VLIV METODY SMOTE NA PŘESNOST BANKROTNÍCH MODELŮ ZALOŽENÝCH NA KONVOLUČNÍCH NEURONOVÝCH SÍTÍCH

THE EFFECT OF THE SMOTE METHOD ON THE ACCURACY OF BANKRUPTCY MODELS BASED ON CONVOLUTIONAL NEURAL NETWORKS

Monika Šebestová, Petr Dostál

Abstrakt: *Tento článek zkoumá vliv metody SMOTE na přesnost predikce bankrotních modelů. Pro predikci bankrotu podniků v České republice byly použity konvoluční neuronové sítě založené na architektuře GoogLeNet. Vstupy do modelů jsou tvořeny finančními ukazateli podniků, jejichž hodnoty jsou převedeny na několik typů obrázků. Z provedeného výzkumu vyplynulo, že použití techniky SMOTE výrazně zvyšuje přesnost klasifikace aktivních a bankrotních podniků, a zároveň snižuje chybu II. druhu, která způsobuje nesprávnou klasifikaci bankrotního podniku za aktivní.*
Klíčová slova: predikce bankrotu, konvoluční neuronové sítě, finanční ukazatele, SMOTE, transfer learning, hluboké učení

Abstract: *This paper analyzes the effect of the SMOTE method on the prediction accuracy of bankruptcy models. Convolutional neural networks based on the GoogLeNet architecture are used for bankruptcy prediction of firms in the Czech Republic. The inputs to the models are composed of financial indicators of enterprises, whose values are converted into several types of images. The research conducted shows that the use of the SMOTE technique significantly improves the accuracy of classification of active and bankrupt enterprises, while reducing the type II error, which is the misclassification of a bankrupt enterprise as active.*

Keywords: *bankruptcy prediction, convolutional neural networks, financial indicators, SMOTE, transfer learning, deep learning*

JEL klasifikace: *C53, C45, G33*

1 ÚVOD

Problematika predikce finančního selhání a následného bankrotu firem je jednou z nejvíce diskutovaných oblastí, kterým je věnována pozornost na teoretické i praktické úrovni. V dnešní době, kdy hospodaření většiny firem v ČR i zahraničí významně ovlivnila pandemie COVID-19, se téma predikce bankrotu podniků stává opět aktuálním – společnosti se prodávají, transformují, dostávají do finanční tísně a krachují. Vzhledem k vyššímu riziku selhání malých a středních podniků, v porovnání s velkými podniky, je třeba vytvářet bankrotní modely určené přímo pro tuto skupinu podniků a zohlednit specifika typická pro tento typ společností (Kou et al., 2021). Do současné doby bylo publikováno mnoho studií, které se zabývají bankrotem podniků pomocí statistických metod, avšak menší pozornost je věnována přístupům založeným na hlubokém učení neuronových sítí. Při sestavování bankrotních modelů se lze často setkat s nedostupností dat. Nízký počet bankrotujících firem v souboru, ve srovnání s dostupností dat aktivních firem, vede k vysoce nevyváženému datovému souboru, což s sebou přináší obtíže při predikci bankrotních podniků a model tak může ignorovat zbankrotované firmy a klasifikovat je jako nekrachující (Garcia, 2022). Pro korekci tohoto nepoměru lze použít metody převzorkování, které změní rozložení tříd v trénovacích datech.

Cílem článku je posoudit vliv metody SMOTE (Synthetic Minority Oversampling Technique) na přesnost bankrotních modelů, vytvořených na základě konvolučních neuronových sítí. Vstupní proměnné modelů jsou tvořeny finančními a makroekonomickými ukazateli. Vzhledem k tomu, že konvoluční neuronové sítě pracují efektivně s analýzou obrazu, jsou vstupní numerická data ukazatelů převedena na obrázky v podobě grafů a zkoumáno, zda navýšení počtu vzorků u nevyvážených datových souborů ovlivní přesnost klasifikace podniků.

2 SOUČASNÝ STAV VĚDECKÉHO POZNÁNÍ

Mezi průkopníky statistických bankrotních modelů patřil Beaver (1966), který predikoval bankrot podniků pomocí jednorozměrné diskriminační analýzy. V následujících letech začaly vznikat bankrotní modely založené na vícerozměrné diskriminační analýze (Altman, 1968; Ohlson, 1980) a regresní analýze (Zmijewski, 1984). V současné době se preferovaným nástrojem pro

predikci bankrotu firem staly metody strojového učení a dostávají se tak do popředí jako metody vhodné pro řešení nelineárních problémů. Jednou z nejpopulárnějších metod jsou neuronové sítě, které patří mezi neparametrické modely a dle Vochozky (2020) jejich přesnost předpovědí v současné době vyniká nad lineárními regresními modely. Naidu a Govinda (2018) navrhli bankrotní predikční model s algoritmem učení založeném na neuronových sítích a náhodném lese (Random Forest). Analyzovali polské podniky za období let 2000-2012 a zjistili, že neuronové sítě s chybou 4,43 % dosahovaly lepších výsledků než náhodný les (chyba 5,19 %). Na základě stejných metod vytvořili predikční modely také Petropoulos et al. (2020). Došli k závěru, že neuronové sítě a náhodný les (Random Forest) dosahují téměř stejných výsledků a zároveň jsou schopny překonat přesnost ostatních metod jako např. Support Vector Machine (SVM), logistickou regresi nebo lineární diskriminační analýzu.

Relativně novou oblastí strojového učení je hluboké učení (Deep Learning). Počet aplikací těchto modelů do oblasti finanční analýzy je však velmi omezený. Výjimku tvoří několik publikací zabývajících se predikcí fluktuace cen akcií, kde jsou hluboké neuronové sítě používány pro analýzu časových řad (Balaji et al., 2018). Hluboké učení pro predikci likvidity na akciovém trhu ve Vietnamu použili ve svých publikacích Khang et al. (2021). Využití hlubokého učení pro predikci bankrotu firem uvedli ve své studii Yeh et al. (2015). Použili k tomu aplikaci hlubokých neuronových sítí typu Deep Belief Network (DBN) na volatilitu cen akcií. Model pro predikci finančního selhání australských podniků pomocí finančních dat vytvořili Elhoseny et al. (2022). K sestavení modelu použili metodu hlubokých neuronových sítí. Z výsledků vyplývá, že navržený model byl schopen predikovat finanční tíseň podniků s 95,8 % přesností.

Mezi významné modely hlubokého učení se řadí konvoluční neuronové sítě (CNN), jejichž představitelem je Szegedy et al. (2015). Publikací, které aplikují CNN sítě na finanční řízení podniků a predikci jejich bankrotů je však velmi málo. Důvodem může být orientace konvolučních neuronových sítí na obrázky, což s sebou přináší omezení ve zpracování numerických dat a finančních výkazů. Jako příklad jedné z mála aplikací lze uvézt publikaci Hosaka (2019), který aplikoval konvoluční neuronovou síť na predikci bankrotu firem pomocí pixelových obrázků ve stupních šedi. Použil k tomu účetní výkazy 102 bankrotních a 2062 aktivních firem za období 4 let. Hodnoty poměrových

finančních ukazatelů převedl na stupně šedi (hodnoty 0-255) a velikost datového souboru navýšil pomocí vážených průměrů, čímž vytvořil další syntetické vstupy. Celkem vytvořil 7520 obrázků, které použil k trénování konvoluční neuronové sítě typu GoogLeNet.

Predikční modely ke svému učení potřebují větší množství dat, které je v oblasti bankrotu firem někdy obtížnější získat. Často se lze setkat s nepoměrem mezi počtem aktivních a bankrotních podniků. Pro zvýšení počtu případů v nevyváženém datovém souboru lze použít techniky převzorkování, kterou je například syntetická technika minoritního převzorkování (SMOTE), představená autorem Chawla, et al. (2002). Metodu použil ve své publikaci Garcia (2022) a dokázal, že vyvážením datového souboru lze zvýšit přesnost klasifikace bankrotních podniků. Smiti a Soui (2020) použili metodu na úpravu polských bankrotních datových souborů. Na nevyváženém datovém souboru zjistili, že u několika metod strojového učení došlo k významnému zvýšení klasifikační výkonnosti. Mezi další autory, kteří metodu SMOTE použili pro zvýšení počtu vzorků při predikci bankrotu, lze zařadit Faris et al. (2020) a Shen et al. (2021). Další používanou metodou k úpravě nevyvážených datových souborů je podvzorkování (undersampling), při kterém dochází k vyvážení datových tříd odebráním vzorků patřících do většinové (negativní) třídy. Podle García et al. (2020) se jedná o méně vhodnou variantu než převzorkování, protože dochází k odstranění důležitých informací.

3 METODOLOGIE

3.1 Příprava dat

3.1.1 Odstranění odlehlých hodnot

Z datového souboru byly pomocí winsorizovaného průměru odstraněny odlehlé hodnoty ukazatelů, jejichž identifikace byla provedena na základě Grubbsova testu (Grubbs, 1969). Při **Grubbsově testu** byla data nejprve vzestupně seřazena způsobem $x_1 \leq x_2 \leq \ldots \leq x_n$ a následně určeno, zda hodnota x_1 příp. x_n představuje odlehlou hodnotu. Grubbsův test odlehlých hodnot je definován následujícím vztahem:

$$G = \frac{\max|x_i - \bar{x}|}{s}, \tag{1}$$

kde čitatel představuje maximální absolutní odchylku od průměru a *s* směrodatnou odchylku.

Samotné odstranění odlehlých hodnot bylo provedeno pomocí winsorizovaného průměru, pomocí kterého je možné nahradit určité procento extrémních hodnot na straně minima a maxima následující méně extrémní hodnotou. Winsorizovaný průměr je vyjádřen vztahem (Meloun a Militký, 2002):

$$\bar{x}_w(\vartheta) = \frac{1}{n}\left[(M+1)\left(x_{(M+1)} + x_{(n-M)}\right) + \sum_{i=M+2}^{n-M-1} x_{(i)}\right], \quad (2)$$

kde M = int (ϑ n/100),

ϑ je procento nahrazených pořádkových statistik,

$x_{(i)}$ jsou pořádkové statistiky (vzestupně setříděné prvky výběru),

n je počet prvků výběru.

3.1.2 Výběr vstupních proměnných

Pro výběr vhodných ukazatelů predikujících bankrot podniků byla použita logistická regrese, která je vyjádřena následujícím vztahem (Hendl, 2012):

$$\ln\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \sum_i \beta_i x_i, \quad (3)$$

kde $P(x)$ je odhad střední hodnoty pravděpodobnosti výskytu sled. jevu,

β_0 a β_i jsou regresní koeficienty,

x_i jsou hodnoty nezávisle proměnných.

Při hledání optimální podmnožiny prediktorů závisle proměnné z množiny potenciálních prediktorů byla použita kroková *dopředná* logistická regrese. Při dopředné krokové regresi dochází k výběru ukazatelů následovně: V prvním kroku je vybrán nejlepší prediktor a zařazen do tvořené množiny prediktorů. Ve druhém kroku je přidána proměnná, která nejvíce zlepšuje predikční schopnost proměnných, které jsou do predikce již zařazeny. V dalším kroku je odstraněna proměnná, jejíž příspěvek pro predikci klesl pod mez významnosti. Tento proces se opakuje, dokud přidáním dalšího prediktoru nedojde k významnému zlepšení predikce. Poté se proces přibírání prediktorů ukončí.

3.2 Syntetická technika minoritního převzorkování (SMOTE)

Nerovnováha mezi počtem vzorků aktivních (451) a bankrotních (69) podniků byla vyřešena pomocí aplikace algoritmu SMOTE, pomocí něhož došlo k převzorkování dat pro obnovení rovnováhy trénovací množiny. Jedná se o statistickou metodu, která vyváženým způsobem generuje nové instance z existujících dat a nemění tak počet případů většiny. Klíčovou myšlenkou metody SMOTE je zavedení syntetických případů namísto pouhé replikace dat z menšinové třídy, čímž dochází k poskytování nových informací pro modely strojového učení (Garcia, 2022).

Metoda SMOTE vytváří syntetická data pomocí algoritmu *k*-nejbližšího souseda. Postup metody je následující (Fernández et al., 2018):

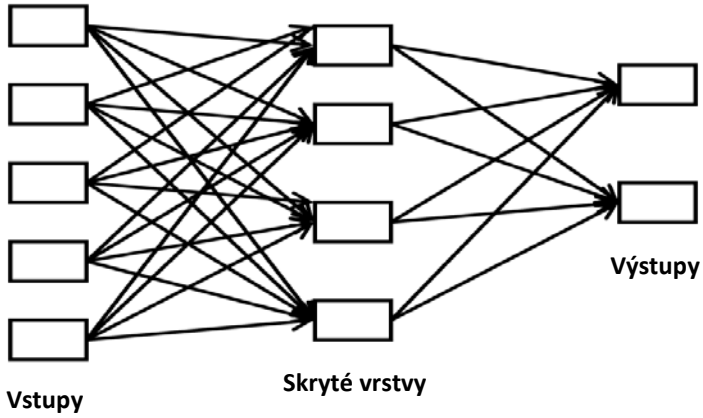
- Stanovení velikosti převzorkování *N* (počet nových případů, nejčastěji uváděno v procentech z počtu dat v menšinové třídě).
- Následně je proveden iterační proces, který se skládá z několika kroků:
  - Z pozitivní (menšinové) množiny je náhodně vybrána instance třídy.
  - Pro zvolenou instanci jsou stanoveny k-nejbližší sousedé *K* (obvykle k = 5).
  - Nakonec je vybráno *N* z těchto *K* instancí pro interpolaci nových syntetických instancí. Pomocí metriky vzdálenosti je vypočítán rozdíl vzdáleností mezi vektorem příznaků (vybranou instancí třídy z bodu 1) a jeho sousedy. Získaný rozdíl je vynásoben libovolnou náhodnou hodnotou z intervalu [0,1] a přičten k předchozímu vektoru příznaků.

3.3 Konvoluční neuronové sítě

Neuronové sítě jsou považovány za nedokonalý model myšlení lidského mozku. Jsou založeny na modelu, který se skládá ze vzájemně propojených skupin umělých neuronů. Trénování sítě spočívá ve vhodném nastavení vah mezi vstupy, jejichž hodnoty jsou na začátku trénování nastaveny náhodně (Dostál, 2012). V praxi se nejčastěji používají vícevrstvé neuronové sítě. Jsou složeny ze vstupní a výstupní vrstvy, mezi nimiž se nacházejí vrstvy skryté. Za nejznámější a nejpoužívanější algoritmus učení neuronových sítí je považována metoda **backpropagation**, která je založena na minimalizaci rozdílu mezi

výstupní a vstupní hodnotou (Munakata, 2008). Struktura vícevrstvé neuronové sítě je uvedena na následujícím obrázku:

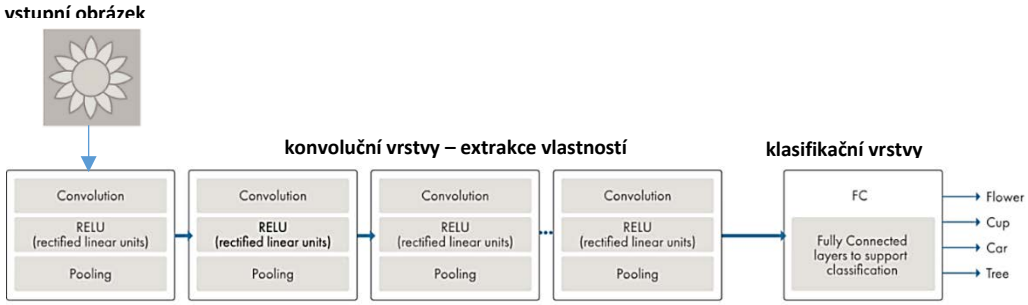
Obrázek 1: Struktura vícevrstvé neuronové sítě



Zdroj: upraveno dle Kabari a Nwachukwu (2012)

Jedním z nástrojů neuronových sítí, který je nejvhodnější použít pro práci s obrazem, jsou konvoluční neuronové sítě (CNN). Jedná se o architektury pro hluboké učení (Deep Learning) a používají se zejména pro rozpoznávání objektů na snímcích (Gao a Lim, 2019). Jádrem CNN sítí jsou dvourozměrné konvoluční vrstvy, díky nimž jsou tyto sítě vhodné pro zpracování souborů 2D dat, jako např. obrázků (Jirkovský, 2018a). Na Obrázku 2 je znázorněna obecná skladba CNN sítě určená ke klasifikaci objektů v obrázcích. Konvoluční vrstvy bývají často doplněny vrstvami ReLU (Rectified Linear Units) a Pooling, jejichž úkolem je úprava výstupů (např. odstranění záporných hodnot) a převzorkování na menší rozměr. Za sadou konvolučních vrstev se nacházejí klasifikační vrstvy, které vyhodnocují rysy extrahované ze vstupních obrázků (Jirkovský, 2018a).

Obrázek 2: Obecná skladba konvoluční neuronové sítě



Zdroj: Jirkovský (2018a)

Stejně jako jiné neuronové sítě se CNN síť skládá ze vstupní vrstvy, výstupní vrstvy a mnoha skrytých vrstev mezi nimi. Vstupní vrstva vkládá do sítě obrázky, které následně normalizuje. Definuje velikost vstupních obrázků a v případě potřeby tuto velikost mění. Dvourozměrná konvoluční vrstva se skládá z neuronů, které se připojují k dílčím oblastem vstupních obrázků a aplikuje na ně klouzavé konvoluční filtry, které jsou schopny se učit. Tyto filtry se aktivují tehdy, jakmile uvidí určitý typ vizuálního prvku, např. hranu nebo skvrnu barvy. Aktivační vrstva ReLU zajišťuje rychlejší a efektivnější trénink tím, že mapuje záporné hodnoty na nulu a kladné hodnoty zachovává. Pooling vrstva zmenšuje velikost obrazu se zachováním obsažených informací. Sama o sobě neprovádí žádné učení, pouze snižuje počet parametrů, které se mají učit v následujících vrstvách. Po konvolučních a pooling vrstvách následuje plně propojená vrstva, která kombinuje všechny naučené informace z předchozí vrstvy. Jedná se o předposlední vrstvu, jejímž výstupem je vektor o počtu tříd, které bude síť schopna predikovat. Tento vektor obsahuje také pravděpodobnost predikce pro každý klasifikovaný obrázek. Poslední vrstva architektury CNN sítí je klasifikační vrstva, která poskytuje výstup klasifikace (Specify Layers of Convolutional Neural Network, © 1994-2022).

3.3.1 Transfer Learning

Proces učení konvoluční neuronové sítě „od nuly“ vyžaduje velké množství trénovacích dat a nastavení milionů váhových koeficientů. Alternativou k tomuto typu učení je využití „předučené“ sítě, která je schopna ze vstupních údajů sama extrahovat charakteristické rysy. Tento postup se nazývá transfer learning a umožňuje doladit parametry předem vytvořených sítí, čímž se trénink stává rychlejší a jednodušší.

Princip fungování metody transfer learning vychází z předpokladu, že prvky klasifikace objektů v obrazových datech jsou v prvních vrstvách sítě relativně shodné. K přizpůsobení sítě pro konkrétní problémy tedy stačí doučit pouze několik posledních vrstev, které rozlišují konkrétní objekty na obrázku (Jirkovský, 2018b). K učení CNN sítí lze použít velké množství předtrénovaných sítí. Mezi nejznámější architektury přenosového učení patří např. GoogLeNet a AlexNet, jejichž struktura umožňuje doladit výkon sítě na vlastních datech. Tyto sítě byly natrénovány na více než milionu obrázků a dokáží klasifikovat objekty do tisíce kategorií. V předloženém článku je použita konvoluční neuronová síť GoogLeNet. Má 22 vrstev a byla vyvinuta výzkumníky ze společnosti Google. Vstupní vrstva architektury přijímá obrázky o rozměrech 224 x 224 pixelů.

4 DATA

Do výzkumu bylo zahrnuto 520 stavebních podniků (451 aktivních a 69 bankrotních podniků). Vybrané podniky musely splňovat následující požadavky: Analyzované podniky sídlí na území České republiky. Podnik je zařazen do kategorie malých a středních podniků (MSP). Jedná se o společnosti s ručením omezeným nebo akciové společnosti. Hlavní podnikatelská činnost společností musí podle CZ-NACE spadat do kategorie F (Stavebnictví). Podnik zařazený do výzkumu musí být aktivní, neaktivní (po úpadku) nebo aktivní v insolvenčním řízení. Informace o jednotlivých účetních závěrkách byly získány z databází AMADEUS a ORBIS. Data aktivních i bankrotních podniků pokrývají období let 2011-2018, přičemž u bankrotních podniků se jedná o účetní výkazy 1-3 roky před bankrotem (k bankrotní události došlo v letech 2012-2019).

S využitím softwaru Statistica byly ze základního datového souboru pomocí winsorizovaného průměru odstraněny odlehlé hodnoty, k jejichž identifikaci byl použit Grubbsův test. Nahrazeno bylo 2,5 % extrémních hodnot na straně minima a maxima (celkem tedy 5 % hodnot) následující méně extrémní hodnotou. Následně bylo z celkového počtu 45 finančních a makroekonomických ukazatelů pomocí logistické regrese vybráno 8 ukazatelů, jejichž p-hodnota byla menší než 0,05. Tyto ukazatele následně tvořily vstupní proměnné v navržených modelech. Seznam finálních proměnných je uveden v Tabulce 1.

Tabulka 1: Redukovaný počet ukazatelů na základě logistické regrese

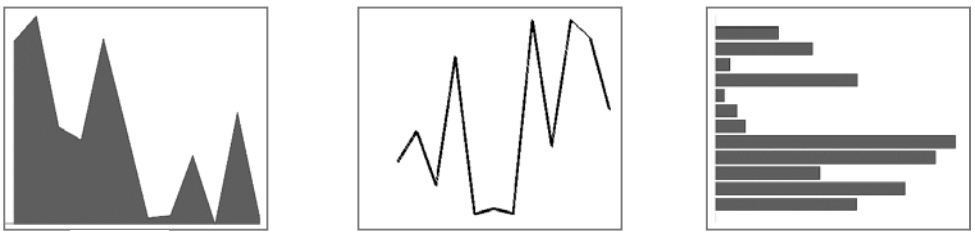
Ukazatel	p-value
Logaritmus aktiv	0,0000
Krátkodobé závazky/celková aktiva	0,0000
EAT/tržby	0,0016
Zásoby/průměrné tržby	0,0002
Celkové závazky/celková aktiva	0,0000
Meziroční růst aktiv	0,0021
CZ<>CA*	0,0000
HDP CELKEM (mil. Kč)	0,0128

* hodnota 1, pokud celkové závazky > celková aktiva;
hodnota 0, pokud celkové závazky < celková aktiva

Zdroj: vlastní zpracování

Vzhledem k tomu, že CNN sítě dosahují nejlepších výsledků ve zpracování obrazu, bylo potřeba hodnoty proměnných každého podniku graficky zpracovat. Cílem bylo predikčnímu modelu předložit obrázek, který by odrážel finanční situaci podniku v čase, přičemž každý z vybraných ukazatelů byl zaznamenán v 3leté časové ose. V softwaru MATLAB od společnosti MathWorks byly z hodnot ukazatelů jednotlivých firem vygenerovány obrázky, které sloužily jako vstupní proměnné do modelů. Pro predikci bankrotu byly použity tři různé typy grafického zpracování vstupních proměnných, jejichž podoba je uvedena na následujícím Obrázku 3.

Obrázek 3: Typy grafického zpracování vstupních dat



AREA PLOT BARH

Zdroj: vlastní zpracování

5 VÝSLEDKY

Predikční modely byly sestaveny ve dvou variantách. V první variantě byl použit originální počet bankrotních podniků (69 vzorků), ve druhé variantě byla

použita metoda SMOTE pro vytvoření syntetických proměnných a navýšení počtu případů v menšinové třídě bankrotních podniků na 173 vzorků. Uvedené modely byly sestaveny na základě stejných parametrů a testováno, která varianta je pro predikci bankrotu firem nejpřesnější.

K modelování predikce bankrotu firem byla použita předtrénovaná konvoluční neuronová síť založená na architektuře GoogLeNet. Vstupní data byla rozdělena na trénovací a validační množinu. Trénink sítě probíhal na 70 % obrázků a validace modelu byla ověřena na zbývajících 30 % obrázků. Trénink neuronové sítě probíhal ve 40 epochách. Po uplynutí maximálního počtu epoch bylo trénování sítě zastaveno a zaznamenána jeho přesnost.

Na základě výše uvedených ukazatelů (Tabulka 1) byly sestaveny modely. Trénovací a validační proces každého modelu byl spuštěn třikrát a do výsledků zaznamenána průměrná hodnota jejich výstupů. Uvedené přesnosti modelů jsou v následujících tabulkách posuzovány pomocí celkové přesnosti a metriky F-measures (vhodnější pro nevyvážené datové soubory):

$$\text{Celková přesnost} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{4}$$

$$\text{F-measures} = \frac{2*Precission*Sensitivity}{Precission+Sensitivity}, \tag{5}$$

kde *TP* je počet správně klasifikovaných bankrotních podniků, *TN* je počet správně klasifikovaných aktivních podniků, *FP* je počet aktivních podniků, které model označil za bankrotní a FN je počet bankrotních podniků, které byly modelem označeny za aktivní, *Precission = TP/*(*TP+FP*), *Sensitivity = TP/*(*TP+FN*). Některé modely mohou mít vyšší charakteristiku Precission, jiné naopak Sensitivity. Charakteristika F-measures z těchto hodnot vytváří harmonický průměr (převrácenou hodnotu aritmetického průměru převrácených hodnot) a umožňuje tak zjistit přesnost modelu s nevyváženou datovou sadou pomocí jednoho čísla (Chen, 2011). U obou ukazatelů přesnosti lze dostáhnout výsledku na intervalu [0,1], kdy 1 značí nejvyšší přesnost.

5.1 Modelování

V této části jsou uvedeny výsledky modelování bankrotních modelů, které byly vytvořeny s využitím softwaru MATLAB verze R2022b.

5.1.1 Modely s původním počtem dat

V následující části jsou uvedeny trénovací a validační přesnosti modelů, jejichž vstupní proměnné pochází z původního počtu vzorků (451 aktivních a 69 bankrotních firem). Celková přesnost klasifikace při učení sítě byla nejvyšší u dvou vytvořených modelů, a to u grafického zpracování typu AREA a PLOT (99,73 %). Nejvyšší přesnosti validace bylo dosaženo u modelu s použitými obrázky typu PLOT (90,38 %). K posouzení kvality predikce bankrotních podniků, byla použita metrika F-measures. Stejně jako u celkové přesnosti dosáhly nejlepších výsledků při tréninku sítě modely s grafickými vstupy typu AREA a PLOT. U validačních dat dokázal nejlépe predikovat bankrotní podniky model s grafickým zpracováním typu PLOT. Výše uvedené výsledky jsou zobrazeny v Tabulce 2.

Tabulka 2: Celková přesnost a F-measures modelů s originálními daty

Typ obrázku	Celková přesnost		F-measures	
	Trénink	Validace	Trénink	Validace
AREA	0,9973	0,8654	0,9895	0,4324
PLOT	0,9973	0,9038	0,9895	0,5714
BARH	0,9780	0,8846	0,9167	0,5500

Zdroj: vlastní zpracování

5.1.2 Modely se syntetickými daty

Učící proces CNN sítí dosahuje lepších výsledků při větším množství trénovacích dat. V analyzovaném vzorku je počet položek pro trénování sítě malý (především u bankrotních podniků). V následující části jsou proto uvedeny trénovací a validační přesnosti modelů, u nichž došlo k navýšení vzorku bankrotních podniků o syntetická data. Nové případy byly vytvořeny pomocí metody SMOTE. Počet aktivních podniků zůstal stejný, zatímco u bankrotních podniků došlo k navýšení vzorku o 1,5násobek (ze 69 na 173 vzorků) a syntetické proměnné byly vypočítány z 5 nejbližších sousedních hodnot. Po aplikaci metody SMOTE dosáhl nejlepších výsledků ve všech sledovaných kategoriích přesností model s grafickým zpracováním vstupů typu PLOT. Výsledné přesnosti modelů po použití metody SMOTE jsou uvedeny v Tabulce 3.

Tabulka 3: Celková přesnost a F-measures modelů s použitím SMOTE

Typ obrázku	Celková přesnost		F-measures	
	Trénink	Validace	Trénink	Validace
AREA	0,9963	0,8984	0,9959	0,8119
PLOT	0,9977	0,9144	0,9959	0,8462
BARH	0,9908	0,8984	0,9835	0,8224

Zdroj: vlastní zpracování

5.2 Porovnání modelů

Z předešlé analýzy přesností modelů vyplývá, že lepších výsledků je dosahováno u modelů s navýšeným počtem vzorků bankrotních podniků pomocí metody SMOTE. Pro potvrzení, zda je tento typ modelů vhodný také pro predikci bankrotu, byla na validačních datech provedena analýza chyby II. druhu a zjišťováno, zda modely nemají ve velké míře tendenci označovat bankrotní podniky za aktivní (Tabulka 4). Tím by společnosti nebyly včas upozorněny na blížící se finanční problémy a tato situace by mohla vézt až k jejich bankrotu. Chyba II. druhu tedy měří podíl případů, ve kterém jsou bankrotující podniky nesprávně identifikovány jako aktivní (falešně negativní) a její výpočet je dán vztahem:

$$\text{Chyba II. druhu} = \frac{FN}{TP+FN}, \tag{6}$$

kde *FN* je počet bankrotních podniků, které byly modelem označeny za aktivní a *TP* je počet správně klasifikovaných bankrotních podniků.

Tabulka 4: Chyba II. druhu u vytvořených modelů

Typ obrázku	Originální data	Rozšířená data (SMOTE)
AREA	0,619	0,211
BARH	0,476	0,154
PLOT	0,524	0,151
Průměr	0,540	0,172

Zdroj: vlastní zpracování

Z výsledků uvedených v Tabulce 4 je vidět patrný rozdíl mezi modely s původním počtem vzorků a modely s vytvořenými syntetickými daty pomocí metody SMOTE. Modely s původním počtem dat dosahují mnohem vyšší chyby

II. druhu než modely rozšířené o syntetická data a ve větší míře označují bankrotní podniky za aktivní, což při důvěřování tomuto typu modelu může firmám přinášet vysoké náklady a ztráty. Jako vhodný typ modelů pro predikci bankrotu firem jsou tedy považovány modely s daty upravenými pomocí metody SMOTE, které jsou schopny lépe odhalit bankrotující podnik a včas tak firmy informovat o možném selhání podnikání.

6 ZÁVĚR

Tento článek se zabýval problematikou predikce bankrotu firem s použitím konvolučních neuronových sítí. Bylo zkoumáno, zda použití syntetické techniky minoritního vzorkování (SMOTE) pro vytvoření syntetických dat má vliv na přesnost predikce. Z hodnot vybraných finančních a makroekonomických ukazatelů byly vygenerovány 3 typy obrázků, které sloužily jako vstupy do modelů. K tvorbě modelů byla použita metoda transfer learning s předučenou sítí GoogLeNet. Na základě zjištěných výsledků lze konstatovat, že s použitím metody SMOTE se zvyšuje přesnost správné klasifikace podniků na aktivní a bankrotní. Současně se výrazně snižuje chyba II. druhu, ve které jsou bankrotní podniky nesprávně označeny za aktivní. Tato chybná predikce může mít pro firmy díky absenci včasného varování negativní důsledky. Nejlepších predikčních výsledků dosáhl model s typem obrázku PLOT, a to na základě trénovacích i validačních dat, včetně nejlepších výsledků v oblasti chyby II. druhu.

Pro další výzkum by bylo vhodné zanalyzovat i jiné předučené architektury CNN sítí a zjistit, zda některá z nich pracuje s obrázky finančních ukazatelů a následnou predikcí bankrotu efektivněji.

POUŽITÉ ZDROJE

[1] ALTMAN, Edward I., 1968. *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*. The Journal of Finance. 23(4), 589-609. ISSN 00221082.

[2] BALAJI, A., D.S. HARISH RAM a Binoy B. NAIR, 2018. *Applicability of Deep Learning Models for Stock Price Forecasting An Empirical Study on BANKEX Data*. Procedia Computer Science. 143, 947-953. ISSN 18770509.

[3] BEAVER, Wh., 1966. *Financial Ratios as Predictors of Failure*. Journal of Accounting Research. (4), 71-111.

[4] CHAWLA, N. V., K. W. BOWYER, L. O. HALL a W. P. KEGELMEYER, 2002. *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research. 16, 321-357. ISSN 1076-9757.

[5] CHEN, Mu-Yen, 2011. *Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches*. Computers. 62(12), 4514-4524. ISSN 08981221.

[6] DOSTÁL, Petr, 2012. *Pokročilé metody rozhodování v podnikatelství a veřejné správě*. Brno: Akademické nakladatelství CERM. ISBN 978-80-7204-798-7.

[7] ELHOSENY, Mohamed, Noura METAWA, Gabor SZTANO a Ibrahim M. EL-HASNONY, 2022. *Deep Learning-Based Model for Financial Distress Prediction*. Annals of Operations Research. Springer. ISSN 0254-5330.

[8] FARIS, Hossam, Ruba ABUKHURMA, Waref ALMANASEER, Mohammed SAADEH, Antonio M. MORA, Pedro A. CASTILLO a Ibrahim ALJARAH, 2020. *Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market*. Progress in Artificial Intelligence. 9(1), 31-53. ISSN 2192-6352.

[9] FERNÁNDEZ, Alberto, Salvador GARCÍA, Mikel GALAR, Ronaldo C. PRATI, Bartosz KRAWCZYK a Francisco HERRERA, 2018. *Learning from Imbalanced Data Sets* [online]. 1. Switzerland: Springer. ISBN 978-3-319-98074-4. Dostupné také z: https://link.springer.com/book/10.1007/978-3-319-98074-4

[10] GAO, Qishuo a Samsung LIM, 2019. *Classification of hyperspectral images with convolutional neural networks and probabilistic relaxation*. Computer Vision and Image Understanding. 188. ISSN 10773142.

[11] GARCIA, John, 2022. *Bankruptcy prediction using synthetic sampling*. Machine Learning with Applications. 9. ISSN 26668270.

[12] GARCÍA, V., J.S. SÁNCHEZ, A.I. MARQUÉS, R. FLORENCIA a G. RIVERA, 2020. *Understanding the apparent superiority of over-sampling through*

an analysis of local information for class-imbalanced data. Expert Systems with Applications. 158. ISSN 09574174.

[13] GRUBBS, Frank E., 1969. *Procedures for Detecting Outlying Observations in Samples*. Technometrics. 11(1), 1-21. ISSN 0040-1706.

[14] HENDL, Jan, 2012. *Přehled statistických metod: analýza a metaanalýza dat*. 4., rozš. vyd. Praha: Portál. ISBN 978-80-262-0200-4.

[15] HOSAKA, Tadaaki, 2019. *Bankruptcy prediction using imaged financial ratios and convolutional neural networks. Expert Systems with Applications*. 117, 287-299. ISSN 09574174.

[16] JIRKOVSKÝ, Jaroslav, 2018a. *Deep learning vs. signály a časové řady*. Automa [online]. 2018(11), 24-26 [cit. 2021-05-27]. ISSN 1210-9592. Dostupné z: https://automa.cz/Aton/FileRepository/pdf_articles/11834.pdf

[17] JIRKOVSKÝ, Jaroslav, 2018b. *Deep Learning a prostředí MATLAB*. Humusoft [online]. 31. 8.2 018 [cit. 2022-08-14]. Dostupné z: https://www.humusoft.cz/blog/20180831-deep-learning/

[18] KABARI, L. G. a E. O. NWACHUKWU, 2012. *Neural Networks and Decision Trees For Eye Diseases Diagnosis.* Advances in Expert Systems [online]. InTech, 2012-12-05 [cit. 2017-01-05]. ISBN 978-953-51-0888-7. Dostupné z: doi:10.5772/51380

[19] KHANG, Pham Quoc, Klaudia KACZMARCZYK, Piotr TUTAK, Paweł GOLEC, Katarzyna KUZIAK, Radosław DEPCZYŃSKI, Marcin HERNES a Artur ROT, 2021. *Machine learning for liquidity prediction on Vietnamese stock market*. Procedia Computer Science. 192, 3590-3597. ISSN 18770509.

[20] KOU, Gang, Yong XU, Yi PENG, Feng SHEN, Yang CHEN, Kun CHANG a Shaomin KOU, 2021. *Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection*. Decision Support Systems. 140. ISSN 01679236.

[21] MELOUN, Milan a Jiří MILITKÝ, 2002. *Kompendium statistického zpracování dat: metody a řešené úlohy včetně CD*. Praha: Academia. ISBN 80-200-1008-4.

[22] MUNAKATA, Toshinori., 2008. *Fundamentals of the new artificial intelligence: neural, evolutionary, fuzzy and more*. 2nd ed. London: Springer, 255 s. ISBN 978-184-6288-395.

[23] NAIDU, Pranav a Kharisma GOVINDA, 2018. *Bankruptcy prediction using neural networks*. In: Second International Conference on Inventive Systems and Control. s. 248-251. ISBN 978-1-5386-0806-7.

[24] OHLSON, James A., 1980. *Financial Ratios and the Probabilistic Prediction of Bankruptcy*. Journal of Accounting Research. 18(1), 109-131. ISSN 00218456.

[25] PETROPOULOS, Anastasios, Vasilis SIAKOULIS, Evangelos STAVROULAKIS a Nikolaos E. VLACHOGIANNAKIS, 2020. *Predicting bank insolvencies using machine learning techniques*. International Journal of Forecasting. 36(3), 1092-1113. ISSN 01692070.

[26] SHEN, Feng, Xingchao ZHAO, Gang KOU a Fawaz E. ALSAADI, 2021. *A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique*. Applied Soft Computing. 98. ISSN 15684946.

[27] SMITI, Salima a Makram SOUI, 2020. *Bankruptcy Prediction Using Deep Learning Approach Based on Borderline SMOTE*. Information Systems Frontiers. 22(5), 1067-1083. ISSN 1387-3326.

[28] *Specify Layers of Convolutional Neural Network*, © 1994-2022. MathWorks [online]. [cit. 2022-08-14]. Dostupné z: https://www.mathworks.com/help/deeplearning/ug/layers-of-a-convolutional-neural-network.html

[29] SZEGEDY, Christian, WEI LIU, YANGQING JIA, et al., 2015. *Going deeper with convolutions*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015, 1-9. ISBN 978-1-4673-6964-0.

[30] VOCHOZKA, Marek, 2020. *Metody komplexního hodnocení podniku*. 2. aktualizované vydání. Praha: Grada Publishing. Finance (Grada). ISBN 978-80-271-1701-7.

[31] YEH, Shu-Hao, Chuan-Ju WANG a Ming-Feng TSAI, 2015. *Deep belief networks for predicting corporate defaults.* 2015 24th Wireless and

Optical Communication Conference (WOCC). IEEE, 2015, 159-163. ISBN 978-1-4799-8868-6.

[32] ZMIJEWSKI, Mark E., 1984. *Methodological Issues Related to the Estimation of Financial Distress Prediction Models*. Journal of Accounting Research. 22, 59-82. ISSN 00218456.

AUTOŘI

Ing. Monika Šebestová, Ústav informatiky, Fakulta podnikatelská, VUT v Brně, Kolejní 2906/4, 612 00 Brno, e-mail: Monika.Sebestova@vut.cz.

prof. Ing. Petr Dostál, CSc., Ústav informatiky, Fakulta podnikatelská, VUT v Brně, Kolejní 2906/4, 612 00 Brno, e-mail: dostalp@vut.cz.

AUTHORS

Ing. Monika Šebestová, Department of Informatics, Faculty of Business and Management, Brno University of Technology, Kolejní 2906/4, 612 00 Brno, Czech Republic, e-mail: Monika.Sebestova@vut.cz.

prof. Ing. Petr Dostál, CSc., Department of Informatics, Faculty of Business and Management, Brno University of Technology, Kolejní 2906/4, 612 00 Brno, Czech Republic, e-mail: dostalp@vut.cz.