDRMUC, J. – HAINZ, C.; *Default Rates in the Loan Market for SMEs: Evidence from Slovakia*. In : Economic Systems, 2010, vyd. 34, č. 2, s. 133 – 147.

GENTRY, J. A. – NEWBOLD, P. – WHITFORD, D.T.; *Classifying bankrupt firms with funds flow components*. In: Journal of Accounting Research, 1985, vyd. 23, č. 1, s. 146–160.

GEPP, A. – KUMAR, K. (2008). *The Role Of Survival Analysis In Financial Distress Prediction* [online]. In : International Research Journal of Finance and Economics, 2008, vyd. 16, s. 13-34, [cit. 2020-20-1]. Dostupné na: https://www.researchgate.net/publication/27826547_The_Role_of_Survival_Analysis_in_Financial_Distress_Prediction

GILBERT, L.R. – MENON, K. - SCHWARTZ, K. B.; *Predicting Bankrupcy For Firms In Financial Distress*. In : Journal of Business Finance & Accounting, 1990, vyd. 17, č. 1, s. 161 – 171.

GRUNERT, J. - NORDEN, L. - WEBER, M.; *The role of non-financial factors in internal credit ratings*. In : Journal of Banking and Finance,2004, vyd. 29, č. 2, s. 509–531.

GRUNERT, J. – NORDEN, L. – WEBER, M.; *The Role of Non-financial Factors in Internal Credit Ratings.* In : Journal of Banking & Finance, 2005, vyd. 29, č.52, s. 509 – 531.

GURNEY, K.; *An introduction to neural networks* [online]. London: UCL Press, 1997. 317 s. ISBN: 0-203-45151-1. [cit.2020-25-1]. Dostupné na: http://www.macs.hw.ac.uk/~yjc32/project/ref-NN/Gurney_et_al.pdf

JOHNSEN, J. – MELICHER, R.W.; *Predicting corporate bankruptcy and financial distress: Information value added by multinomial logit models*. In : Journal of Economics and Business, 1994, vyd. 46, č. 4, s. 269 – 286.

KÁČER, M. – OCHOTNICKÝ, P. – ALEXY. M.; *The Altman's Revised Z'-Score Model, Non-financial Information and Macroeconomic Variables: Case of Slovak SMEs* [online]. In : Ekonomický časopis, 2019, vyd. 67, č.4, s. 335 – 366, [cit. 2019-21-12]. Dostupné na: https://www.sav.sk/journals/uploads/0517110904%2019%20Alexy%20+%20SR.pdf

KAISER, J., *Dealing with Missing Values in Data* [online]. In: Journal of Systems Intergation, 2011, vyd. 1, s. 42-51. [cit.2020-02-03]. Dostupné na: http://si-journal.org/index.php/JSI/article/viewFile/178/255

KEASY, K. – WATSON, R.; *Non-Financial Symptoms and the Prediction of Small Company Failure: A Test of Argenti's Hypotheses*. In : Journal of Business Finance & Accounting, 1987, vyd. 14, č. 3, s. 335 – 354.

KEASY, K. – WATSON, R.; *Financial Distress Prediction Models*: A Review of Their Usefulness. In : British Journal Of Management, 1991, vyd. 2, č. 2, s. 89 – 102.

KIM, Y. S. – SOHN, S. Y., Kim, Y.S., *Managing Loan Customers Using Misclassification Patterns Of Credit Scoring Model*. In : Expert Systems with Applications, 2004, vyd. 26, s. 567-573.

KINGMA, D. P – BA, J., *Adam: A Method for Stochastic Optimization* [online]. In: ICLR, 2017. [cit.-2020-04.04.09]. Dostupné na: https://arxiv.org/pdf/1412.6980.pdf

KUO, Y. F.; *A Study On Service Quality Of Community Websites.* In : Total Quality Management and Business Excellence,2003, vyd. 14, s. 461-473.

LAITINEN, T. – KANKAANPAA, M. - *Comparative analysis of failure prediction methods: the Finnish case*. In : Taylor & Francis Journals, 1999, vyd. 8, č. 1, s. 67 – 92.

LONG, G., *Jupyter: Intro to Data Science – Lecture 5 Regression in Python* [online]. In : CUNY City College, 2018. [cit.2020-18-04-2020]. Dostupné na: https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1285&context=cc_oers

LORENZ, F. O., Teaching About Influence in Simple Regression.In : Teaching Sociology, 1987, vyd. 15, č. 2, s. 173-77.

LUENGO, J. – GARCÍA, S. – HERRERA, F., *On The Choice Of The Best Imputation Methods For Missing Values Considering Three Groups Of Classification Methods* [online]. Publikované: 14.06.2011. [cit.2020-16-02]. Dostupné na: http://sci2s.ugr.es/MVDM/index.php

MELOUN, M. – MILITKÝ, J. – HILL, M.; *Statistická analýza vícerozměrných dat v příkladech.* 1. vyd. Praha : Academia, nakladatelství Československé akademie věd, 2012. 750 s. ISBN 978-80-200-2071-0.

MENSAH, Y.; *An Examination Of The Stationarity Of Multivariate Bankruptcy Prediction Models - A Methodological Study*. In : Journal of Accounting Research, 1987, vyd. 22, č. 1, s. 380-395.

MORAVČÍK T.; *Inovatívne modelovacie metódy – Neurónové siete*. In: Forum Statisticum Slovakum. Bratislava: Slovenská štatistická a demografická spoločnosť, 2010, roč. 6, č. 6 , s. 56-59. ISSN 1336-7420.

NEOPHYTOU, E. – MAR MOLINERO, C.; *Predicting Corporate Failure in the UK: A Multidimensional Scaling Approach* [online]. Southampton: School of Management, University of Southampton, 2001, č.1, [cit. 2019-20-12]. ISSN 1356 – 3548. Dostupné na: https://eprints.soton.ac.uk/35733/1/01-172.pdf

NG, A. Y., Feature Selection, L1 vs. L2 Regularization, And Rotational Invariance [online]. In: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, s. 78 – 86. [cit.2020-25-03]. Dostupné na: https://dl.acm.org/doi/pdf/10.1145/1015330.1015435

ODURO, R. – ASIEDU, M.; *A Model To Predict Corporate Failure In The Developing Economies: A Case Of Listed Companies On The Ghana Stock Exchange* [online]. In : Journal of Applied Finance & Banking, 2017, vyd. 7, č. 4, s. 79 – 92, [cit. 2019-28-1]. Dostupné na: https://www.researchgate.net/publication/319270961_A_model_to_predict_corporate_failure_in_the_developing_economies_A_case_of_listed_companies_on_the_Ghana_Stock_Exchange

OHLSON, J. A.; *Financial Ratios And The Probabilistic Prediction Of Bankrupcy*. In : Journal of Accounting Research, vyd. 18, č. 1, 1980.

PEEL, M. J. – PEEL, D. A.; *A multi-logit approach to predicting corporate failure: some evidence for the UK corporate sector* [online]. In. Omega International Journal of Management Science, 1989, vyd. 13, č. 3, s. 309–318, [cit. 2019-10-12]. Dostupné na: https://www.researchgate.net/publication/24080149_Predicting_corporate_failure_empirical_evidence_for_the_UK

PEEL, M. J – WILSON, N.; *Working Capital and Financial Management Practices in the Small Firm Sector*. In : International Small Business Journal, 1996, vyd. 14, č. 2, s. 52-68.

PLATT, H.D. - PLATT, M. B.; *Developmpment Of A Class Of Stable Predictive Variables: The case of bankrupcy prediction.* In : Journal of Business Finance & Accounting, 1990, vyd. 17, č. 1, s. 31 – 51.

PLATT, H. D. – PLATT, M. B. – PEDERSEN, J.G.*; Banrupcy Discrimination With Real Variables*. In : Journal of Business Finance & Accounting, 1994, vyd. 21, č. 4, s. 491 – 510.

POWERS, D. M. W., *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation* [online]. In : Journal of Machine Learning Technologies, 2011, vyd. 2, č. 1, s. 37 – 63. [cit. 2020-25-1]. Dostupné na: https://dspace2.flinders.edu.au/xmlui/bitstream/handle/2328/27165/Powers%20Evaluation.pdf?sequence=1&isAllowed=y

ROJAS, R.; *Neural Networks*, Berlin Springer Berlin Heidelberg, 2013. 502s. ISBN 978-3-642-61068-4.

SCHALKOFF, J. R.; *Artificial neural networks*. New York : McGraw-Hill College, 1997. 448 s. ISBN 0-07-057118-X.

SCHMIDHUBER J.; *Deep Learning in Neural Networks: An oveview*. In: Neural Networks, 2015, vyd. 61, s. 85 – 117. ISSN: 0893-6080.

SCHROEDER, K.; *After prediction* [online]. [cit.2019-11-20]. Dostupné na: http://www.kschroeder.com/weblog/after-prediction

STEIN, R., *Benchmarking Default Predicton Models:Pitfalls and Remedies In Model Validation*. In : Moody´s Investor Service, 2007.

TANG, J. - ALELYANI, S. - LIU, H., *Feature Selection For Classification: A Review*. In : Data Classification: Algorithms And Applications, 2014, vyd. 37.

TIERNEY B.; *Understanding, Building and Using Neural Network Machine Leaning Models using Oracle 18c* [online]. [cit.2019-11-11]. Dostupné na: https://developer.oracle.com/databases/neural-network-machine-learning.html

VAN SELST, M. – JOLICOEUR, P., *A Solution To The Effect Of Sample Size On Outlier Elimination*. In : Quarterly Journal of Experimental Psychology, 1994, vyd. 47, č. 3, s. 631-650.

WANG, S. – YAO, X., *Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models* [online]. In : IEEE Symposium on Computational Intelligence and Data Mining, 2009. [cit.2020-15-02]. Dostupné na : https://www.cs.bham.ac.uk/~syw/documents/papers/CIDMShuo.pdf

WANG, D. – ZHENG, T.. *Transfer learning for speech and language processing*. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015, s. 1225 – 1237.

WATSON, M. R.; *A Discrete Model Of Bacterial Metabolism.* In : Computer Applications in the Biosciences. 1986, vyd. 2, č. 1, s. 23-27.

WILSON, N. – OCHOTNICKÝ, P. – KÁČER, M.; *Creation and Destruction in Transition Economies: The SME Sector in Slovakia*. In : International Small Business Journal, 2016, vyd. 34, č. 5, s. 579 – 600.

YALE., *C/Floating Point* [online]. Publikované: 17.06.2014. [cit.2020-15-04]. Dostupné na: http://www.cs.yale.edu/homes/aspnes/pinewiki/C(2f)FloatingPoint.html

ZHANG, C. – WOODLAND, P. C.*; Parameterised Sigmoid and ReLU Hidden Activation Functions for DNN Acoustic Modelling* [online]. In: ISCA, 2015, vyd. 6, č. 10, s. 3224 – 3228. [cit.2020-10-1]. Dostupné na: http://cslt.riit.tsinghua.edu.cn/mediawiki/images/b/b5/Parameterised_Sigmoid_and_ReLU_Hidden_Activation_Functions_for_DNN_Acoustic_Modelling.pdf