Acta Informatica Pragensia, 2019, 8(1), 18–37
DOI: 10.18267/j.aip.123
Original Article

Dolovanie dát z bankového sektora

Data Mining from the Banking Sector´s Data

Anna Biceková [1], **Ľudmila Pusztová** [1]

Abstrakt

Predkladaný príspevok sa zaoberá problematikou bankrotov podnikov a definuje spôsoby akými je možné tomuto nežiadúcemu stavu predísť. V súčasnosti medzi tieto spôsoby patria hlavne moderné prístupy z oblasti získavania znalostí a dolovania v dátach, ktoré podnikom dokážu pomôcť v mnohých smeroch. V rámci praktickej aplikácie metód dolovania v dátach s cieľom predikovať budúci stav podniku, boli použité dáta finančných ukazovateľov poľských spoločností. V predkladanom článku sme využili algoritmy vhodné na predikciu bankrotov – rozhodovacie stromy, ktoré poskytujú jednoduchú interpretáciu výsledkov. V niektorých experimentoch sme využili aj metódy výberu atribútov, LASSO alebo PCA metódu. Postup práce sa riadi metodológiou CRISP-DM, ktorá ponúka popis dôležitých krokov potrebných pri rôznych analytických úlohách. Súčasťou článku je aj analýza súčasného stavu, ktorá predstavuje riešenia danej problematiky inými autormi. Po vyhodnotení všetkých modelov sme dospeli k záveru, že algoritmus C5.0 je na 97,07 % schopný predikovať zbankrotovanie respektíve nezbankrotovanie podniku, pričom použitie metód výberu atribútov nebolo potrebné.

Kľúčové slová: Predikcia bankrotov, dolovanie v dátach, CRIPS-DM metodológia, rozhodovacie stromy.

Abstract

This paper deals with the prediction of company bankruptcies and defines how this undesirable state can be prevented. Currently, these methods include modern approaches from the area of data mining that can help companies in many ways. In a practical application of data mining methods for predicting the future state of a company, financial indicators of Polish companies were used. In the analyses, we used algorithms suitable for bankruptcy prediction – decision trees that provide a simple interpretation of results. In some experiments, we also used attribute selection methods, LASSO, or the PCA method. The workflow is governed by the CRISP-DM methodology, which describes the important steps needed for different analytical tasks. Part of the article is an analysis of the current state, which presents solutions to this problem suggested by other authors. After evaluating all models, we concluded that the C5.0 algorithm is capable of predicting a company's bankruptcy or non-bankruptcy with 97.07 % accuracy, without the use of attribute selection methods.

Keywords: Bankruptcy prediction, Data mining, CRISP-DM methodology, Decision trees.

[1] Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic
✉ anna.bicekova@tuke.sk

1 Úvod

Pre rôzne druhy podnikov je z ekonomického hľadiska dôležitý neustály rast, spokojnosť zákazníkov, udržanie si postavenia na trhu a v neposlednom rade aj dosahovanie zisku, ktorý je výsledkom činnosti podniku. Všetky vyššie uvedené aspekty môžu byť ohrozené nepriaznivými činiteľmi a rozhodnutiami, ktoré spôsobujú problémy a v najhoršom prípade postupný úpadok podniku. Bankroty obchodných spoločností, výrobných podnikov alebo bánk, predstavujú negatívny vplyv v ekonomike krajiny.

Bankrot, krach alebo úpadok podniku je nežiaduci stav, kedy podnik dlhodobo nie je schopný plniť si svoje záväzky a zanikne. Bankrot je v spoločnosti všeobecne chápaný ako problém nielen národnej, ale aj svetovej ekonomiky. Dôkazom je mnoho finančných kríz - najznámejšia za posledné obdobie bola kríza v roku 2008 (Ivashina & Scharfstein, 2010), ktorej podľahli známe aj neznáme firmy a banky po celom svete. V súčasnej dobe, kedy využitie informačných technológií je nevyhnutnosťou pre prežitie ekonomických subjektov na trhu je dôležité, aby podniky vedeli aplikovať tieto metódy a týmto spôsobom dokázali predísť stavu bankrotu. Bankroty podnikov môžu byť spôsobené mnohými faktormi, ako sú zlé investičné rozhodnutia, zlé investičné prostredie, nízky cash flow a podobne (Dwyer & Tkac, 2009). Údaje plynúce z finančných ukazovateľov a vhodné metódy predikcie dokážu úspešne napovedať podniku, v akom stave sa môže v budúcnosti nachádzať resp. kedy je potrebné zaviesť príslušné opatrenia.

1.1 Vývoj predikcie bankrotov

Úpadok podniku je dlhodobo diskutovanou problematikou a predmetom výskumu viacerých autorov už od začiatku 19. storočia. Kľúčovými faktormi pre zistenie finančného zdravia podniku sú finančné ukazovatele získané z viacerých finančných výkazov firmy. Historický vývoj bankrotov môžeme rozčleniť na dve etapy, a to pred a po roku 1966 (Delina & Packová, 2013). Počas tohto obdobia autori dospeli k rôznym záverom v otázke, ktorý z finančných ukazovateľov vplýva resp. dokáže najpresnejšie detegovať finančné zdravie podniku, a tým lepšie predpovedať vznik potencionálneho bankrotu.

1.1.1 Obdobie pred rokom 1966

Prvé vedecké články a štúdie týkajúce sa analýzy predikcie bankrotov sú známe už od roku 1930. Jednotliví experti sa sústredili najmä na jednotlivé finančné pomery ziskových a krachujúcich firiem, ktoré medzi sebou porovnávali.

FitzPatrick (1932) porovnal 13 pomerových ukazovateľov ziskových a skrachovaných podnikov. Výsledkom bolo zistenie, že vo väčšine prípadov vykazovali úspešné firmy priaznivé pomery a naopak neúspešné firmy nepriaznivé pomery, a to všetko pri rešpektovaní vtedajších trendov v rámci vývoja finančných ukazovateľov. Porovnaním týchto podnikov dospel autor k záveru, že medzi dva dôležité a významné pomery, ktorým by sa mala venovať zvýšená pozornosť pri posudzovaní stavu podniku, patrí pomer *Vlastné imanie* a *záväzky* a pomer *Čistý zisk* a *Vlastné imanie*.

Smith a Winakor (1935) rozoberali finančné pomery 183 skrachovaných podnikov rôzneho druhu zamerania. Prišli k názoru, že oveľa schopnejším ukazovateľom pri predikcii finančných problémov firmy je pomer *Pracovný kapitál* a *Celkové aktíva*. Taktiež zistili, že čím bližšie je podnik k bankrotu, tým viac klesá pomerový ukazovateľ *Obežné aktíva* a *Celkové aktíva*.

V roku 1942 bola publikovaná ďalšia štúdia zaoberajúca sa malými výrobnými podnikmi (Merwin, 1942). Jej výsledkom bolo zistenie, že upadajúce podniky prejavujú začínajúce známky krachu už štyri alebo päť rokov pred konečným bankrotom. Významným

ukazovateľom bankrotu firmy je podľa neho pomer *Čistý pracovný kapitál* a *Celkové aktíva*, ukazovateľ *Krátkodobej likvidity* a pomer *Vlastného imania* a *Celkových záväzkov*.

1.1.2 Obdobie po roku 1966

Zieba et al. (2016) uvádzajú, že šesťdesiate roky 20. storočia priniesli zásadný prelom pri zisťovaní známok zlyhávania fungovania podniku. Začali sa používať *predikčné modely*, ktoré priniesli inovatívne postupy a obohatili tak statickosť doterajších metód. Nové predikčné modely dokázali stanoviť riziko bankrotu pre každú firmu v každom okamihu (Dallas, 2013).

Beaver v roku 1966 skúmal 79 skrachovaných a ziskových firiem z 38 oblastí priemyslu (Beaver, 1966). Testoval predikčnú silu jednotlivých pomerových ukazovateľov, napríklad *Čistý zisk* a *Celkové záväzky* mali najvyššiu predikčnú silu, a to presnosť 92 % jeden rok dopredu pred samotným zlyhaním podniku. Načrtol taktiež možnosť, že použitie viacerých ukazovateľov môže mať vyššiu predikčnú schopnosť ako použitie len jedného ukazovateľa, čím začal novú etapu vývoja predikčných bankrotových modelov.

Najznámejším autorom predikčného bankrotového modelu je profesor Edward Altman, ktorý je zároveň aj expertom v tejto oblasti. V roku 1968 vytvoril pomocou viacnásobnej diskriminačnej analýzy päť faktorový model pre výrobné podniky, známy ako *Altmanov Z-skóre model*, ktorého prvá podoba vyzerala nasledovne (Altman, 1968):

$$Z = 0{,}012.X_1 + 0{,}014.X_2 + 0{,}033.X_3 + 0{,}006.X_4 + 0{,}999.X_5, \quad (1)$$

kde ukazovateľ

X_1: *čistý pracovný kapitál/celkový majetok,*

X_2: *nerozdelený zisk/celkový kapitál,*

X_3: *zisk pred úrokmi a zdanením/celkový kapitál,*

X_4: *trhová hodnota vlastného kapitálu/cudzí kapitál,*

X_5: *obrat(tržby)/celkový kapitál.*

Podstatou modelu bolo pomocou diskriminačnej analýzy určiť jednotlivým pomerovým ukazovateľom váhy. Z-skóre znamená, že podniku bol určený stupeň podľa toho, či výsledok z rovnice *presiahol* deliacu hranicu (prosperujúci podnik) alebo *nepresiahol* (podniku hrozí bankrot). Altmanov model pre predikciu bankrotu na jeden rok dopredu bol úspešný na 95 %, na obdobie dva roky dopredu s úspešnosťou 72 % a na obdobie tri roky dopredu to bola úspešnosť 48 %.

Neskôr v 90. rokoch 20. storočia s vývojom moderných technológií už nebolo možné používať jednoduché lineárne modely a do výskumu predikcie vstúpila umelá inteligencia a strojové učenie.

1.2 Analýza existujúcich prác

Téma bankrotov a ich predikcia je zaujímavou a stále vyhľadávanou oblasťou pre mnohých odborníkov. Snažia sa nájsť najlepšie techniky, ktoré by mohli pomôcť firmám, manažérom alebo investorom pri odhadovaní stavu podniku na trhu v budúcnosti a tým predísť rôznym komplikáciám vedúcim k finančným problémom. Bankrotné a prediktívne modely sa považujú

za systémy včasného varovania založené na analýze vybraných ukazovateľov so schopnosťou detekovať hrozbu vo finančnom zdraví spoločnosti (Rybarova et al., 2016).

Za zmienku stojí spomenúť aj najznámejší krach štvrtej najväčšej investičnej banky *Lehman Brothers* počas globálnej krízy v 2008. Banka s takmer 160-ročnou tradíciou vyhlásila bankrot, ktorý spustil vlnu finančných problémov bánk a firiem po celom svete. Bankrot *Lehman Brothers* spôsobil najmä rizikový obchod s hypotekárnymi úvermi a spôsobil reťazovú reakciu s ďalšími činnosťami banky (Ringner, 2008). Počas uplynulého desaťročia použili viacerí autori vo svojich štúdiách rôzne techniky umelej inteligencie na predpovedanie bankrotu (Chen, 2011). V nasledujúcej časti príspevku sme sa venovali piatim štúdiám, ktoré priamo súvisia s našou vybranou dátovou množinou.

V práci čínskych vedcov Fan et al.(2017) bola riešená otázka predikcie bankrotov na rovnakom datasete poľských spoločností, aký sme použili aj my. Rozhodli sa riešiť hlavne problém súvisiaci so skreslenými dátami, na ktoré sa snažili aplikovať niekoľko algoritmov na zistenie anomálií. Tento krok by im pomohol k tomu, aby dáta s menšinovou triedou (trieda 1) model nezaradil do triedy prosperujúcich podnikov (trieda 0). Na zistenie takýchto odchýlok použili tri rôzne modely, konkrétne *Viacnásobnú Gaussovu distribúciu*, *One-class SVM* a *izolovaný les* (predpokladá, že anomálie majú 2 kvalitatívne vlastnosti – obsahujú malú časť objektov a hodnoty atribútov sú iné od normálnej triedy. Podstatou *1. experimentu* bolo použiť metódy detekcie anomálií na zistenie najlepšieho predikčného modelu. V *2. experimente* použili štyri modely kontrolovaného učenia. Dáta, ktoré mali k dispozícii rozdelili na tréningovú a testovaciu množinu v pomere 60:40 a *použili 5-násobnú krížovú validáciu.* Zistili, že model *izolovaný les* (isolation forest) v rámci metód odhaľovania odchýlok mal najlepšie výsledky, napr. v klasifikačnom prípade *1st Year* mal izolovaný les hodnotu *mean* (MN) **0.93** a neurónové siete zo všetkých 7 porovnávaných modelov najnižšiu hodnotu a to **0.84**. Poukázali, že aj modely odhaľovania anomálií môžu vyriešiť problém nevybalansovaných a skreslených dát, pretože modely kontrolovaného učenia často nedokážu vyriešiť problém.

Autori Hardinata et al. (2018) vo svojej práci taktiež používali rovnaké dáta. Zamerali sa hlavne na *umelé neurónové siete* (ANN) v rámci strojového učenia, konkrétne implementáciu **Jordan Recurrent Neural Networks** (JRNN) na klasifikáciu bankrotov. Prvotne klasifikovali poľské spoločnosti na zbankrotované a nezbankrotované pomocou JRNN a získaný model použili na predikciu bankrotov v ďalšom období. Neurónové siete považujú za vhodnejšie pre predikciu v oblasti dolovania v dátach, pretože ANN majú schopnosť vybrať dôležité informácie z veľkých dát. Výsledky zo štúdie ukázali, že priemerná klasifikačná *presnosť* modelu je **81,3785 %** s 5 neurónmi v skrytej vrstve.

Zieba et al. (2016) navrhli nový prístup v oblasti predikcie bankrotov, ktorý využíva metódu *extreme gradient boosting* (XGB) na učenie súboru rozhodovacích stromov. Cieľom výskumu bolo identifikovať najlepší klasifikačný model pre každý z 5 dátových množín. Do úvahy vzali 16 klasifikačných modelov, medzi ktoré patrili napr. *lineárna diskriminačná analýza*, *viacvrstvový perceptrón so skrytou vrstvou, logistická regresia*, *AdaBoost, náhodný les* (randomforest) a *extreme gradient boosting*. Na vyhodnotenie modelov použili *AUC krivku* a na testovanie kvality rôznych nástrojov trénovacích parametrov použili *10-násobnú krížovú validáciu.* Výsledky experimentu prezentovali v podobe *priemeru* a *štandardnej odchýlky* pre každý z piatich klasifikačných prípadov. Okrem porovnávaných výsledkov medzi predchádzajúcimi modelmi sa zamerali na výsledky modelov *XGBE*, *XGB* a *EXGB*, ktoré predstavujú rozšírenie XGB modelu. Na základe použitia *Wilcoxsonovho testu* p-hodnoty autori vyhodnotili, že najlepší klasifikačný model je EXGB s najvyššími hodnotami priemeru (MN) – **0.959**, pre porovnanie XGB malo hodnotu **0.945** alebo náhodný les s hodnotou **0.851**, z pomedzi všetkých 16 klasifikačných modelov.

V publikácií Moreira et al. (2018) sa autori zaoberali všeobecnou dátovou analýzou určenou pre študentov spísanou vo forme návodu, ako postupovať pri vlastných projektoch. Ako jeden z príkladov použili dáta poľských spoločnosti a postupovali metodológiou CRISP-DM. Počas fázy prípravy dát sa rozhodli riešiť problém nepomeru medzi triedami, a teda vymazania viac ako 800 riadkov (záznamov) s triedou 0 (firiem ktoré nezbankrotovali). Následne hodnoty atribútov normalizovali a chýbajúce hodnoty v atribútoch nahradili priemerom hodnôt daného stĺpca. Vo fáze modelovania si autori vybrali tri modely. K-najbližší susedov s k=15 a použili Euklidovskú vzdialenosť ako mieru vzdialenosti. Dostupné dáta rozdelili na ténovaciu a testovaciu v pomere 70:30 a ako vstupné atribúty použili *Attr6, Attr11, Attr24, Attr27* a *Attr60*. Použili taktiež aj algoritmus rozhodovacieho stromu C4.5 a algoritmus random forest, v ktorom vygenerovali500 stromov. Pre všetky tieto algoritmy použili 10-násobnú krížovú validáciu. Z dosiahnutých výsledkov vyplynulo, že algoritmus *random forest* mal najlepšiu presnosť, a to 98,47 %. Pre porovnanie presnosť algoritmu C4.5 bola 98,30 % a k-NN metódy 98,21 %.

Nagaraj a Sridhar (2015) sa zaoberajú vytvorením najvhodnejšieho modelu na predikciu bankrotov, ktorý neskôr použili na vývoj systému na podporu rozhodovania. Použité dáta nie sú tie isté, aké používame my, líšia sa najmä počtom atribútov (7 stĺpcov) a nie sú to pomery resp. finančné ukazovatele, ale **kvalitatívne vlastnosti podniku** ako napríklad: *konkurencia* (competitiveness), *manažérske riziko* (management risk), *dôveryhodnosť* (credibility), *priemyselné riziko* (industrial risk), *finančná flexibilita* (financial flexibility), *prevádzkové riziko* (operating risk) a *stav bankrotu* označený ako *class*. Prvých 6 atribútov má nominálne hodnoty – triedy s úrovňami *positive*, *average* a *negative*, tieto hodnoty potom transformovali na numerické atribúty ako *1*, *0.5* a *0*. Pre atribút class sú to hodnoty *non-bankruptcy* alebo *bankruptcy*. Pomer triedy class bol vyrovnaný, 107 záznamov ako *bankrupt* a 143 záznamov ako *non-bankrupt*. Aj napriek nízkemu počtu záznamov autori nemali problém s nevyrovnanosťou dát, ako je to v našom prípade. Dáta si rozdelili na tréningovú a testovaciu množinu v pomere 70:30. Pre zistenie najpresnejšieho modelu autori porovnávali niekoľko algoritmov, ktoré zahŕňali *logistickú regresiu*, *náhodný les*, *naivný Bayes*, *SVM metódu* a *neurónové siete*. Na overenie úspešnosti každého klasifikátora použili 10-násobnú krížovú validáciu. Jednotlivé algoritmy porovnávali na základe štyroch metrík výkonnosti, a to *úspešnosť* v %, pomery *true-positive* a *true-negative* a *presnosť* modelov. Z výsledkov, ktoré získali bol najúspešnejší model *SVM* s úspešnosťou **99,6 %**, *neurónové siete* s úspešnosťou **98,6 %** a tretím najlepším modelom bol *naivný Bayes* s úspešnosťou **98,3 %**. Celkovo všetky modely mali vysoké percento presnosti aj úspešnosti a to nad 90 %.

Medzi ďalšie štúdie, ktoré pracovali s údajmi o bankrotoch patrili napríklad (Zhou & Elhag, 2007) ktorí na skúmanie vzťahov medzi závislými premennými používali chi-kvadrát test; (Jardin & Severin 2011), ktorí použili metódy k-najbližších susedov na predikciu bankrotov alebo Atiya (2001), ktorý vytvoril model predikcie bankrotov pomocou neurónových sietí. Úspešnosť metód dolovania v dátach na predikciu bankrotov potvrdzuje aj (Bellovary et al., 2007), ktorá v Tabuľke 1 zhrnula počet najpoužívanejších modelov od 60. rokov 20. storočia.

Tab. 1. *Najpoužívanejšie modely od 60. rokov 20. storočia. Zdroj: (Bellovary et al., 2007).*

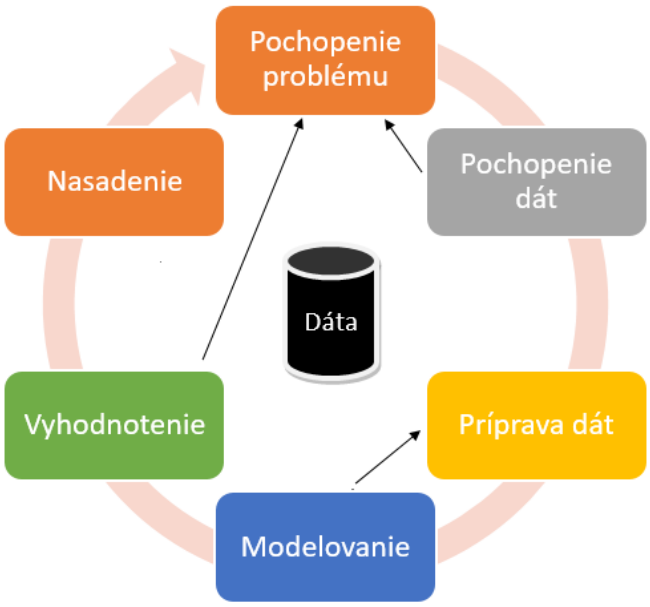
	Diskriminačná analýza	**Logit model**	**Probit model**	**Neurónové siete**	**Ďalšie**
60. roky	2	0	0	0	1
70. roky	22	1	1	0	4
80. roky	28	16	3	1	7

	Diskriminačná analýza	Logit model	Probit model	Neurónové siete	Ďalšie
90. roky	9	16	3	35	11
Začiatok 21. storočia	2	3	0	4	3

2 Metódy a materiály

Základným pojmom pre správne pochopenie získavania nových vedomostí je *objavovanie znalostí*, ktorý popisuje proces nájdenia znalostí a užitočných informácií z veľkého objemu dát, ktoré nám pomáhajú pri rozhodovaní v rôznych situáciách (Moreira et al., 2018). Keďže je tento proces *iteratívny* a *interaktívny*, je možné jednotlivé kroky procesu v prípade potreby opakovať. Dôležitou charakteristikou je aj jeho *multidisciplinárnosť*, ktorá napovedá o rôznych oblastiach, kde sa proces objavovania znalostí môže uplatniť (hlavne štatistika, databázové systémy a umelá inteligencia). Na štandardizáciu procesu objavovania znalostí vznikali postupom času rôzne metodiky, ako napríklad *SEMMA, 5A, ASUM-DM* alebo *CRISP-DM*.

Z vyššie vymenovaných metodík patrí medzi najznámejšiu CRISP-DM (*Cross Industry Standard Process for Data Mining*), ktorá poskytuje prehľad o celom životnom cykle vybranej analytickej úlohy (Paralič, 2003, str. 5). Samotná metodika bola spustená v roku 1996 v rámci Európskeho výskumného projektu. Považuje sa za štandardný model a nenadväzuje na konkrétny softvér. Využiteľnosť CRISP-DM metodiky sa primárne orientuje na rozsiahle projekty, kde pomáha správne viesť celé analytické projekty rýchlejšie, efektívnejšie a s použitím menej finančných prostriedkov. Okrem návrhu postupu ponúka aj sprievodcu nad potencionálnymi problémami, ktoré sa môžu vyskytovať v reálnych úlohách. Celý postup metodiky je možné popísať na základe nasledujúceho modelu.



Obr. 1. *Fázy modelu CRISP-DM. Zdroj: (Paralič, 2003).*

Tento model je idealizovaný sled udalostí, ktorý sa skladá zo šiestich fáz, konkrétne *pochopenie problému, pochopenie dát, príprava dát, modelovanie, vyhodnotenie* a *nasadenie*. Jednotlivé šípky označujú najdôležitejšie a najčastejšie závislosti medzi týmito fázami. Tento model

poskytuje návod krok pod kroku, avšak ich poradie nie je presné. V prípade potreby je možné vrátiť sa späť na predchádzajúce kroky a opakovať určité akcie. Paralič (2003) uvádza, že z jednotlivých krokov CRISP-DM metódy je *najdôležitejšou fázou* **pochopenie cieľa** a *časovo najnáročnejšou fázou* **príprava dát.**

- **Pochopenie problému** – prvá a najdôležitejšia etapa procesu. V tomto kroku je dôležité pochopiť daný problém *z obchodného hľadiska* a z *hľadiska dolovania v dátach.* V úvode je potrebné zhodnotiť súčasnú situáciu riešenej problematiky, pomocou ktorej sa dokážu odhaliť faktory, prostriedky alebo obmedzenia, ktoré môžu ovplyvniť celkový výsledok projektu. Okrem toho je nutné stanoviť si *kritériá úspešnosti*, ktoré napomôžu pri vyhodnocovaní daného riešenia. V rámci tohto kroku je vhodné spomenúť *vybrané techniky dolovania v dátach* pre konkrétnu úlohu a jednotlivé *vyhodnocovacie metódy.*
- **Pochopenie dát** – popisuje proces získavania potrebných dát, ako aj ich základných informácií. Dôležitými informáciami pre dôkladné pochopenie dát je napríklad *počet dostupných záznamov, označenie stĺpcov* a *ich význam, rozsah hodnôt* (minimálna a maximálna hodnota), *priemer hodnôt* v každom atribúte, *typy jednotlivých dát* (numerický, binárny) a pod. (Moreira et al., 2018). Rovnako dôležité je aj zistenie *kvality dát*, teda ich konzistencia a možnosť výskytu chýbajúcich hodnôt. Poslednou úlohou je *prieskum dát* pomocou štatistických analýz a jednoduchých vizualizácií, ktoré poskytnú informácie o vzájomných vzťahoch medzi atribútmi, distribúciu kľúčových premenných alebo rôzne jednoduché štatistiky.
- **Príprava dát** – zahŕňa činnosti, ktoré vedú k vytvoreniu výslednej dátovej množiny adekvátnej k stanovenému cieľu úlohy. Pri *výbere dát* je dôležité vybrať tie, ktoré sa následne používajú na samotnú analýzu a modelovanie. Výber sa vzťahuje na atribúty ako aj na záznamy. *Čistenie dát* závisí od kvality dostupných dát. Keďže dáta prichádzajú z reálneho prostredia, nikdy nie sú bezchybné (môžu byť zašumené, nekonzistentné, obsahovať chýbajúce/prázdne hodnoty). Práve úloha čistenia dát sa zaoberá týmito problémami, v prípade chýbajúcich hodnôt je nutné ich doplniť alebo odstrániť. V úlohe *konštrukcie dát* dochádza napríklad ku transformovaniu hodnôt atribútov, generovanie nových záznamov a atribútov. *Integrácia dát* sa vykonáva vtedy, ak dáta pochádzajú z rôznych tabuliek a je potrebné ich zlúčiť do jednej.
- **Modelovanie** – dochádza k aplikovaniu najlepších a najvhodnejších techník pre modelovanie na pripravené dáta. Keďže väčšina techník ma odlišné požiadavky na dáta, interakcia s fázou prípravy je nevyhnutná. Po vytvorení modelov na pripravenej množine je dôležité jednotlivé modely ohodnotiť pomocou rôznych kritérií, ktoré sa uvádzajú vo fáze pochopenia problému.
- **Vyhodnotenie** – zameriava sa na vyhodnotenie výsledkov modelov vo všeobecnosti z pohľadu obchodných cieľov stanovených na začiatku procesu. Sústredí sa hlavne na overenie, či boli stanovené ciele splnené. V tomto kroku je vhodné zamerať sa aj na neúspešné úlohy, ktoré boli počas modelovania zanedbané. Rovnako sa stanovujú aj ďalšie kroky – buď sa rozhodne o ukončení projektu alebo sa prejde do nasledujúcej fázy.
- **Nasadenie** – ak sa vo váze vyhodnotenia rozhodne o pokračovaní projektu, je dôležité dosiahnuté výsledky práce upraviť do zrozumiteľnej podoby pre prijímateľa. Taktiež tu dochádza k monitorovaniu a organizácií výsledkov. Výstupom fázy je celkové posúdenie projektu a celkové zjednotenie z pohľadu dosiahnutých cieľov, vzniknutých problémov alebo potencionálnych krokov, ktoré by sa mohli v rámci riešenej problematiky ešte vytvoriť.

2.1 Použité metódy

Nasledujúca časť príspevku sa venuje stručnému prehľadu metód použitých vo fáze prípravy dát a modelovania, ktoré boli aplikované pri analýze nami vybranej vzorky.

2.1.1 Rozhodovacie stromy

Rozhodovací strom je model, ktorý je ľahko pochopiteľný a jasne prezentuje získané výsledky. Využíva sa hlavne pri klasifikačných úlohách, ktorých cieľom je predikovať nominálny cieľový atribút. Učenie modelu rozhodovacieho stromu prebieha na základe záznamov trénovacej množiny. Vytvorený model je následne použitý v rámci záznamov, ktoré ešte nemajú stanovenú cieľovú triedu. Moreira et al. (2018) uvádzajú, že medzi základné charakteristiky rozhodovacieho stromu patrí *koreňový uzol* (počiatočný bod, od ktorého sa vetvia ďalšie uzly), *medziľahlý uzol* (predstavuje vybraný testovací atribút), *listový uzol* (konečný bod, ktorý obsahuje klasifikovanú hodnotu cieľového atribútu) a *hrana* (test na atribút z predchádzajúceho uzla). Po vygenerovaní takéhoto modelu je možné vyčítať rozhodovacie pravidlá, ktoré začínajú v koreňovom uzle a končia listom.

Na zvýšenie efektívnosti a presnosti modelu sa používajú rôzne techniky orezávania ako *pre-prunning* (orezanie modelu počas jeho generovania) a *post-prunning* (orezanie modelu až po jeho vytvorení).

V praxi sa využívajú rôzne typy algoritmov rozhodovacích stromov, ako napríklad *C4.5, C5.0, CART* alebo *random forest.*

- C4.5 – poskytuje veľmi dobré výsledky, pracuje s numerickými atribútmi, dokáže pracovať s chýbajúcimi hodnotami niektorých atribútov a je schopný využívať techniky orezávania.
- C5.0 – je zlepšená verzia C4.5 algoritmu. Popri výhodách z C4.5 poskytuje rýchlejší výpočet, lepšiu prácu s pamäťou a taktiež aj schopnosť pracovať so spojitými a diskrétnymi hodnotami.
- CART – vie pracovať s extrémnymi hodnotami, generuje binárne ako aj regresné stromy.
- Random forest – po vygenerovaní viacerých rozhodovacích stromov dochádza k ich spojeniu z dôvodu vylepšenia presnosti modelu.

Po vygenerovaní rozhodovacieho stromu je dôležité daný model ohodnotiť. V prípade použitia viacerých modelov/algoritmov je vhodné ich medzi sebou porovnať. Vo všeobecnosti sa na vyhodnotenie modelov používajú metriky vypovedajúce o kvalite použitých modelov (Moreira et al., 2018), ako napríklad:

- *Presnosť* – meria podiel správne klasifikovaných pozitívnych prípadov voči všetkým prípadom, ktoré boli pozitívne klasifikované (správne aj nesprávne).
- *Úspešnosť* – meria celkovú úspešnosť klasifikácie modelu (podiel správne klasifikovaných prípadov voči všetkým)
- *Chyba* – meria podiel nesprávne klasifikovaných prípadov voči všetkým.
- *Návratnosť* – meria podiel správne klasifikovaných pozitívnych prípadov voči všetkým pozitívnym prípadov (správne aj nesprávne klasifikovaných).
- *Špecifickosť* – meria podiel správne klasifikovaných negatívnych prípadov voči všetkým negatívnym prípadov (správne aj nesprávne klasifikovaných).

- *AUC* (Area Under the Curve) – oblasť pod krivkou – kvantifikuje celkovú schopnosť modelu rozlišovať medzi správne a nesprávne klasifikovanými prípadmi. Čím je hodnota AUC vyššia, tým vie model lepšie klasifikovať prípady.

Hodnoty jednotlivých metrík je možné vyčísliť z kontingenčnej tabuľky (viď Tab. 2), ktorá obsahuje presné počty správne a nesprávne klasifikovaných tried pre každý model.

Tab. 2. *Kontingenčná tabuľka. Zdroj: Autori.*

		Skutočná trieda	
		P	**N**
Predpovedaná trieda	**P**	TP	FN
	N	FP	TN

Hodnota *TP* (True Positive) predstavuje počet všetkých správne klasifikovaných pozitívnych prípadov; *TN* (True Negative) predstavuje počet správne klasifikovaných negatívnych prípadov; *FP* (False Positive) predstavuje počet nesprávne klasifikovaných pozitívnych prípadov a *FN* (False Negative) predstavuje počet nesprávne klasifikovaných negatívnych prípadov. Platí, že na hlavnej diagonále sa nachádzajú počty *správne klasifikovaných prípadov* a na vedľajšej diagonále počty *nesprávne klasifikovaných prípadov*.

2.1.2 Redukcia počtu atribútov

Redukcia počtu atribútov sa vykonáva vo fáze prípravy dát a je veľmi úspešná v prípade, že dátová množina obsahuje veľký počet dimenzií. Zredukovaný počet dimenzií umožní lepšiu prácu a väčší prínos pri modelovaní. Existuje niekoľko metód pre redukciu dimenzií (Fonti, 2017):

- *PCA* – metóda redukcie počtu dimenzií vhodná pre množiny s veľkým počtom atribútov a vzájomnou koreláciou. Zabezpečuje výber tých atribútov, ktoré ponúkajú najväčší informačný prínos a zároveň sú navzájom lineárne nezávislé. Táto metóda prvotne vyhľadá komponenty, ktoré sú *eigenvektory* predstavujúce smer najväčšieho rozptylu dát. Každý *eigenvektor* má prislúchajúcu *eigenhodnotu*, pričom za najviac prínosný sa považuje komponent s *eigenhodnotou* väčšou ako 1.
- *LASSO* – patrí medzi regresnú metódu a používa sa nie len na výber atribútov ale aj na regularizáciu. Metóda aplikuje proces penalizácie a zabezpečí, aby dôležité atribúty pre modelovanie mali nenulovú hodnotu. Dôležitým faktorom je parameter λ, ktorý kontroluje silu penalizácie. Čím je vyššia hodnota λ, tým viac atribútov bude mať nulovú hodnotu a dôjde k redukcii dimenzií.

Korelácie medzi atribútmi – na redukciu počtu atribútov je možné využiť aj korelačné vzťahy alebo vzťahy závislostí. Metóda patrí medzi *filtračné metódy* a je založená na rôznych štatistických testoch, ako napríklad Pearsonov korelačný koeficient (pre numerické atribúty) alebo *Chi-kvadrát test* (pre kategoriálne atribúty). Závislosť sa porovnáva medzi dvoma atribútmi a z výsledných korelačných koeficientov c je možné následne zistiť silu korelácie (Moreira et al., 2018), ktorá môže byť *slabá* ($c < 0.5$), *stredná (* $0.5 \leq c < 0.8$) alebo *silná* ($0.8 \leq c$).

2.1.3 Použité dáta

Dáta použité v tomto príspevku pochádzajú z repozitára Machine Learning Repository (Tomczak, 2016), ktoré pozostávajú z piatich dátových súborov. Dátové súbory predstavujú hodnoty finančných ukazovateľov poľských spoločností v priebehu piatich rokov. Každý dátový súbor obsahuje 64 finančných ukazovateľov (v súboroch označené ako „Attr"), pričom počet záznamov je odlišný. Jednotlivé finančné ukazovatele sú numerického typu, s výnimkou posledného cieľového atribútu, ktorý je binárny. Každý záznam (riadok) predstavuje konkrétnu poľskú spoločnosť, ktorých názov však nie je známi. Na nasledujúcom Obrázku 1 sa nachádza výrez dátovej množiny. Podrobnejší popis jednotlivých súborov sa nachádza v podkapitole 3.2.

Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8	Attr9	Attr10	Attr11	Attr12	Attr13	Attr14
0.2023500	0.46500	0.240380	1.51710	-14.5470	0.51069000	0.253660	0.91816	1.15190	0.42695	0.253660	0.545610	0.178650	0.253660
0.0300730	0.59563	0.186680	1.33820	-37.8590	-0.00031864	0.041670	0.67890	0.32356	0.40437	0.042199	0.075493	0.145630	0.041670
0.2578600	0.29949	0.665190	3.22110	71.7990	0.00000000	0.318770	2.33200	1.67620	0.69841	0.324530	1.064400	0.197450	0.318770
0.2271600	0.67850	0.042784	1.08280	-88.2120	0.00000000	0.285050	0.47384	1.32410	0.32150	0.381070	0.551560	0.227940	0.285050

Obr. 1. *Výrez dátovej množiny. Zdroj: Autori.*

3 Analytický proces

Celý analytický proces bol realizovaný pomocou metodológie CRISP-DM, pričom na vykonanie jednotlivých fáz bol použitý programovací jazyk R.

3.1 Pochopenie problému

Ako sme už naznačili, bankrot je nežiaduci stav, kedy podnik dlhodobo nie je schopný plniť si svoje záväzky a zanikne. Predikcia bankrotov sa usiluje o zníženie tohto vplyvu tým, že dokáže na určité obdobie dopredu zhodnotiť „finančné zdravie" firmy.

Biznis cieľom tejto úlohy bolo pomôcť napr. investorovi rozhodnúť sa, či si pre svoje investičné plány vyberie danú firmu alebo nie. Banky poskytujú úver na základe predpovedania stavu danej firmy na trhu. Situácia na trhu a v samotnej ekonomike sa neustále mení a pre podnik resp. manažment podniku je veľmi dôležité vedieť, v akom stave sa podnik bude nachádzať po troch, štyroch alebo piatich rokoch na základe súčasných finančných ukazovateľov.

Cieľom úlohy z **pohľadu dolovania v dátach** bolo potrebné nájsť model, ktorý na základe finančných ukazovateľov poľských spoločností bude schopný predikovať *zbankrotovanie (1)* alebo *nezbankrotovanie (0)* firmy. Na tvorbu tohto modelu sme použili konkrétne úlohy dolovania v dátach – *klasifikáciu* a *asociačné pravidlá.* Jednotlivé klasifikačné modely sme prvotne generovali na trénovaceja následne vyhodnotili na testovacej množine.

Na hodnotenie modelov sme používali vygenerovanú kontingenčnú tabuľku, z ktorej bolo možné vyčítať počet zbankrotovaných/nezbankrotovaných firiem a taktiež aj kvalitatívne metriky ako *presnosť*, *hodnotu AUC oblasti pod krivkou, ROC krivka*, *úspešnosť klasifikačného modelu* alebo *chyba klasifikačného modelu*. Okrem kvality modelu nás zaujímala aj *dôležitosť atribútov* v jednotlivých modelov.

3.2 Pochopenie dát

Dátová množina vybraná na predstavenú úlohu pochádzala z repozitára dát (Tomczak, 2016) určených na dátovú analýzu. Autori, ktorí sa podieľali na vytvorení týchto dát, si nevybrali túto

oblasť náhodne. Zaujal ich hlavne fakt, že v Poľsku v roku 2004 (Zieba et al., 2016) mnoho podnikov výrobného charakteru zbankrotovalo. Práve tento fakt podnietil ich záujem vytvoriť dáta, ktoré by boli použiteľné na výskum v problematike predikcie bankrotov.

Časový rámec zozbieraných dát sa pohybuje od roku 2000 do roku 2013 a sú rozdelené do piatich množín na základe obdobia predpovede bankrotu. Každá množina obsahuje rôzny počet záznamov spoločností, rovnaký počet a význam atribútov avšak rôzne hodnoty jednotlivých atribútov. Počet atribútov je **65 atribútov**, z ktorých prvých 64 predstavuje finančné ukazovatele, ako napr. čistý zisk k celkovým aktívam a pod. Posledný atribút *class* je trieda (faktor) označujúci stav bankrotu, teda 0 ako podnik v stave „nezbankrotuje“ a 1 ako podnik v stave „zbankrotuje“. Tabuľka 3 popisuje hodnoty atribútov z datasetu *1st Year*.

Tab. 3. *Hodnoty atribútov datasetu 1st Year. Zdroj: Autori.*

Názov atribútu	**Popis**	**Typ atribútu**	**Rozsah hodnôt v atribúte**	
			Minimum	**Maximum**
Attr1	čistý zisk/celkové aktíva	numerický	-256,89	94,28
Attr2	celkové záväzky/celkové aktíva	numerický	-72,162	441,5
Attr3	pracovný kapitál/celkové aktíva	numerický	-440,5	1
Attr4	obežné aktíva/krátkodobé záväzky	numerický	0	1017,8
Attr5	[(hotovosť + krátkodobé CP + pohľadávky - krátkodobé záväzky)/(prevádzkové náklady - odpisy)] * 365	numerický	-2 722100	990900
Attr6	nerozdelený zisk/celkové aktíva	numerický	-397,89	303,67
Attr7	EBIT / celkové aktíva	numerický	-189,56	453,77
Attr8	účtovná hodnota VI/celkové pasíva	numerický	-141,41	1452,2
Attr9	tržby/celkové aktíva	numerický	0	3876,1
Attr10	VI/celkové aktíva	numerický	-440,55	1099,5
Attr11	(hrubý zisk + mimoriadne položky + finančné náklady)/celkové aktíva	numerický	-189,45	453,78
Attr12	hrubý zisk/krátkodobé záväzky	numerický	-23,207	331,46
Attr13	(hrubý zisk + odpisy)/tržby	numerický	-607,42	13315
Attr14	(hrubý zisk + úrok)/celkové aktíva	numerický	-189,56	453,77
Attr15	(celkové pasíva * 365)/(hrubý zisk + odpisy)	numerický	-5611900	3599100
Attr16	(hrubý zisk + odpisy)/celkové pasíva	numerický	-42,322	405,33
Attr17	celkové aktíva/celkové pasíva	numerický	-0,4129	1529,9
Attr18	hrubý zisk/celkové aktíva	numerický	-189,56	453,77
Attr19	hrubý zisk/tržby	numerický	-622,06	2156,8
Attr20	(zásoby * 365)/tržby	numerický	0	7809200
Attr21	tržby(n) / tržby(n-1)	numerický	-1325	27900

Názov atribútu	Popis	Typ atribútu	Rozsah hodnôt v atribúte	
			Minimum	Maximum
Attr22	zisk z prevádzkových činností/celkové aktíva	numerický	-216,8	454,64
Attr23	čistý zisk/tržby	numerický	-634,59	2156,8
Attr24	hrubý zisk (za 3 roky)/celkové aktíva	numerický	-189,56	831,66
Attr25	(VI - ZI)/celkové aktíva	numerický	-459,56	1353,3
Attr26	(čistý zisk + odpisy)/celkové pasíva	numerický	-21,793	612,88
Attr27	zisk z prevádzkových činností/finančné náklady	numerický	-14790	2040800
Attr28	pracovný kapitál / dlhodobý majetok	numerický	-490,09	1570
Attr29	logaritmus celkových aktív	numerický	0,1761	9,3861
Attr30	(celkové pasíva - hotovosť)/tržby	numerický	-149,07	152860
Attr31	(hrubý zisk + úrok)/tržby	numerický	-622,06	2156,8
Attr32	(krátkodobé záväzky * 365)/náklady na predané výrobky	numerický	0	351630
Attr33	prevádzkové náklady/krátkodobé záväzky	numerický	0	884,2
Attr34	prevádzkové náklady/celkové pasíva	numerický	-280,26	884,2
Attr35	zisk z predaja/celkové aktíva	numerický	-169,47	445,47
Attr36	celkové tržby/celkové aktíva	numerický	0	3876,1
Attr37	(obežné aktíva - zásoby)/dlhodobé záväzky	numerický	-525,5	398920
Attr38	konštantný kapitál/celkové aktíva	numerický	-440,55	1099,5
Attr39	zisk z predaja/tržby	numerický	-701,63	2156,5
Attr40	(obežné aktíva - zásoby -pohľadávky)/krátkodobé záväzky	numerický	-101,27	1014,6
Attr41	celkové pasíva/[(zisk z prevádzkovej činnosti + odpisy)*(12*365)]	numerický	-77,791	813,14
Attr42	zisk z prevádzkovej činnosti/tržby	numerický	-701,63	2156,8
Attr43	ročné pohľadávky + obrat zásob v dňoch	numerický	0	30393000
Attr44	(pohľadávky*365)/tržby	numerický	0	22584000
Attr45	čistý zisk/zásoby	numerický	-256230	5986,8
Attr46	(obežné aktíva - zásoby)/krátkodobé záväzky	numerický	-101,26	1017,8
Attr47	(zásoby*365)/náklady na predané výrobky	numerický	0	62233
Attr48	EBITDA (zisk z prevádzkovej činnosti - odpisy)/celkové aktíva	numerický	-218,42	405,59
Attr49	EBITDA/tržby	numerický	-9001	31,639
Attr50	obežné aktíva/celkové pasíva	numerický	0	261,5

Názov atribútu	Popis	Typ atribútu	Rozsah hodnôt v atribúte	
			Minimum	Maximum
Attr51	krátkodobé záväzky/celkové aktíva	numerický	0	441,5
Attr52	(krátkodobé záväzky * 365)/náklady na predané výrobky	numerický	0	453,96
Attr53	VI/dlhodobý majetok	numerický	-130,47	180440
Attr54	konštantný kapitál/dlhodobý majetok	numerický	-122,03	180440
Attr55	pracovný kapitál	numerický	-800470	4398400
Attr56	(tržby – náklady na predané výrobky)/tržby	numerický	-1108300	1
Attr57	(obežné aktíva - zásoby – krátkodobé záväzky)/(tržby – hrubý zisk - odpisy)	numerický	-315,37	126,67
Attr58	celkové náklady/celkové tržby	numerický	0	1108300
Attr59	dlhodobé záväzky/VI	numerický	-327,97	119,58
Attr60	tržby/zásoby	numerický	0	2137800
Attr61	tržby/pohľadávky	numerický	0	21110
Attr62	(krátkodobé záväzky * 365)/tržby	numerický	0	25016000
Attr63	tržby/krátkodobé záväzky	numerický	0	1042,2
Attr64	tržby/dlhodobý majetok	numerický	0	294770
class	stav bankrotu	binárny	0	1

Distribúcia záznamov do tried je pre každý dataset rôzna. Prvý dataset **1st Year** obsahuje finančné pomery z 1. roku predpovedaného obdobia a vyjadruje stav bankrotu po 5 rokoch. Dataset obsahuje 7027 podnikov (záznamov), z ktorých je 6756 podnikov nezbankrotovaných a 271 podnikov zbankrotovaných. Druhý dataset **2nd Year** obsahuje finančné pomery z 2. roku predpovedaného obdobia a vyjadruje stav bankrotu po 4 rokoch. Dataset obsahuje 10 173 podnikov, z ktorých je 9773 podnikov nezbankrotovaných a 400 podnikov zbankrotovaných. Tretí dataset **3rd Year** obsahuje finančné pomery z 3. roku predpovedaného obdobia a vyjadruje stav bankrotu po 3 rokoch. Dataset obsahuje 10 503 podnikov, z ktorých je 10 008 podnikov nezbankrotovaných a 495 podnikov zbankrotovaných. Štvrtý dataset **4th Year** obsahuje finančné pomery zo 4. roku predpovedaného obdobia a vyjadruje stav bankrotu po 2 rokoch. Dataset obsahuje 9 792 podnikov, z ktorých je 9277 podnikov nezbankrotovaných a 515 podnikov zbankrotovaných. Piaty dataset **5th Year** obsahuje finančné pomery z 5. roku predpovedaného obdobia a vyjadruje stav bankrotu po 1 roku. Dataset obsahuje 5 910 podnikov, z ktorých je 5500 podnikov nezbankrotovaných a 410 podnikov zbankrotovaných.

V tejto fáze sme zisťovali *závislosti medzi jednotlivými numerickými atribútmi* navzájom. Pre tento účel nám poslúžila *korelačná matica* obsahujúca korelačné koeficienty a na samotný výpočet metóda *Pearsonovho korelačného koeficientu*. Nakoľko hodnoty závislostí boli rôzne, zamerali sme sa hlavne na hodnoty závislostí <0,8;1> a <-0,8;-1>. Korelačné vzťahy sme robili zvlášť na samostatných množinách. Medzi najvyššie závislosti, ktoré boli spoločné pre všetkých päť dátových množín, patrili napríklad:

- **Attr1:** čistý zisk/celkové aktíva a **Attr7:** EBIT/celkové aktíva,
- **Attr7:** EBIT/celkové aktíva a **Attr14:** (hrubý zisk + úrok)/tržby,
- **Attr2:** celkové záväzky/celkové aktíva a **Attr10:** VI/celkové aktíva,
- **Attr32:** (krátkodobé záväzky * 365)/náklady na predané výrobky a **Attr52:** (krátkodobé záväzky * 365)/náklady na predané výrobky.

Pre každú dátovú množinu sme zistili aj základné štatistiky numerických atribútov ako napríklad *minimálna* a *maximálna hodnota*, *priemer hodnôt* a *medián*, *1 kvartil*, *3 kvartil* a samozrejme aj *počet chýbajúcich hodnôt* (NA) pre každý atribút. Z analýzy chýbajúcich hodnôt môžeme konštatovať, že sa v našich datasetoch vyskytovali atribúty s veľkým počtom týchto hodnôt. Medzi atribúty s najvyšším počtom NA hodnôt patrili napríklad Attr21, Attr27, Attr45 a Attr60. Atribút Attr37 však obsahoval najvyšší počet chýbajúcich hodnôt **vo všetkých piatich dátových množinách.**

3.3 Príprava dát

V rámci tejto fázy sme sa zamerali na vytvorenie dátovej množiny, ktorá by bola v podobe vhodnej na modelovanie a získanie kvalitných a relevantných výsledkov pre našu úlohu. Prvotne sme sa rozhodli *spojiť jednotlivé datasety* do jednej množiny, keďže počet a typ atribútov bol v samotných množinách rovnaký. Ich spojením sme dostali 43 405 záznamov. Následne sme sa zamerali na *problém chýbajúcich hodnôt*. Atribút s najvyšším počtom chýbajúcich hodnôt - Attr37 *(obežné aktíva - zásoby) / dlhodobé záväzky* – sme sa rozhodli vymazať. Pri atribútoch s nižším počtom chýbajúcich hodnôt sme hodnoty nahrádzali dvoma spôsobmi, konkrétne pomocou **metódy k-NN** s počtom k=4 a **priemerom hodnôt** daného atribútu. Podobným spôsobom, teda nahradením chýbajúcich hodnôt priemerom, postupovali aj autori v štúdií Moreira et al. (2018).

Keďže naše dáta obsahovali veľký počet atribútov, použili sme aj niektoré *metódy výberu atribútov,* ktoré budú mať najväčší prínos na predpovedanie bankrotov do budúcnosti. Použili sme **metódy *PCA, LASSO*** a taktiež **korelačné vzťahy metri atribútmi.** V Tabuľke 4 sú uvedené počty atribútov vybraných jednotlivými metódami *feature selection* na množinách s obidvoma úpravami NA hodnôt.

Tab. 4. *Počet atribútov vybraný metódami feature selection. Zdroj: Autori.*

Spôsob úpravy NA hodnôt	Metóda	Počet atribútov
k-NN	Korelácie	46
	PCA	23
	LASSO	34
priemer	Korelácie	37
	PCA	23
	LASSO	34

3.4 Modelovanie

Vo fáze modelovania sme na vytvorené množiny aplikovali algoritmy rozhodovacích stromov, pričom sme si túto fázu rozdelili na 4 experimenty z hľadiska použitia *typu metódy na výber atribútov*. Pri vyhodnotení uvádzame len výsledné metriky najlepších modelov.

1. **Experiment** – výber atribútov pomocou metódy **LASSO** a použitie algoritmov rozhodovacích stromov *C4.5*, *C5.0*, *random forest* a *CART*. Vytvorených bolo 32 modelov v rôznych pomeroch trénovacej a testovacej množiny. Z dôvodu nevybalansovaných dát v atribúte *class,* sme na trénovacej množine pre naučenie modelu použili tzv. *sampling metódu* (vzorkovanie), teda *under sampling*, *over sampling*, ale aj naučenie modelu bez vzorkovania. Dôležité informácie o najlepšom modely tohto experimentu sa nachádzajú v tabuľke 5.

Tab. 5. *Metriky najlepšieho modelu v 1. experimente. Zdroj: Autori.*

Algoritmus	C4.5	
Vzorkovanie	žiadne	
Pomer	80/20	
Dôležité atribúty	Attr27, Attr41, Attr34	
Presnosť 0/1	97.14 % / 69.96 %	
Úspešnosť/chyba	96.35 % / 3.65 %	
	Nezbankrotuje	**Zbankrotuje**
Nezbankrotuje	8 187	241
Zbankrotuje	76	177

2. **Experiment** – výber atribútov pomocou metódy **PCA** a použitie algoritmov rozhodovacích stromov *C4.5*, *C5.0*, *random forest* a *CART*. Vytvorených bolo 52 modelov v rôznych pomeroch. Použitie vzorkovania ako v 1. experimente. Dôležité informácie o najlepšom modely tohto experimentu sa nachádzajú v tabuľke 6.

Tab. 6. *Metriky najlepšieho modelu v 2. experimente. Zdroj: Autori.*

Algoritmus	C5.0	
Vzorkovanie	žiadne	
Pomer	70/30	
Dôležité atribúty	Attr35, Attr56, Attr34	
Presnosť 0/1	95,49 % / 64,18 %	
Úspešnosť/chyba	95,33 % / 4,67 %	
	Nezbankrotuje	**Zbankrotuje**
Nezbankrotuje	12 370	589
zbankrotuje	24	43

3. **Experiment** – výber atribútov na základe **korelačných koeficientov** z korelačnej matice a použitie algoritmov rozhodovacích stromov *C4.5*, *C5.0*, *random forest* a *CART*. Vytvorených bolo 48 modelov v rôznych pomeroch. Použitie vzorkovania ako v 1. a 2. experimente. Dôležité informácie o najlepšom modely tohto experimentu sa nachádzajú v tabuľke 7.

Tab. 7. *Metriky najlepšieho modelu v 3. experimente. Zdroj: Autori.*

Algoritmus	C5.0	
Vzorkovanie	žiadne	
Pomer	90/10	
Dôležité atribúty	Attr27, Attr41, Attr34	
Presnosť 0/1	96,52 % / 95,24 %	
Úspešnosť/chyba	96,5 % / 3,5 %	
	Nezbankrotuje	**Zbankrotuje**
Nezbankrotuje	4 128	149
zbankrotuje	3	60

4. **Experiment** – použitie **všetkých atribútov** okrem atribútu Attr37 generovaných pomocou algoritmov rozhodovacích stromov *C4.5*, *C5.0*, *random forest* a *CART*. Vytvorených bolo 28 modelov v rôznych pomeroch. V tomto experimente sme neaplikovali vzorkovanie. Dôležité informácie o najlepšom modely tohto experimentu sa nachádzajú v tabuľke 8.

Tab. 8. *Metriky najlepšieho modelu v 4. experimente. Zdroj: Autori.*

Algoritmus	Random forest	
Vzorkovanie	žiadne	
Pomer	90/10	
Dôležité atribúty	Attr27, Attr34, Attr46	
Presnosť 0/1	96,83 % / 90,24 %	
Úspešnosť/chyba	96,71 % / 3,29 %	
	Nezbankrotuje	**Zbankrotuje**
Nezbankrotuje	4 123	135
zbankrotuje	8	74

3.5 Vyhodnotenie

V tomto kroku sme vyhodnocovali všetky vytvorené modely. Zo všetkých vytvorených modelov sme vybrali *40 najlepších*, ktoré boli natrénované pomerne správne a ich aplikovanie na testovacej množine prinieslo priaznivé výsledky.

Vyhodnotenie prinieslo tieto zaujímavé zistenia:

- Z hľadiska *presnosti pre triedu 0* (nezbankrotované podniky), mali všetky experimenty hodnoty vyššie ako **95,18 %.**
- *Najhoršie výsledky* mali modely, v ktorých boli chýbajúce hodnoty nahradené k-NN metódu s výnimkou algoritmu random fotest (ak bola použitá aj metóda vzorkovania).
- Z celkového hľadiska nemala metóda pod-vzorkovania žiadny vplyv na výsledné presnosti modelov.
- V prípade algoritmu *CART* bolo nutné pri každom modely použiť metódu vzorkovania, pretože v opačnom prípade sa model preučil.
- *Najnižšie AUC hodnoty* mali modely vygenerované algoritmom **C4.5** (ktoré mali súčasne aj najnižšiu presnosť) a algoritmom **CART** (pri ktorom dochádzalo k preučeniu).
- *Najvyššie AUC hodnoty* bola dosiahnutá len na modeloch vygenerovaných algoritmom **random forest.**
- *Najnižšiu chybu resp. najvyššiu úspešnosť* mal model vygenerovaný algoritmom **C5.0** na množine s NA hodnotami nahradenými priemerom atribútov.
- *Z hľadiska metód výberu atribútov* modely v 1., 2. a 3. experimente dosiahli väčšinou slabé presnosti. Pri metóde **LASSO** na množine s NA hodnotami nahradenými priemerom atribútov pre model **C4.5** bola presnosť veľmi dobrá v porovnaní s množinou, na ktorej boli NA hodnoty nahradené pomocou k-NN. Bolo to aj z dôvodu, že v oboch prípadoch bol počet atribútov rôzny. V modeli **C5.0,** v ktorom sme atribúty vybrali pomocou korelácií na množine s hodnotami priemeru, boli výsledky presnosti *výborné*.
- *Celkovo najlepšie výsledky* dosiahol **4. experiment** na rôznych modeloch, v ktorom sme použili všetky atribúty. Na základe toto sme teda zhodnotili, že veľmi dobré výsledky modelov sme dosiahli aj bez použitia metód na výber atribútov.
- Medzi najdôležitejšie atribúty v jednotlivých čiastkových výsledkoch modelov patril najmä:
    - **Attr24** (hrubý zisk (za 3 roky)/celkové aktíva),
    - **Attr27** (zisk z prevádzkových činností/finančné náklady),
    - **Attr34** (prevádzkové náklady/celkové pasíva),
    - **Attr41** (celkové pasíva/[(zisk z prevádzkovej činnosti + odpisy)*(12*365)]),

Za **najlepší model,** vzhľadom ku všetkým hodnotiacim metrikám, sme určili model rozhodovacieho stromu C5.0 vygenerovaný na množine, kde chýbajúce hodnoty boli nahradené priemerom daných atribútov a pomer trénovacej a testovacej množiny bol 90:10. Ako vstupné atribúty boli použité všetky dostupné, takže bez použitia metód výberu atribútov. Zistili sme, že z celkového počtu 4 340 podnikov bolo správne klasifikovaných 4213, čo znamená, že na základe tohto modelu by bola známy stav bankrotu v budúcnosti na 97,1 %. Zhrnutie dôležitých informácií tohto modelu sa nachádzajú v tabuľke 9.

Tab. 9. *Hodnoty metrík najúspešnejšieho modelu. Zdroj: Autori.*

Úspešnosť	97,07 %
Chyba modelu	2,93 %
Presnosť pre triedu 0	97,33 %
Presnosť pre triedu 1	87,27 %
AUC hodnota	0,8715

Senzitivita	0,9966
Špecifickosť	0,4593
Dôležitosť atribútov	**Attr27(100 %)**, Attr34(92.55 %), Attr41(91.78 %)

Z tohto modelu sme následne vygenerovali pravidlá, ako napríklad:

- **AK** (Attr27) *zisk z prevádzkových činností/finančné náklady*> 1 096.9 **A** (Attr34) *prevádzkové náklady/krátkodobé záväzky*> 0.58 **A** (Attr34) *prevádzkové náklady/krátkodobé záväzky* <= 0.85 **A** (Attr22) *zisk z prevádzkových činností/celkové aktíva* <= 0.117, **POTOM** 0 (nezbankrotuje). **INAK AK** (Attr22) *zisk z prevádzkových činností/celkové aktíva* > 0.117, **POTOM** podnik zbankrotuje.
- **AK** (Attr27) *zisk z prevádzkových činností/finančné náklady*> 1 096.9 **A** (Attr34) *prevádzkové náklady/krátkodobé záväzky* <= 0.581 **A** (Attr56) *(tržby – náklady na predané výrobky)/tržby* <= 0.2197 **A** (Attr9) *tržby/celkové aktíva* > 0.716 **A** (Attr9) *tržby/celkové aktíva* <= 1.117, **POTOM** podnik zbankrotuje.
- **AK** (Attr27) *zisk z prevádzkových činností/finančné náklady* <= 1 096.9 **A** (Attr41) *celkové pasíva/[(zisk z prevádzkovej činnosti + odpisy)*(12*365)]* <= -0.006 **A** (Attr58) *celkové náklady/celkové tržby* <= 0.975 **A** (Attr5) *[(hotovosť + krátkodobé CP + pohľadávky – krátkodobé záväzky)/(prevádzkové náklady - odpisy)]*365* > 204.93, **POTOM** podnik nezbankrotuje.
- **AK** (Attr27) *zisk z prevádzkových činností/finančné náklady* <= 1 096.9 **A** (Attr41) *celkové pasíva/[(zisk z prevádzkovej činnosti + odpisy)*(12*365)]* <= -0.006 **A** (Attr58) *celkové náklady/celkové tržby* > 0.975 **A** (Attr34) *prevádzkové náklady/krátkodobé záväzky* <= 0.011, **POTOM** podnik nezbankrotuje.

4 Diskusia a záver

Hlavným cieľom predkladaného príspevku bolo nájsť vhodný model, ktorý by bol schopný najlepšie a najpresnejšie predpovedať bankrot podniku a napomôcť tak investorom pri lepšom rozhodovaní. Využitím moderných techník a postupov objavovania znalostí sme použili dáta jednotlivých finančných ukazovateľov poľských spoločností pri tvorbe predikčných modelov s cieľom získať čo najlepšie výsledky. V rámci modelovania sme zaznamenali najlepšie výsledky prostredníctvom algoritmu C5.0 s presnosťou 97,33 % v triede 0 (nezbankrotuje) a 87,27 % presnosťou v triede 1 (zbankrotuje). O úspechu tohto modelu rozhodla aj najnižšia chyba klasifikácie spomedzi všetkých vytvorených experimentov s hodnotou 2,93 %.

Pri popisovaní prípadových štúdií sme v rámci analýzy súčasného stavu narazili na porovnateľnú štúdiu, v ktorej autori použili rovnaké dáta ako my v našej analýze. Z výsledkov našich experimentov a výsledkov štúdie bolo možné porovnať model random forest so všetkými použitými atribútmi a chýbajúcimi hodnotami nahradenými priemerom. V našom experimente sme dosiahli lepšiu presnosť o 14,5 %. V zmienenej štúdii bola presnosť v triede 1 - 75,76 %, náš výsledok dosiahol presnosť v triede 1 - 90,26 %. Presnosť v triede 0 z našich výsledkov predstavovala 96,73 %, čo bolo nižšie o 1,74 % ako v štúdii. Veľmi dobrým výsledkom v našom experimente predstavovala metrika senzitivity (TPR) s hodnotou 99,82 %, pričom v porovnávanej štúdii to bola nižšia hodnota - 99,45 %. Rovnako v tejto štúdií pomocou metódy forward selection identifikovali ako najdôležitejšie finančné pomery **Attr6** nerozdelený zisk/celkové aktíva, **Attr11** (hrubý zisk + mimoriadne položky + finančné náklady)/celkové aktíva, **Attr24** (hrubý zisk (za 3 roky)/celkové aktíva), **Attr27** (zisk z prevádzkových činností/finančné náklady) a **Attr60** (tržby/zásoby). V našej analýze medzi tieto ukazovatele

patrili už vyššie spomenuté, konkrétne **Attr24** (hrubý zisk (za 3 roky)/celkové aktíva), **Attr27** (zisk z prevádzkových činností/finančné náklady), **Attr34** (prevádzkové náklady/celkové pasíva) a **Attr41** (celkové pasíva/[(zisk z prevádzkovej činnosti + odpisy)*(12*365)]).

V budúcej práci sa chceme zaoberať použitiu zložitejších metód na tieto dáta, konkrétne metódam *extreme gradient boosting, neurónové siete, support vector machine* ale aj *k-najbližších susedov,* ako aj ich aplikácii na ďalšie dátové vzorky zaoberajúce sa finančnými ukazovateľmi.

Poďakovanie

Táto publikácia vznikla vďaka podpore projektu FEI-2018-52 financovaného Fakultou elektrotechniky a informatiky, Technickej univerzity v Košiciach, Agentúry na podporu výskumu a vývoja na základe Zmluvy č. APVV-16-0213 a Vedeckej grantovej agentúry MŠVVaŠ SR a SAV, projekt č. 1/0493/16.

Zoznam použitej literatúry

Altman, E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4), 589-609.

Atiya, A.F. (2001). Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. *IEEE Transactions on Neural Networks*, 12(4), 929–935. doi: 10.1109/72.935101

Beaver, W.H. (1966) Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, 4, 71–111.

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education*, 33, 1–42.

Dallas, G. (2013). Principal Component Analysis 4 Dummies: Eigenvectors, Eigenvalues and Dimension Reduction. Retrieved May 15, 2019, from https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/

Delina, R., & Packová, M. (2013). Validácia predikčných bankrotových modelov v podmienkach SR. *E+M Ekonomie a Management*, 16(3), 101–112.

Dwyer, G.P., & Tkac, P.A. (2009). The financial crisis of 2008 in fixed income markets. *Journal of International Money and Finance*, 28(8), 1293–1316. doi: 10.1016/j.jimonfin.2009.08.007

Fan, S., Liu, G., & Chen, Z. (2017). Anomaly detection methods for bankruptcy prediction. In *Proceedings of the 4th International Conference on Systems and Informatics* (pp.1456–1460). New York: IEEE. doi: 10.1109/ICSAI.2017.8248515

FitzPatrick, P. (1932). A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies. *The Certified Public Accountant*, in three issues: October, 598–605; November, 656–662; December, 727–731.

Fonti, V. (2017): Feature selection using LASSO. Retrieved May 15, 2019, from https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

Hardinata, L., Warsito, B., & Suparti, A. (2018). Bankruptcy prediction based on financial ratios using Jordan Recurrent Neural Networks: a case study in Polish companies. *Journal of Physics: Conference Series*. 1025 (1), 1–6. doi: 10.1088/1742-6596/1025/1/012098

Chen, M.Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, 62(12), 4514–4524. doi: 10.1016/j.camwa.2011.10.030

Ivashina, V., & Scharfstein, D. S. (2010). Bank Lending During the Financial Crisis of 2008. *Journal of Financial Economics*, 97(3), 319–338. doi: 10.1016/j.jfineco.2009.12.001

Jardin, P., & Séverin, E. (2011). Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of financial failure model. *Decision Support Systems*, 51(3), 701–711. doi: 10.1016/j.dss.2011.04.001

Merwin, C. L. (1942). *Financing small corporations in five manufacturing industries*. New York: National Bureau of Economic Research.

Moreira, J., Carvalho, A., & Horvath, T. (2018). *A general introduction to data analytics*. Hoboken: John Wiley.

Nagaraj, K., & Sridhar, A. (2015). A predictive system for detection of bankruptcy using machine learning techniques. Retrieved May 15, 2019, from https://arxiv.org/abs/1502.03601

Paralič, J. (2003). Objavovanie znalostí v databázach. Košice: Elfa.

Ringner, M. (2008). What is principal component analysis?. *Nature Biotechnology*, 26(3), 303–304.

Rybarova, D., Braunova, M., & Jantosova, L. (2016). Analysis of the Construction Industry in the Slovak Republic by Bankruptcy Model. *Procedia – Social and Behavioral Sciences*, 230, 298–306. doi: 10.1016/j.sbspro.2016.09.038

Smith, R., & Winakor, A. (1935). Changes in Financial Structure of UnsuccessfUl Industrial Corporations. In *Bureau of Business Research, Bulletin No. 51*. Urbana: University of Illinois Press.

Tomczak, S. (2016). Polish Companies Bankruptcy Data, Data Set. *UCI – Machine Learning Repository*. Retrieved May 15, 2019, from https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

Zieba, M., Tomczak, S., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 58, 93–101. doi: 10.1016/j.eswa.2016.04.001

Zhou, A., & Elhag, T.M.S. (2007). Apply Logit analysis in Bankruptcy Prediction. In *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization* (pp. 302–308). Stevens Point: WSEAS.

The article has been reviewed. | **Received:** 20 May 2019 | **Accepted:** 28 June 2019

Academic Editor: Stanislava Mildeova