doc. Ing. Klára Antlová, Ph.D.
katedra informatiky
Ekonomická fakulta
Technická univerzita v Liberci

Oponentský posudek disertační práce:

Modelování skórovacích karet pro rozhodování v nebankovních finančních institucích

Autor: Ing. Monika Papoušková
Školitel: doc. Ing. Petr hájek, Ph.D.

Univerzita Pardubice,
Fakulta ekonomicko – správní
Ústav systémového inženýrství a informatiky

Obor programu: Aplikovaná informatika

Rozsah disertační práce: 109 stran textu + přílohy

Cíle práce, aktuálnost tématu a jeho významu pro obor

Cílem předkládané disertační práce je návrh modelu skórovacích karet, které umožní ohodnotit a predikovat chování dlužníka pomocí heterogenních kombinací metod strojového učení v rámci spotřebitelských úvěrů. Téma disertační práce je velmi aktuální a výsledky práce budou přínosem pro obor Aplikovaná informatika.

Postup řešení, použité metody, splnění cílů

Splnění cílů disertační práce je založeno na návrhu modelu predikce pravděpodobnosti defaultu a případné očekávání ztráty úvěru. Pro modelování byly použity individuální klasifikátory jak homogenní, tak i heterogenní kombinace modelů. Použité metody pro modelování pravděpodobnosti neplacení závazku jsou tradiční statistické metody, rozhodovací stromy a rozhodovací lesy a neuronové sítě. Z testovaných klasifikačních metod byla zvolena jako nejvhodnější Metoda Stacking s meta klasifikátorem náhodný les, a proto byla použita ve dvoufázovém modelu očekávané ztráty (tabulka 10, str. 78). Navržený model byl následně konfrontován se současnými modely skórovacích karet a s modelem používaným v konkrétní nebankovní instituci (při využití dat z případové studie) a dosáhl lepších výsledků (tabulka 12, str.83). Lze konstatovat, že vytčený cíl disertační práce byl splněn.

Stanovisko k výsledkům disertační práce a původnímu konkrétnímu přínosu

Přínos práce je založen na návrhu heterogenní kombinace modelů pravděpodobnosti selhání. Zvolená kombinace klasifikátorů přináší vyšší přesnost a tím i nižší náklady spojené s chybnou klasifikací v porovnání se současným stavem řešení. Modelování úvěrového rizika

pomocí kombinace heterogenních modelů expozice při defaultu je unikátní. Navržený model umožňuje nejen predikci pravděpodobnosti defaultu, ale i ekonomický dopad defaultu úvěru.

Vyjádření k formální úpravě, jazykové úrovni, systematičnosti, přehlednosti, habilitační práce

Práce je rozdělena do osmi kapitol, kde první kapitola uvádí vysvětlením úvěrového a kreditního skórováním. Současně lze i tuto kapitolu chápat i jako velmi podrobnou literární rešerši. Následující druhá kapitola uvádí hlavní a dílčí cíle práce. Třetí kapitola je věnována metodologii vytváření skórovacích karet. Čtvrtá kapitola popisuje metody odhadu pravděpodobnosti selhání. Pátá kapitola obsahuje popis ukazatelů výkonnosti vhodných pro hodnocení v oblasti řízení úvěrového rizika. Šestá kapitola je zaměřená na modelování skórovacích karet spotřebitelských úvěrů. Sedmá kapitola přináší diskusi a osmá shrnuje dosažené výsledky a přínosy práce. Z formálního hlediska je práce velmi pečlivě zpracovaná, práce je srozumitelná, bez překlepů a je přehledně logicky uspořádána, oceňuji rovněž poměrně obsáhlou literární rešerši.

Závěr

Disertační práce se zabývá problematikou, která je aktuální. V práci je nutné ocenit velmi dobrou úroveň znalostí autorky, jejichž uplatnění díky systematickému vědeckému zkoumání odborné literatury a vhodného metodického postupu, přispívá k řešení dané problematiky. **Doporučuji tuto disertační práci k obhajobě** a v případě úspěšné obhajoby navrhuji udělit akademický titul „philosophiae doctor“ (Ph.D.) Ing. Monice Papouškové.

Otázky k diskusi:
Má Vámi navržený model nějaké nevýhody? Bude model využíván v praxi?

V Liberci dne 22.5.2019

doc. Ing. Klára Antlová, Ph.D.

Oponentní posudek disertační práce

Název práce: Modelování skórovacích karet pro rozhodování v nebankovních finančních institucích

Autorka: Ing. Monika Papoušková
Školitel: doc. Ing. Petr Hájek, Ph.D., Fakulta ekonomicko-správní, Univerzita Pardubice
Oponent: doc. RNDr. Pavel Pražák, Ph.D., Fakulta informatiky a managementu, Univerzita Hradec Králové

Téma, aktuálnost tématu, struktura a obsah práce

Práce se věnuje tématu modelování úvěrového rizika především v nebankovních institucích. Vzhledem k současné zadluženosti a problémům s vypořádáním úvěrových produktů nezanedbatelné části obyvatel České republiky lze téma řešené v práci považovat za užitečné, potřebné a aktuální. Práce poměrně komplexně popisuje problematiku úvěrového skórování se zaměřením na modelování pravděpodobnosti defaultu a expozice při defaultu a snaží se integrovat metody strojového učení, přístupy a charakteristiky do jednotného metodologického postupu.

Předložená disertační práce včetně příloh má 123 stran. Obsahuje 8 číslovaných kapitol, 2 nečíslované kapitoly nazvané Úvod a Závěr, dále 3 seznamy, které souhrnně uvádějí přehled obrázků, tabulek a zkratek. V závěru práce je uveden seznam autorčiných publikací a důležitá příloha, týkající se popisu dat, které byly v práci použity.

První kapitola s názvem Úvěrové skórování se věnuje problematice automatického hodnocení klientů žádajících úvěrový produkt. Popisuje různá hlediska, přístupy a metody, které jsou k této činnosti využívány. Zaměřuje se také na popis tvorby různých skórovacích karet a na s. 22 zavádí vztah (1) pro očekávanou ztrátu, který je později v šesté kapitole použit při tvorbě vlastního modelu. Tato kapitola uvádí také rešerši literatury zabývající se modelováním pravděpodobnosti defaultu, především pak té literatury, která se zabývá kombinací možných metod, viz Tabulka 2. Další část rešerše se týká literatury zabývající modelováním ztráty při defaultu a expozice při defaultu, viz přehled v Tabulce 3. V závěru kapitoly se uvádí, kde je možné hledat informace o potenciálních klientech úvěrových produktů.

V druhé kapitole s názvem Cíle disertační práce je uveden jednak hlavní cíl práce a dále 7 dílčích cílů, viz popis níže v tomto posudku.

Kapitola třetí nese název Metodologie vytváření skórovacích karet. Text se zaměřuje na možné použití dataminingové metodologie CRISP-DM v problematice tvorby skórovacích karet.

Čtvrtá kapitola s názvem Modelování skórovacích karet rozvíjí tu část metodologie CRISP-DM, která se týká problematiky modelování. Autorka zde uvádí poměrně rozsáhlý přehled možných metod modelování. Text se pak omezuje na relativně stručný popis tradičních statistických metod, rozhodovacích stromů a lesů, algoritmů podpůrných vektorových strojů, vícevrstvé neuronové sítě typu perceptron. Dále se zde věnuje algoritmům možných kombinací jednotlivých modelů (Bagging, Boosting, Decorate, RandomSubSpace, Rotační les, Náhodný les). Závěr kapitoly popisuje heterogenní kombinaci modelů, zvláště algoritmu Stacking.

V páté kapitole s názvem Hodnocení modelů se autorka věnuje popisu kritérií hodnocení modelů, zvláště jde o Giniho koefient, ROC křivku, Kolmogorov-Smirnovovu statistiku, koeficient přesnosti klasifikace a také koeficienty MAE, RMSE a koeficient determinace R^2 pro hodnocení regresních modelů.

Šestou kapitolu s názvem Modelování skórovacích karet spotřebitelských úvěrů lze považovat za klíčovou část předložené disertační práce. Jedná se o popis vývoje (návrhu a tvorby) a testování vlastního modelu autorky, včetně porovnání s jinými modely a s jistým současným modelem anonymní nebankovní společnosti.

Sedmá kapitola se věnuje diskusi a osmá kapitola shrnuje přínosy disertační práce.

Cíl práce, výzkumné otázky, soulad s oborem studia

Podle s. 34 je cílem práce *„navrhnout model skórovacích karet, který umožní modelování pravděpodobnosti defaultu a expozice při defaultu pomocí heterogenních kombinací metod strojového učení.“* K dosažení uvedeného hlavního cíle bylo stanoveno 7 dílčích cílů a výzkumných problémů: 1) Návrh modelu pravděpodobnosti defaultu využívající podvzorkování majoritní třídy nedefaultních úvěrů a heterogenní kombinaci klasifikátorů. 2) Návrh metriky pro hodnocení výsledku klasifikace modelu pravděpodobnosti defaultu. 3) Návrh modelu expozice při defaultu využívající heterogenní kombinaci regresorů. 4) Návrh dvoufázového modelu úvěrového rizika spotřebitelských úvěrů skládající se z modelu pravděpodobnosti defaultu a modelu expozice při defaultu. 5) Verifikace navrženého modelu na reálných datech české nebankovní finanční instituce. 6) Provést srovnávací analýzu výsledků navrženého modelu se současnými modely, resp. srovnat jednofázové a dvoufázové modely predikce úvěrového rizika. 7) Provést srovnávací analýzu výsledků navrženého modelu se současným modelem vybrané nebankovní instituce. Podrobněji jsou tyto dílčí cíle uvedeny na s. 34 a 35.

Dílčí kroky jsou vhodně voleny a vedou k dosažení cíle. Práce svojí tématikou patří do oboru Aplikovaná informatika doktorského studijního programu Aplikovaná informatika.

Použité vědecké metody v disertační práci

Práce vychází především z rešerše a analýzy velmi rozsáhlých literárních zdrojů o předmětné problematice. Vzhledem k rozsáhlosti možných návrhů modelů skórovacích karet lze orientaci v literatuře a provedenou analýzu považovat za náročný úkol. Velká pozornost je věnována metodologii CRISP-DM s hlavním důrazem na fázi modelování. Nicméně pro úspěšnou tvorbu modelu je pozornost také věnována získání, porozumění a přípravě relevantních dat, které poskytla anonymní nebankovní instituce. Důležitou metodou je také návrh a provedení hodnocení modelů, kde se autorka omezuje na klasické a osvědčené kvantifikátory. Z textu práce je také patrné, že autorka provedla řadu simulací, které jí umožnily provést srovnání různých modelů a nastavení modelů. Potenciál, který mají metody a jejich skupiny pro řešení problematiky, je v práci poměrně podrobně rozebrán. Metody je možno v souhrnu považovat za adekvátní a správné. Ačkoli se v jisté části jedná o metody klasické, práce obsahuje i méně obvyklé metody heterogenní kombinace dílčích modelů. Literární zdroje, které autorka při zpracování práce použila, jsou uvedeny a v textu citovány.

Splnění cílů práce

Problematika řešená v práci je díky struktuře a charakteru dat poměrně náročná, existují s ní různé zkušenosti a je navíc dynamická. Experimenty, které lze s modely a jejich parametry provést nemohou být ze své podstaty úplné a ukončené. Popsaný metodický rámec lze však dobře použít pro další zlepšení skórovacích karet úvěrů, kterými se práce zabývá. Na základě textu šesté kapitoly lze konstatovat, že stanovený cíl i formulované dílčí cíle práce byly splněny. K hlavnímu cíli práce a podobně ke všem dílčím krokům vedoucím k jeho dosažení přispěla autorka svým vlastním příspěvkem. Postupy a metody jsou aplikovány správně.

Přesnost práce, formální stránka práce

Celková struktura práce je čitelná a dobře navržena. V práci se tak lze snadno orientovat. Jazyk textu je srozumitelný a neobsahuje mnoho překlepů. Hloubka zpracování obsahu jednotlivých kapitol je však různá. Zatímco první kapitola, která se věnuje rešerši a současnému stavu poznání, je podrobně a díky tabulkám zdařile a přehledně zpracována, čtvrtá kapitola, která se věnuje modelování, a pátá kapitola, která se věnuje hodnocení modelů, by mohly být zpracovány do větší hloubky s případným konkrétním zaměřením na modelování skórovacích karet.

První kapitola pracuje s některými pojmy, které nejsou definovány nebo jsou definovány až později v textu. Příkladem nedostatečně vysvětleného pojmu na začátku textu je pojem expozice při defaultu (EAD), který je použit např. na s. 14, 15 a 22, ale je částečně vysvětlen na s. 27 a podrobně až na s. 75 v kapitole šesté. Z tohoto popisu je zřejmé, že se jedná o poměr z intervalu (0, 1), případně o poměr vyjádřený v %, takže je třeba se zamyslet nad konkrétním vyjádřením EAD na posledním řádku Tabulky 7 na s. 73, který daný charakter nemá. Podobně ztráta při defaultu (LGD) je vysvětlena až na s. 70 ve vztahu (12). Bylo by užitečné, aby byly v úvodní kapitole uvedeny všechny potřebné pojmy a případně také odkazy na další kapitoly, kde jsou popsány metody, o kterých se v první kapitole píše, ale které zde není možné uvést, např. většina pojmů z Tabulky 2 na s. 25 je uvedena ve čtvrté kapitole, což při četbě první kapitoly není zřejmé. V šesté kapitole, která se věnuje konkrétní aplikaci a konkrétním datům, lze očekávat, že se již nové pojmy víceméně nebudou vyskytovat. Bylo by také vhodné důsledně uvádět definiční obory vstupních veličin a obory hodnot výstupních veličin, např. ve vztahu (1) s. 22 se lze pouze dohadovat, jaká je hodnota případné očekávané ztráty (EL). V tomto vztahu je uveden znak $\times$ a naopak ve vztazích (11), (13) a (14) je uveden znak *, žádný z těchto znaků se však v české literatuře pro vyjádření součinu reálných čísel nepoužívá. V Tabulce 1 na s. 20 je uveden znak <=, ten je možné použít při psaní kódu, v textu by zřejmě měl být uveden dobře definovaný symbol uspořádání $\leq$. Domnívám se, že v těchto tradičních zápisech není třeba být inovativní, ale spíše naopak jednotně dodržovat zavedené značení.

Je pochopitelné, že není možné detailně vysvětlit všechny modely uvedené ve čtvrté kapitole. Přesto by bylo vhodné, aby byl popis přesnější a lépe odpovídal hloubce poznání autorky. Bylo by užitečné, pokud by zde byl komentář, jak lze metodu použít pro řešený problém modelování skórovacích karet a provázat tuto kapitolu s odstavcem 3.3, který popisuje některé vstupní proměnné. Např. popis Bayesovské sítě, uvedený na s. 48, je v práci založen především na Bayesově vzorci (5). Zde by bylo užitečné vysvětlit, co představují uzly tam zmíněného grafu, co představují hrany grafu a jak se případně odvozují pravděpodobnosti jevu, který reprezentuje koncový uzel. Ve vztahu k řešenému problému by pak bylo vhodné uvést, co je možným vstupem dané sítě. Podobný komentář je možné uvést i k jiným uvedeným modelům.

Návrhy a doporučení

- V odstavci 5.4 uvést, které regresní modely budou na základě MAE, RMSE nebo R^2 upřednostněny.
- Řadu nalezených výsledků bylo možné získat použitím software Weka 3.8, to je stručně uvedeno na s. 68. Bylo by vhodné popsat, proč bylo použito toto prostředí a uvést tento software také v nadpisu Tabulky 6 na s. 68. V této souvislosti by bylo možné doporučit, aby místo strohého konstatování „*Zdroj: vlastní*“ např. pod Tabulkou 9 na s. 77, Tabulkou 10 na s. 78 atp. uvést např. „*Zdroj: vlastní uspořádání a výpočet v software Weka 3.8*.“
- V Tabulce 7 na s. 73 uvést odkaz na přílohu na s. 110, která podrobněji uvedenou tabulku doplňuje a vysvětluje některé hodnoty z této tabulky.

Výsledky práce, poznatky a přínosy práce

Práce přináší výsledky v oblasti teoretické, metodologické i praktické. Za hlavní teoreticko-metodologický výsledek a inovaci lze považovat konceptuální návrh dvoufázového modelu rizika poskytnutí úvěrových produktů (odstavec 6.1). V práci se dále podařilo vypracovat návrh heterogenní kombinace modelů pravděpodobnosti defaultu, navrhnout heterogenní kombinace modelů expozice při defaultu, porovnat různé přístupy k vybalancování datového souboru (odstavec 6.3 a Tabulka 9), navrhnout metriku klasifikace defaultu (odstavec 6.1 a vztah (11), dále odstavec 6.3 a vztah (14)). Za hlavní praktický výsledek lze považovat ověření navržených modelů na reálných datech české nebankovní finanční instituce (odstavec 6.3 a Tabulka 10) a provedení srovnávací analýzy výsledků navrženého modelu se současnými používanými modely (odstavec 6.3 a Tabulka 12, odstavec 6.4), včetně kalibrace skórovací karty (odstavec 6.5).

Možné otázky a náměty do diskuse při obhajobě

Práce ukázala, že problematika modelování skórovacích karet je rozsáhlá, náročná, obtížně modelovatelná a potřebná. V důsledku toho musí být některé závěry poněkud obecné. Těchto skutečností si je autorka dobře vědoma.

- Nastavení metod použitých při modelování není jednoznačné -v práci je uvedeno např. v Tabulce 6 na s. 68. Jak se k danému nastavení došlo? Byl proveden výpočet i pro jiná nastavení a s jakým výsledkem?
- Při hodnocení kvality a specifikace regresních modelů se používají také informační kritéria (např. Akaikeho (AIC) nebo Schwarz-Bayesovo (BIC)). Bylo by možné tato kritéria použít i pro hodnocení regresních modelů uvedených v práci a doplnit tak kritéria z odstavce 5.4, nebo tuto činnost automaticky provádí použitý software?
- Bylo by možné a rozumné nějak (i uměle) snížit počet vstupních atributů v Tabulce 7 na s. 73? Zkoumala se multikolinearita kvantitativních proměnných vstupních dat? Měl tento jev nějaký vliv na uvažované regresní modely?
- V algoritmu C4.5 na s. 49 je v bodě 1) zmíněno kritérium ukončení, v bodě 3) je pak uveden informační zisk. Lze na základě použité literatury kritérium ukončení tohoto algoritmu konkretizovat a dovysvětlit pojem informační zisk?

Závěr

Práce splňuje nároky kladené na disertační práce. Doporučuji, aby Ing. Monice Papouškové byl po úspěšné obhajobě udělen titul Ph.D.

Hradec Králové 18. 5. 2019

Pavel Pražák

Univerzita Konštantína Filozofa v Nitre
Fakulta prírodných vied
Katedra informatiky

POSUDOK ZÁVEREČNEJ PRÁCE

Téma: MODELOVÁNÍ SKÓROVACÍCH KARET PRO ROZHODOVÁNÍ V NEBANKOVNÍCH FINANČNÍCH INSTITUCÍCH
Typ záverečnej práce: Doktorandská záverečná práca
Študijný program: Aplikovaná informatika
Študijný obor: Aplikovaná informatika
Autor: Ing. Monika Papoušková
Oponent: prof. RNDr. Michal Munk, PhD.

Práca sa zaoberá úlohou klasifikácie v doméne financií, kde autorka na riešenie tohto data mining-ového (DM) problému používa pestrú paletu metód a prístupov k modelovaniu pravdepodobnosti default-u a expozície pri default-e. Téma práce je aktuálna, práve podcenenie úverového rizika viedlo k ostatnej finančnej kríze.

Práca je prehľadná, veľmi dobre čitateľná, pomerne náročná problematika je zrozumiteľne popísaná. Kladne hodnotím, že práca má aplikačnú doménu, výsledky je možné použiť na modelovanie úverového rizika a pravdepodobne výsledky práce budú použité pre konkrétnu nebankovú finančnú spoločnosť. Oceňujem dotiahnutie teoretickej práce až do aplikačných výstupov.

V analýze súčasného stavu pozitívne hodnotím prehľadne zosumarizované štúdie z hľadiska modelovania pravdepodobností default-u, straty a expozície pri default-e.

V kapitole 2 je podrobne rozpísaný cieľ práce ako aj úlohy (čiastkové ciele) potrebné na dosiahnutie stanoveného cieľa. Úlohy vyplývajúce z cieľa práce sú podrobne rozpísané a riešené v ďalšej časti práce. Oceňujem presnú a jasnú formuláciu ako aj reálnosť jednotlivých úloh (čiastkových cieľov), ktorých riešenie viedlo k naplneniu stanoveného cieľa práce.

V kapitole 3 je podrobne popísaný proces modelovania skórovacích kariet. Na riadenie procesu autorka použila metodiku CRISP-DM, ktorá na jednom mieste sprostredkováva postup od porozumenia problému, získania dát, cez predspracovanie dát, modelovanie, až po vyhodnotenie a využitie výsledkov. Oceňujem túto časť práce, ktorá výrazne prispieva

k sprehľadneniu celého procesu vytvárania skorovacích kariet. Vo fáze porozumenia dátam mi chýba jasné určeniu typu DM problému/úlohy získavania znalostí (KD) a tým pádom aj zúžený návrh metód na riešenie daného problému. Naopak pozitívne hodnotím spracovanie fázy predspracovania dát, ktorá predstavuje časovo najnáročnejšiu fázu celého procesu. V tomto prípade závislá/vysvetľovaná premenná je dichotomická (default/ne-default).
Je to bežná prax aj v prípade bankových inštitúcií alebo v tomto prípade sa rozlišuje viacero úrovní zlyhania (default)?

Kapitoly 4, 5, 6 predstavujú nosnú časť práce, kde sú popísané metódy a prístupy (ensembling) k modelovaniu, techniky hodnotenia modelov a koncepčný rámec pre modelovanie úverového rizika. Spracovanie homogénnych/heterogénnych kombinácií modelov vhodne dopĺňajú pseudokódy jednotlivých techník. Rovnako oceňujem rozsah z hľadiska kombinácií použitých metód, techník a prístupov k modelovaniu dát. Napriek rozsahu, práca nestráca na prehľadnosti. Pozitívne hodnotím vyhodnotenie a porovnanie výsledkov získaných rôznymi prístupmi k riešeniu úlohy klasifikácie v predmetnej oblasti ako aj porovnanie navrhovaného modelu so súčasným modelom nebankovej finančnej inštitúcii. Aj keď závery z porovnávania sú jasné, samotné porovnávanie na prvý pohľad až také zrejmé nie je.
Prečo ste testovali rozdiely medzi dvoma vzorkami, keď sa ponúka testovanie rozdielov medzi viacerými vzorkami?
Identifikovali ste homogénne skupiny ($p > 0{,}05$) alebo štatisticky významné rozdiely ($p \leq 0{,}05$)?

Výsledky práce boli publikované v časopisoch a zborníkoch z konferencií zameraných práve na predmetnú problematiku. Najvýznamnejšie výsledky práce boli publikované v *Springer* sérii *Smart Innovation, Systems and Technologies* a v impaktovanom časopise *Decision Support Systems* (Q1, CCC) vydavateľstva *Elsevier*. Práve v spomínanom časopiseckom výstupe sú publikované výsledky popísané v kapitole 6, ktoré môžeme označiť, z hľadiska prínosu práce, za kľúčové. Myslím, že by bolo hodnotné publikovať aj samotné dáta, samozrejme za predpokladu, že budú vhodne predspracované/anonymizované, ako aj prístupy k modelovaniu dát, ktoré sú hodnotné a prínosné aj v iných doménach, kde sa rieši úloha klasifikácie.
Nezvažovali ste popri research article v Elsevier publikovať aj data article (Data in Brief/Mendeley Data) a method article (MethodsX)?

Celkovo prácu hodnotím ako veľmi dobrú, práca pôsobí transparente z hľadiska predspracovania, modelovania ako aj vyhodnotenia a porovnania výsledkov. O čom svedčia aj publikačné výstupy autorky.

Práca spĺňa požiadavky kladené na tento typ záverečnej práce a prácu odporúčam v predloženej podobe obhajovať a po jej úspešnej obhajobe navrhujem, aby Ing. Monike Papouškovej bol udelený akademický titul Philosophiae Doctor (PhD.) v študijnom programe Aplikovaná informatika, oboru Aplikovaná informatika.

V Nitre, 20.05. 2019

prof. RNDr. Michal Munk, PhD.