UNIVERZITA PARDUBICE
FAKULTA EKONOMICKO-SPRÁVNÍ
ÚSTAV SYSTÉMOVÉHO INŽENÝRSTVÍ A INFORMATIKY

MODELOVÁNÍ SKÓROVACÍCH KARET PRO ROZHODOVÁNÍ V NEBANKOVNÍCH FINANČNÍCH INSTITUCÍCH

DISERTAČNÍ PRÁCE

Autor: Ing. Monika Papoušková

Školitel: doc. Ing. Petr Hájek, Ph.D.

Pardubice 2019

UNIVERSITY OF PARDUBICE
FACULTY OF ECONOMICS AND ADMINISTRATION
INSTITUTE OF SYSTEM ENGINEERING AND INFORMATICS

MODELLING CREDIT SCORECARDS FOR DECISION-MAKING IN NON-BANK FINANCIAL INSTITUTIONS

DISSERTATION THESIS

Author: Ing. Monika Papoušková

Supervisor: doc. Ing. Petr Hájek, Ph.D.

Pardubice 2019

Abstrakt

V předložené disertační práci jsou uvedeny možnosti modelování skórovacích karet. Nejprve je představen problém a základní pojmy z oblasti modelování skórovacích karet. Dále je zde uveden přehled současného stavu řešení v této oblasti. Další část práce se zabývá metodami strojového učení, které jsou vhodné pro predikci pravděpodobnosti defaultu a expozice při defaultu. Těmito metodami jsou jak individuální metody, tak kombinace modelů. Kombinace modelů jsou založeny na kombinaci stejných individuálních metod (kombinace homogenních modelů) nebo různých individuálních metod (kombinace heterogenních modelů). Jako individuální metody jsou v této práci použity čtyři skupiny metod strojového učení, a to tradiční statistické metody, rozhodovací stromy, podpůrné vektorové stroje a neuronové sítě. Tyto metody jsou pak použity jako základní algoritmy v homogenních a heterogenních kombinacích modelů. Pomocí těchto metod je realizováno modelování skórovací karty. Je navržen model skórovacích karet, který umožňuje modelování pravděpodobnosti defaultu a expozice při defaultu pomocí heterogenních kombinací metod strojového učení. Výsledky jednotlivých metod jsou mezi sebou porovnány. Efektivnost navrženého výsledného modelu je prezentována jak porovnáním se současnými přístupy k modelování skórovacích karet, tak se současným modelem zvolené nebankovní finanční instituce.

Klíčová slova

Úvěrové riziko, úvěrové skórování, kombinace modelů, očekávaná ztráta, expozice při defaultu, pravděpodobnost defaultu

Abstract

The possibilities of score cards modelling is presented in this thesis. At first, the problem of fundamental terms of the score cards modelling is presented. Afterwards, the overview of the state-of-the-art in this field is introduced. The next part deals with the machine learning methods that are suitable for the prediction of the probability of default and the exposure at default. These methods are both individual methods and ensemble learning. Ensemble learning is based on a combination of the same individual methods (homogeneous ensemble) or different individual methods (heterogeneous ensemble). Four groups of the machine learning methods are used in this thesis. These are traditional statistical methods, decision trees, support vector machines and neural networks. These methods are used as the basic algorithms in homogeneous and heterogenous ensembles learning. The score cards modelling is realised using these methods. There is the score card model proposed that allows the modelling of the probability of default and the exposure at default using heterogeneous ensemble methods of machine learning. The results of individual methods are compared with each other. The effectiveness of the proposed model is presented in comparison with both the state-of-the-art in the score card modelling and the current model of the selected non-bank financial institution.

Keywords

Credit risk, credit scoring, ensemble learning, expected loss, exposure at default, probability of default

Prohlašuji:

Tuto práci jsem vypracovala samostatně. Veškeré literární prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu použité literatury.

Byla jsem seznámena s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 9/2012, bude práce zveřejněna v Univerzitní knihovně a prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 20. 3. 2019 Ing. Monika Papoušková

Poděkování

Ráda bych poděkovala panu doc. Ing. Petru Hájkovi, Ph.D. za odborné vedení práce, rady a připomínky při vypracování disertační práce. Dále chci poděkovat své rodině za podporu po celou dobu studia.

OBSAH

Seznam obrázků

Seznam tabulek

Seznam zkratek

Acc	Přesnost
AMT	Alternující modelovací strom
AUC	Plocha pod ROC křivkou
avg	průměr
BRKI	Bankovní registr klientských informací
CBS	Credit Bureau Score
CDT	Credal rozhodovací strom
CRISP-DM	CRoss-Industry Standard Proces for Data Mining
CRÚ	Centrální registr úvěrů
EAD	Expozice při defaultu
ForestPA	Rozhodovací strom s penalizací proměnných
FN	Chybně negativní případy
FNR	Podíl chybně negativních případů
FP	Chybně pozitivní případy
FPR	Podíl chybně pozitivních případů
KS	Kolmogorov-Smirnov
LGD	Ztráta při defaultu
LogR	Logistická regrese
LR	Lineární regrese
LTD	Podíl vyplacená částka / volné zdroje
maj	majority, většina
MAE	Střední absolutní chyba
MC	Náklady chybné klasifikace
NN	Neuronová síť
NRKI	Nebankovní registr klientských informací
PD	Pravděpodobnost defaultu
ROC	Receiver Operating Characteristic
RF	Náhodný les
RMSE	Odmocnina ze střední kvadratické chyby
RotForest	Rotační les
SVM	Podpůrné vektorové stroje
SVR	Podpůrná vektorová regrese

TN	Správně negativní případy
TNR	Podíl správně negativních případů
TP	Správně pozitivní případy
TPR	Podíl správně pozitivních případů

Úvod

Zadlužování je fenoménem dnešní společnosti. V dnešní době je dluh vnímán jako prostředek, který slouží k rychlejšímu dosažení cílů. Ve snaze zvýšit si životní úroveň se mnoho lidí přiklání k financování svých potřeb úvěrem. Existuje velké množství úvěrových produktů a stejně tak existuje mnoho finančních institucí a zprostředkovatelů, kteří úvěr poskytují. Lidé s nedostatečnou bonitou jsou v bankovních institucích odmítáni a přicházejí k nebankovním finančním institucím, které své úvěrové produkty poskytují s větší ochotou. Vyšší rizikovost těchto žadatelů však znamená vyšší cenu úvěru. Žadatelé, kteří přicházejí žádat o úvěr do nebankovních finančních institucí, mají mnohdy nízkou finanční gramotnost, neodhadnou své finanční možnosti, nevěnují pozornost smluvním podmínkám a často jim nerozumí, proto je potřebné, aby se věřitelé zabývali tzv. zodpovědným úvěrováním, které lze definovat jako zodpovědnost věřitele půjčovat takovou výši prostředků dlužníkovi, která nebude znamenat jeho nadměrnou zadluženost. Úvěrové produkty by měly odpovídat potřebám klientů a odrážet jejich schopnost splácet své závazky, čehož může být dosaženo v případě, že všichni věřitelé i zprostředkovatelé jednají čestně, férově a profesionálně jak před uzavřením úvěrové smlouvy, během jejího trvání, tak i po jejím ukončení. Pro věřitele to znamená, že by měli správně hodnotit bonitu klienta a správně vyhodnotit vhodnost produktu.

Hodnocení bonity žadatele je klíčové. Podhodnocení bonity žadatele znamená zvýšení pravděpodobnosti nesplacení úvěru (tzv. default). Některé finanční instituce ovšem neprovádí řádné hodnocení bonity z důvodu urychlení procesu poskytnutí úvěru se snahou získat co nejvíce klientů. Podobnou motivaci mají i někteří zaměstnanci těchto finančních institucí, kteří jsou odměňování za počet a objem uzavřených úvěrů, ale už ne za budoucí kvalitu těchto žadatelů, budoucích klientů. Úvěrové (kreditní) riziko, tedy riziko, které vyplývá z neschopnosti nebo neochoty dostát svým závazkům, by však mělo být dominantním zájmem věřitelů, neboť díky jeho aktivnímu řízení chrání sebe i žadatele o úvěr před budoucími problémy.

Vzhledem ke zvyšujícím se požadavkům na tvorbu nástrojů, které budou schopné automaticky rozhodnout, zda bude žadatel, tj. budoucí klient svůj úvěr splácet nebo ne, dochází k rozvoji metod na bázi strojového učení. Ty se ve srovnání s expertním hodnocením úvěrového rizika (tzv. úvěrové skórování) ukazují jako efektivnější z důvodu vyšší přesnosti

a nižších nákladů na hodnocení. V oblasti úvěrového skórování jsou těmito metodami například neuronové sítě, podpůrné vektorové stroje, náhodné stromy či lesy nebo kombinace modelů. Kombinace modelů strojového učení, ať už homogenní či heterogenní, spojují dílčí modely v jeden komplexní model. Tyto kombinace modelů profitují z diverzity dílčích modelů a snižují tak riziko přeučení. V případě modelování úvěrového rizika se ukázaly jako velmi efektivní také z toho důvodu, že umožňují modelovat různé rizikové profily klientů.

Úvěrové skórování je obvykle založeno na několika desítkách vstupních proměnných, které tvoří tzv. skórovací kartu. Současné skórovací karty jsou založeny na odhadu pravděpodobnosti defaultu klienta. Ta však tvoří pouze jednu složku úvěrového rizika. Dalšími jsou velikost expozice a ztráta při defaultu. Pro finanční instituce jsou tyto parametry úvěrového rizika důležité z toho důvodu, že i částečně nesplacené úvěry mohou být v konečném důsledku ziskové. Pravděpodobnost defaultu proto neinformuje o celkové ztrátě způsobené defaultem klienta.

Cílem této disertační práce je proto navrhnout takový model skórovacích karet, který umožní modelování pravděpodobnosti defaultu a expozice při defaultu pomocí heterogenních kombinací metod strojového učení.

V první kapitole jsou definovány základní pojmy z oblasti úvěrového skórování a zhodnocen současný stav řešení v této oblasti. Druhá kapitola stanovuje cíle doktorské disertační práce. Další kapitola popisuje metodologii vytváření skórovacích karet. Následující kapitola se zabývá metodami, které jsou v současné době používané při modelování úvěrového rizika. Pátá kapitola uvádí metriky, kterými lze porovnávat jednotlivé modely mezi sebou. Další kapitola si klade za cíl modelovat jak pravděpodobnost defaultu, tak očekávanou ztrátu úvěru. K modelování pravděpodobnosti defaultu, resp. očekávané ztráty jsou použity jak individuální klasifikátory, resp. regresory, tak homogenní a heterogenní kombinace modelů. Efektivnost navrženého modelu je porovnána jednak se současnými modely skórovacích karet a dále se současným modelem skórovací karty používaným v nebankovní finanční instituci, jejíž reálná data jsou použita pro verifikaci navrženého modelu. K porovnání jsou použity jak tradiční metriky přesnosti predikce, tak ekonomické ukazatele nákladů a ziskovosti. Sedmá kapitola diskutuje dosažené výsledky vzhledem k současnému stavu poznání. Osmá kapitola pak prezentuje původní vědecké přínosy a aplikační přínosy disertační práce.

1 Úvěrové skórování

Tato kapitola se zabývá úvěrovým neboli kreditním skórováním. Literatury na toto téma existuje mnoho, neboť úvěrovému skórování a jeho modelování byla v posledních letech věnována stále větší pozornost. Kromě nových a přesnějších metod modelování se objevují nové problémy, jako je modelování dalších parametrů úvěrového rizika nebo kalibrace skórovacích karet. Tato kapitola se soustředí jednak na literaturu, která položila základní otázky k problematice automatizovaného ohodnocení (skórování) klientů z hlediska jejich bonity, resp. pravděpodobnosti defaultu (probability of default), kdy na počátku byly využívány tradiční statistické metody jako například diskriminační analýza či logistická regrese. Dále se tato část práce zaměřuje na současné přístupy k modelování úvěrového skórování, které se snaží co nejvíce zvýšit přesnost predikce pravděpodobnosti defaultu pomocí kombinace modelů (ensemble learning). Dále je zde uvedena základní literatura, která se zabývá modelováním nejen pravděpodobnosti defaultu, která je pro úvěrové společnosti klíčová, ale také modelováním ztráty při defaultu (loss given default) a expozice při defaultu (exposure at default), neboť tyto proměnné jsou významnou složkou současných modelů, které se využívají při řízení úvěrového rizika.

Další část kapitoly se zabývá základní terminologií, která souvisí s úvěrovým skórováním, např. s definicí dobrého a špatného úvěru či s konkrétními typy skórovacích karet, které lze v praxi využít. Jsou zde uvedeny problémy, které jsou spojeny s výstavbou skórovací karty. Dále jsou zde zmíněny základní informace o úvěrových registrech, které jsou pro úvěrové společnosti cenným zdrojem dat.

1.1 Historie

Úvěr je definován jako množství peněz vypůjčených jednotlivcem či podnikatelským subjektem od věřitele, které mají být splaceny v plné výši i s úroky (Siddiqi 2012). Splácení těchto vypůjčených peněz probíhá obvykle ve splátkách v pravidelných intervalech. Pokud takovéto půjčování peněz není řádně kontrolováno, může vést až k finanční krizi. Příkladem může být laxní poskytování úvěrů v USA, které značně přispělo k finanční krizi v letech 2007 – 2009.

Úvěrové skórování (credit scoring) znamená proces přiřazování kvantitativních měr, neboli skóre potencionálnímu dlužníkovi, kdy toto skóre představuje odhad budoucí výkonnosti

úvěru (Feldman 1997). Ve skutečnosti je úvěrové skórování proces používaný pro predikci dobrého a špatného klienta. Dobrým klientem je ten, který řádně splácí a neselže se splácením úvěru. Špatný klient je takový, který při splácení selže, tzn. „defaultuje“. Základem úvěrového skórování je model, který na základě vstupních parametrů dokáže odlišit dobré a špatné klienty. Výsledkem modelu je skórovací karta (scorecard), skóre nebo klasifikátor, tzn. nástroj, který shrnuje mnoho prediktivních vlastností do jediného modelu.

Tradiční metody používané k vývoji skórovacích karet jsou diskriminační analýza, logistická regrese či rozhodovací stromy (Thomas et al. 2017). Dle účelu, pro který se skórovací karta vytváří, lze rozlišit tyto typy karet (Nguyen 2015):

- Aplikační skórovací karta (application scorecard);
- Behaviorální skórovací karta někdy nazývaná jako výkonnostní skórovací karta (behavioral scorecard, performance scorecard);
- Skórovací karta podvodů (fraud scorecard).

Za poskytování úvěrů byli dříve zodpovědní úředníci, kteří se rozhodovali subjektivně. Časem došlo k představení sekvenčního rozhodovacího procesu pro podniky, které chtěly získat úvěr (Mehta 1968).

Několik autorů potvrdilo skutečnost, že klíčovou výhodou automatických metod úvěrového skórování je doba zpracování. V roce 1985 P. Chalos porovnával výsledky mezi rozhodováním úředníka a modelu o poskytnutí úvěru a dospěl k závěru, že úvěrové skórování (model) předčilo úředníka. Ačkoliv úvěrová komise označila za nejlepší metodu rozhodování úředníka, doba zpracování úvěrového skóre byla v případě modelu významně nižší (Chalos 1985). W. Alexander v roce 1989 poukázal na klíčovou výhodu úvěrového skórování, kterou je čas potřebný k lustraci žadatele o úvěr. Odhadoval, že model by potřeboval jen 5 až 6 minut k hodnocení žadatele (Alexander 1989). Na základě empirické studie kanadské banky K. Leonard v roce 1995 srovnával počet dní potřebných pro lustraci žadatele o úvěr před a po zavedení úvěrového skórování. Bez modelu zabralo hodnocení žadatele 9 dní. S modelem to byly pouze 3 dny (Leonard 1995). Z. Bilgin a U. Yavas rovněž doporučili upřednostnit používání automatického úvěrového skórování (Bilgin a Yavas 1995). V roce 2000 M. Banasiak a G. Kiely doporučili použití úvěrového skórování finančním institucím při klasifikaci úvěrů pro zvýšení výkonnosti a kratší dobu zpracování (Banasiak a Kiely 2000).

Výhodou úvěrového skórování oproti manuálnímu hodnocení úředníkem bylo také snížení nákladů na poskytnutí úvěru (Barefoot 1996). Další předností úvěrového skórování byla vysoká přesnost. D. Witkowska poukázala na slabiny spoléhání se na úvěrové úředníky jako například náklady na školení, dlouhou dobu zpracování, nepřesnost a doporučila automatizaci rozhodování při řízení úvěrového rizika (Witkowska 2006). Někteří autoři zjistili, že kombinace ručního a automatického rozhodnutí by mohla vést k ještě přesnějšímu rozhodování, což uvedl například R. Edmister (Edmister 1988).

Poslední klíčovou výhodou úvěrového skórování oproti subjektivnímu rozhodování jsou různé možnosti aplikace. R. Avery a kol. zmínil, že úvěrové skórování by se mohlo využívat pro stanovení cen úvěrů resp. nastavení úrokových sazeb (Avery et al. 2000). A. Sandler a kol. rovněž uvedl, že úvěrové skórování by mohlo být použito pro stanovení ceny úvěru (Sandler et al. 2000). L. Punchová uvedla hned několik možností, jak využít skórovací kartu, a to pro účely schválení/zamítnutí žádosti, nastavení úrokové sazby úvěru a predikci ziskovosti úvěrového portfolia (Punch 2000).

První číselné skórovací systémy byly vyvinuty pro zásilkový obchod a byly založeny na diskriminační analýze (Capon 1982). V roce 1936 R. Fisher použil diskriminační analýzu ke klasifikaci druhového jména rostlin kosatců na základě čtyř pozorovaných znaků (Fisher 1936). D. Durand v roce 1941 naznačil, že pomocí diskriminační analýzy by bylo možné rozlišit dobré a špatné úvěry (Durand 1941). J. Myers a W. Cordner vytvořili další jednoduchou studii na posuzování úvěrů pomocí této statistické metody (Myers a Cordner 1957). Historicky je tedy diskriminační analýza nejstarší technikou, která je prezentována v literatuře. Po diskriminační analýze následovala lineární regrese, která otevřela dveře i pro jiné techniky jako logistickou regresi, neparametrické metody vyhlazování, matematické programování, Markovovy řetězce, expertní systémy, genetické algoritmy, neuronové sítě a podmíněné modely nezávislosti. Od roku 1960 bylo úvěrové skórování studováno ve stále větší míře, a to zejména kvůli limitovanému použití diskriminační analýzy (Rosenberg a Gleit 1994). Nejznámější modely, zejména časově proměnné modely, byly prezentovány v 60. a 70. letech 20. století. Příkladem je Cyert Davidson Thompsonův model pro podezřelé účty z roku 1962 (Cyert et al. 1962) či Bierman Hausmanův model poskytování úvěrů (Bierman a Hausman 1970). V 80. letech byla diskriminační analýza odložena. Jedná se totiž o parametrickou metodu, jejíž předpoklady většinou nejsou na reálných úvěrových datech splněny.

Zájem se obrátil směrem k expertním systémům. Autoři začali poukazovat na další problémy, které vyžadovaly další výzkum. V roce 1982 N. Capon přezkoumal různé problémy, kterým se čelilo v procesu implementace úvěrového skórování (Capon 1982). Průzkum v roce 1998 ukázal, že přibližně 82 % bank využívá expertních systémů jako nástroj pro spotřebitelské a hypoteční úvěry. Tento průzkum odhadoval, že náklady na vybudování a implementaci skórovacích karet byly mezi 50 000 a 100 000 dolary (Sakai 1998).

Zájem o úvěrové skórování nezískali pouze vědci, matematici, statistici či informatici, ale také ekonomové či ekonometři. Mezi nejznámější výzkumníky z oblasti ekonomie byl H. Bierman a W. Hausman, kteří představili úvěrovou politiku včetně části, která souvisela s procesem poskytování úvěrů. Čím konzervativnější úvěrová politika, tím se banka snaží podstupovat co nejnižší riziko a omezuje tak své prodeje a tím i zisk (Bierman a Hausman 1970). H. Bierman a W. Hausman se zaměřili hlavně na proces poskytování, zatímco J. Stiglitz a A. Weiss (Stiglitz a Weiss 1983) se zaměřili více na účinky, které souvisely s úrokovými sazbami. W. Greene v roce 1998 popsal několik důležitých problémů, které souvisely s úvěry na kreditních kartách. Jeho článek se snažil odhadnout pravděpodobnost selhání (defaultu) na úvěrech z kreditních karet a rozšířil svou analýzu o odhad výdajů spotřebitele (Greene 1998). Y. Sakai přezkoumal Greeneův článek a dodal, že problém defaultu není pouze v dlužníkovi samém, ale také ve společnosti. Pokud je ekonomika stabilní a dobře funguje, tak se méně lidí dostává do problémů se splácením. Společnosti se ve skutečnosti zabývají mnohem více ziskem než defaultem. Y. Sakai naznačil, že budoucí výzkum se bude věnovat zisku, konkrétně jaké zisky se budou vztahovat k defaultu (Sakai 1998). Tyto studie ukázaly, že úvěrové skórování závisí také na úvěrové politice, úrokových sazbách nebo ekonomických faktorech. Nemělo by se zaměřovat pouze na predikci defaultu, ale také na zisk.

1.2 Úvěrové skórování v řízení rizik a typy skórovacích karet

Systém řízení rizik by měl zahrnovat systém pro skórování, úvěrová pravidla a výjimky. Systém úvěrového skórování je soubor modelů, které jsou vytvořeny za účelem zlepšení alokace úvěrů. Použitím správného modelu se finanční instituce může zaměřit na konkrétní cíl a minimalizovat rizika. Například použitím kombinace tří modelů: aplikačního modelu a behaviorálního modelu pro krátké a dlouhé období, může finanční instituce definovat profil žadatelů, kteří by měli být schváleni a s jakou výší úvěrového limitu. Závislou proměnnou

při výstavbě modelu je proměnná, která označuje dobrého a špatného klienta/žadatele. Tato proměnná je tedy binární. Špatnými klienty jsou oproti těm dobrým ti, kteří defaultují, tj. selžou se splácením svého úvěru. Jinak řečeno, podle Basilejského výboru pro bankovní dohled znamená default 90 a více dní po splatnosti. Dle potřeby lze využít i jiný počet dní, nejčastěji 30 či 60 dní v prodlení. Dobří klienti jsou takoví klienti, kteří pravidelně a řádně platí své závazky.

Terminologie „dobrý“ a „špatný“ byla použita například v pracích autorů J. Banasik a kol. (Banasik et al. 2001), W. Boyes a kol. (Boyes et al. 1989), M. Chen a S. Huang (Chen a Huang 2003), M. Desai a kol. (Desai et al. 2004) a v mnoha dalších. Tuto terminologii tedy využívá většina autorů. Méně častou záležitostí je rozlišování úvěrů do více kategorií. Příkladem může být A. Steenackers a M. Goovaerts (Steenackers a Goovaerts 1989), kteří rozlišovali „dobrý“, „špatný“ a „zamítnutý“ (good, bad, refused) úvěr. N. Šarlijová a kol. (Šarlija et al. 2004) zase rozlišovala „dobrý“, „chabý/mizerný“ a „špatný“ (good, poor, bad) úvěr. Další názvosloví použili Y. Kim a S. Sohn v roce 2004, kteří úvěry nejprve rozdělili do dvou skupin na „dobré úvěry“ a „špatné úvěry“ (good credit, bad credit) (Kim a Sohn 2004). Pak byli klienti pomocí modelu klasifikováni do čtyř skupin. První skupinou byli klienti, kteří nemají prodlení, a není pravděpodobné, že se v budoucnu zpozdí. Druhá skupina zahrnovala klienty, kteří nemají prodlení, ale je pravděpodobné, že v budoucnosti ho mít budou. Třetí skupinu tvořili klienti, kteří byli v současné době delikventní (v prodlení), ale zaplatí a čtvrtá skupina byla složena z klientů, kteří jsou v současné době delikventní a nezaplatí. Autoři odvodili charakteristiky klientů v každé skupině a navrhli strategii řízení, která odpovídala charakteristikám skupin (Kim a Sohn 2004).

Skóre má několik možných aplikací. S. Frame a kol. (Frame et al. 2001) tyto aplikace shrnuli:

1) stanovení úvěrového limitu,
2) automatické zamítání žádostí o úvěr,
3) úvěrové podmínky, kdy na základě výše rizika je přiřazena cena,
4) srovnání skóre zjištěného pomocí úvěrových registrů a interního modelu.

Ve většině literatury se ovšem úvěrové skórování používá pro zamítání žadatelů a mnohem méně se používá pro stanovení úvěrového limitu. Pokud jde o nastavení úvěrového limitu, existuje možnost vytvoření modelu profitability (ziskovosti), včetně behaviorálního skóre.

Kvůli složitosti zohlednění různých finančních prvků, lze budoucí zisk z jednotlivého klienta obtížně definovat.

Tato podkapitola si klade za cíl popsat různé druhy skórovacích karet, konkrétně se jedná o aplikační, behaviorální a skórovací karty podvodů. Existují i další typy skórovacích karet jako tzv. „churn scorecard“, která se vytváří s cílem určit, kteří žadatelé mají největší pravděpodobnost zůstat po dlouhou dobu u jedné organizace (finanční instituce). Dalším typem může být „collection scorecard“, která se snaží odhadovat, jak dobře bude dlužník vymáhán. K výstavbě skórovacích karet jako zmíněná „churn“ skórovací karta či „collection“ skórovací karta se využívají stejné metody, pouze s jinými proměnnými.

1.2.1 Aplikační skórovací karta

Aplikační skórování při hodnocení úvěrového rizika žadatele spojuje vlastnosti žádosti a bonitu žadatele. Platební vzory se statisticky mohou identifikovat po minimálně tří až šestiměsíčním období (v závislosti na typu úvěru). Toto skóre je dáno rizikem defaultu, které je spojené s každým klientem v závislosti na jeho aplikačních datech a úvěrové historii. Příkladem velmi jednoduché aplikační skórovací karty je tabulka 1. Například, pokud rodinný stav žadatele je rozvedený, tzn. hodnota v tabulce je 2, pak odpovídající výsledek je 45 bodů. Skóre je součtem dílčích skóre pro každou hodnotu proměnné. Čím vyšší skóre, tím je nižší úvěrové riziko žadatele. Z příkladu aplikační skórovací karty (tabulka 1) vyplývá, že pro rozvedeného žadatele ve věku 33 let bude konečný výsledek 99 bodů, naproti tomu ovdovělý žadatel ve věku 33 let bude mít 146 bodů.

Tabulka 1: Příklad aplikační skórovací karty

Proměnná	Hodnota	Interpretace	Skóre
Věk žadatele	<= 24	Méně než 24 let	16
	<= 30	25 – 30 let	33
	<= 40	31 – 40 let	54
	<= 50	41 – 50 let	47
	<= 60	51 – 60 let	62
	> 60	Více než 60 let	55
Rodinný stav	0	Ženatý / vdaná	89
	1	Svobodný / svobodná	38
	2	Rozvedený / rozvedená	45
	3	Druh / družka	80
	4	Vdovec / vdova	92

Zdroj: vlastní

Ve skutečnosti skórovací karty obsahují větší počet proměnných, než je uvedený počet v tabulce výše.

1.2.2 Behaviorální skórovací karta

Behaviorální skórování využívá nedávného chování klientů, aby předpověděl potenciální riziko, že klient selže se splácením. Aplikační skórování zahrnuje aplikační data o klientovi (věk, výše příjmu, vzdělání, délka zaměstnání aj.), zatímco behaviorální skórování využívá proměnné, které se vztahují k historii klienta, což jsou proměnné, které jsou důsledkem splácení a chování klienta. Chování je jedním z prvků, které vysvětluje pravděpodobnost defaultu klienta. V literatuře se doporučuje použít období 12 a 24 měsíců (García 2017). Musí být nadefinované datum ukončení pozorování a stav klienta (selhal / neselhal, default / nedefault, špatný / dobrý klient). K tomuto datu se definuje proměnná „výsledek“. Datum zahájení pozorování je obvykle stanoveno na 12 – 24 měsíců před datem ukončení pozorování. Toto období je v cizojazyčné literatuře nazýváno „performance period“ (Thomas et al. 2002). Všechny údaje o klientech v tomto období se mohou využít pro modelování.

1.2.3 Vymáhací skórovací karta

Skórovací karta pro vymáhání (collection scorecard) se zaměřuje na predikci aktivit souvisejících s vymáháním, které by měla finanční instituce poskytující úvěry provádět. Využívají se podobné proměnné jako při behaviorálním skórování. Tyto skórovací karty se snaží předpovědět nejlepší možnou akci, kterou lze udělat s klientem, který je v prodlení. Možnosti akcí jsou například (Thomas et al. 2017):

1) pokračovat/nevšímat si toho/čekat,
2) předání vymáhací agentuře,
3) zahájení právního procesu (žaloba),
4) odprodej pohledávky,
5) odepsání pohledávky z účetnictví.

Skórovací karta pro vymáhání udává pravděpodobnost vymožení škody/dlužné částky. Některé akce vedou ke stejným výsledkům, proto konečné rozhodnutí závisí na lidském faktoru a s tím souvisejících nákladech (Šarlija et al. 2004).

1.2.4 Skórovací karta podvodů

Podvody mohou být detekovány zavedením skórovací karty, která předpovídá, kteří klienti budou nejspíše podvodníci/selžou se splácením úmyslně. Podvodem v této souvislosti se myslí jakékoliv jednání, které vede k neoprávněnému získání úvěru. Skórovací karty pro podvody jsou vytvářeny stejnými metodami jako třeba aplikační skórovací karty. Vychází ze zkušeností z dřívějších případů a jsou založeny na hypotéze, že bude sledován stejný trend. Výsledkem je binární výstup, a to buď „opravdový" klient, nebo podvodník.

Hlavním rozdílem je, že talentovaní podvodníci a jejich žádosti o úvěr vypadají originálně. Proto vyhodnocení prevence proti podvodům za použití skórovacích karet podvodů neprokázalo smysl. Tyto karty neprokázaly schopnost odlišit skutečné žádosti od těch podvodných (Šarlija et al. 2004).

Podvodné žádosti je možné odhalit/detekovat až po určité době sledování chování klientů. Při výstavbě skórovací karty podvodů je podstatné definovat profil podvodného žadatele. Sledovat se mohou běžné transakce, četnosti, typické hodnoty, druhy zakoupených zboží/služeb, typy transakcí, rovnováha platební historie, výdajové vzory na denní, týdenní či měsíční bázi (Šarlija et al. 2004; Thomas et al. 2017).

1.3 Současné přístupy

Modelování úvěrového rizika se zabývá vývojem empirických modelů, které podporují rozhodování v komerčním či retailovém (spotřebitelském) úvěrování. Vzhledem k ekonomickému významu spotřebitelských úvěru, jejichž objem stále roste, si modelování úvěrového rizika v retailu získává v literatuře stále větší pozornost. Ke správě portfolia spotřebitelských úvěrů využívají bankovní i nebankovní finanční instituce řadu opatření. Mezi nejdůležitější lze zařadit očekávanou ztrátu, která je vyjádřena takto (Nazemi et al. 2017; Papoušková a Hájek 2019):

$$EL = PD \times EAD \times LGD, \tag{1}$$

kde:

EL	očekávaná ztráta (expected loss),
PD	pravděpodobnost defaultu (probability of default),
EAD	expozice při defaultu (exposure at default),
LGD	ztráta při defaultu (loss given default).

Pravděpodobnost defaultu je známa také jako úvěrové skóre, které odkazuje na pravděpodobnost, že se klient dostane do defaultu (prodlení se splácením úvěru) v určitém časovém horizontu. Expozice v defaultu představuje očekávanou velikost expozice v době selhání a ztráta při defaultu je část dané expozice, u které se očekává, že bude ztracena, pokud se klient dostane do defaultu (Papoušková a Hájek 2019).

Existuje velké množství literatury, která se zaměřuje na odhad pravděpodobnosti defaultu pomocí strojového učení (Harris 2015; Louzada et al. 2016). Modelování pravděpodobnosti defaultu je klasifikační úloha zaměřená na predikci, zda úvěr bude v defaultu nebo ne. Výsledek, tzn. zda úvěr je nebo není v defaultu, závisí na příslušném klasifikátoru. Pravděpodobnost defaultu není postačujícím ukazatelem pro posouzení úvěrového rizika, jelikož je nutné vzít v úvahu také ekonomickou ztrátu při defaultu. Část nesplácených úvěrů totiž může být splacena zcela, částečně nebo vůbec, což závisí například na vymáhacím procesu dané finanční instituce (Loterman et al. 2012). Z tohoto důvodu literatura z poslední doby naznačuje, že by se pozornost měla také soustředit na jiné klíčové problémy jako je kalibrace skórovací karty a modelování ztráty a expozice při defaultu (Lessmann et al. 2015).

Modelování ztráty a expozice při defaultu je pro účastníky úvěrového trhu zajímavé, neboť tyto parametry vstupují do výpočtu kapitálových požadavků a rizikově vážených aktiv. Dále se tyto ukazatele používají například při stresovém testování (Tong et al. 2016). Predikce těchto parametrů je náročná kvůli nenormálnímu rozdělení těchto dvou parametrů. V posledních letech se z toho důvodu zvýšil zájem o modelování expozice při defaultu (Leow a Crook 2016) i ztráty při defaultu (Yao, Crook, & Andreeva, 2017). Tyto modely však byly použity odděleně od modelování pravděpodobnosti defaultu, což neumožňuje modelovat celkovou očekávanou ztrátu. Většina těchto studií byla navíc omezena na použití jednotlivých metod predikce (prediktory) ztráty a expozice při defaultu. Málo pozornosti bylo věnováno kombinaci modelů (ensemble learning) (Nazemi et al. 2017) i přesto, že nedávné srovnávací studie prokázaly, že kombinace modelů z hlediska přesnosti predikce překonává jednotlivé prediktory v modelování úvěrového skóre (Abellán a Castellano 2017).

1.3.1 Predikce pravděpodobnosti defaultu

Úvěrové skórování má za cíl přiřadit úvěrové skóre žádosti o úvěr. Toto skóre je založené na predikci klasifikačního algoritmu. Skóre je tedy reprezentováno předpokládanou pravděpodobností defaultu. Komplexní porovnání současných klasifikačních algoritmů je možné

najít v článku S. Lessmana a kol. z roku 2015 (Lessmann et al. 2015). Tato srovnávací studie použila jak jednotlivé klasifikátory (vícevrstvá neuronová síť, podpůrné vektorové stroje, extrémní učící stroje aj.), tak kombinace modelů (náhodné lesy, stochastický gradientní boosting, aj.). Autoři dospěli k závěru, že kombinace modelů v úvěrovém skórování překonává jednotlivé klasifikátory. Přesněji řečeno, klasifikátory náhodné lesy (random forest) a selekce kombinace modelů (ensemble selection) vycházely nejlépe ze všech použitých homogenních a heterogenních kombinací modelů, což je také v souladu s předchozími souvisejícími studiemi (Hsieh a Hung 2010; Yu et al. 2010; Zhang et al. 2010).

Homogenní kombinace modelů jako bagging nebo boosting představují jeden klasifikační algoritmus, který se liší ve způsobu manipulace s trénovacími daty nebo vstupními a výstupními proměnnými. Finální klasifikace se obvykle získá kombinací základní metody a většinového nebo váženého hlasování (majority / weighted voting) (Tsai a Hung 2014). Efektivní základní metodou mohou být na základě nedávných empirických důkazů rozhodovací stromy (Abellán a Castellano 2017).

Na rozdíl od homogenních metod kombinují heterogenní metody různé klasifikační algoritmy pro dosažení vyšší rozmanitosti základních metod. Takto se mohou základní metody lépe vzájemně doplňovat (Ala'raj & Abbod, 2016a, 2016b). Kromě jednoduchého nebo váženého hlasování (simple / weighted voting) lze použít i složitější algoritmy pro získání kombinace modelů (Dzeroski a Zenko 2004). Například hluboké neuronové sítě ukazují dobrou výkonnost v kombinování jednotlivých predikcí extrémních učících se strojů (Yu et al. 2016).

Je také možné optimalizovat proces výběru základních metod (base learners). Předmětem výběru základních metod je výběr nejlepší podskupiny těchto metod z dané množiny. Konkrétně kombinace modelů v takovém případě buď maximalizují přesnost klasifikace, nebo optimalizují rozmanitost klasifikátorů (Florez-Lopez a Ramon-Jeronimo 2015).

Ačkoli kombinace modelů ve výše uvedených studiích obecně převyšují jednotlivé klasifikátory co se týká přesnosti klasifikace, nedávná literatura naznačuje, že i hluboké neuronové sítě (deep neural networks) mohou být z hlediska přesnosti klasifikace konkurenceschopné (Tomczak a Zieba 2015; Luo et al. 2017). To lze přisuzovat skutečnosti, že tyto modely umožňují aproximovat jakékoli rozdělení dat ve vstupním a výstupním prostoru, který je reprezentován žadateli o úvěr (Papoušková a Hájek 2019).

Dalším problémem v modelování pravděpodobnosti defaultu pomocí strojového učení je skutečnost, že většina reálných datových souborů je nevyvážená ve prospěch dobrých úvěrů, tj. úvěrů, které nejsou v defaultu. Použitím technik k vyvážení trénovacích dat lze však přesnost zvýšit (Marqués et al. 2013; Crone a Finlay 2012). Crone a Finlay upřednostnili nadvzorkování menšinové třídy (oversampling) (Crone a Finlay 2012), zatímco podvzorkování většinové třídy (undersampling) upřednostnil Brown a Mues (Brown a Mues 2012). Konkrétně výsledky autorů Brown a Mues ukazují, že kombinace modelů jako náhodné stromy a gradientní boosting lze dále zpřesnit podvzorkováním většinové třídy. Přístup podvzorkování jedné ze tříd byl také použit pro kontrolu silně nevyvážených tříd úvěrů v dalších nedávných studiích (He et al. 2018). Alternativně mohou být použity k převzorkování klasifikátory citlivé na náklady, které přiřazují vyšší váhu případům v menšinové třídě (Bahnsen et al. 2015).

Tabulka 2 je přehledem studií pro modelování pravděpodobnosti defaultu pomocí kombinace modelů.

Tabulka 2: Přehled studií modelování pravděpodobnosti defaultu pomocí kombinace modelů

Studie v chronologickém řazení	Metoda
(Hsieh a Hung 2010)	Neuronové sítě, podpůrné vektorové stroje, Bayesovská síť
(Martens et al. 2010)	Podpůrné vektorové stroje
(Twala 2010)	Kombinace modelů
(Yu et al. 2010)	Podpůrné vektorové stroje
(Zhang et al. 2010)	Vertikální bagging model rozhodovacích stromů
(Zhou et al. 2010)	Podpůrné vektorové stroje
(Li et al. 2011)	Podpůrné vektorové stroje
(Finlay 2011)	Neuronové sítě, bagging, boosting
(Ping a Yongheng 2011)	Neuronové sítě, podpůrné vektorové stroje
(Wang et al. 2011)	Bagging, boosting, Stacking (kombinace logistické regrese, rozhodovacích stromů, neuronových sítí a podpůrných vektorových strojů)
(Yap et al. 2011)	Logistická regrese, rozhodovací stromy
(Yu et al. 2011)	Neuronové sítě, podpůrné vektorové stroje
(Akkoç 2012)	Hybridní adaptivní neuro-fuzzy inferenční systém, diskriminační analýza, logistická regrese, neuronové sítě
(Brown a Mues 2012)	Logistická regrese, neuronové sítě, rozhodovací stromy
(Hens a Tiwari 2012)	Podpůrné vektorové stroje

(Li et al. 2012)	Podpůrné vektorové stroje, kombinace modelů
(Marqués et al. 2012a)	C4.5 rozhodovací strom, neuronová síť typu perceptron, logistická regrese, Bayesovská síť, kombinace modelů
(Marqués et al. 2012b)	Neuronové sítě, podpůrné vektorové stroje, kombinace modelů
(Kruppa et al. 2013)	Kombinace modelů
(Abellán a Mantas 2014)	Neuronové sítě, kombinace modelů
(Tsai 2014)	Neuronové sítě, kombinace modelů
(Tsai a Hung 2014)	Neuronové sítě (hybridní neuronové sítě)
(Harris 2015)	Shlukovací podpůrné vektorové stroje, podpůrné vektorové stroje
(Tomczak a Zieba 2015)	Klasifikační Restricted Boltzmann Machines
(Lessmann et al. 2015)	Neuronové sítě, Bayesovská síť, rozhodovací stromy, metoda nejbližšího souseda, logistická regrese, diskriminační analýza, extrémní učící stroje, neuronové sítě, podpůrné vektorové stroje, homogenní kombinace modelů (alternující rozhodovací strom, bagging, boosting, náhodné stromy, náhodné lesy), heterogenní kombinace modelů
(Ala'raj a Abbod 2016a)	Neuronové sítě, podpůrné vektorové stroje, náhodné lesy, rozhodovací stromy, Bayesovská síť
(Ala'raj a Abbod 2016b)	Neuronové sítě, podpůrné vektorové stroje, náhodné lesy, rozhodovací stromy, Bayesovská síť, logistická regrese
(Louzada et al. 2016)	Logistická regrese, neuronové sítě, rozhodovací stromy, podpůrné vektorové stroje, diskriminační analýza, Bayesovská síť, lineární regrese, kombinace modelů
(Yu et al. 2016)	Hluboké neuronové sítě, extrémní učící stroje
(Abellán a Castellano 2017)	Kombinace modelů (bagging, boosting, random subspace, Decorate, rotační les)
(Luo et al. 2017)	Hluboké neuronové sítě, Restricted Boltzmann Machines, Logistická regrese, neuronová síť typu perceptron, podpůrné vektorové stroje

Zdroj: vlastní

1.3.2 Predikce ztráty při defaultu a expozice při defaultu

Do nedávné doby bylo modelování úvěrového rizika pomocí metod strojového učení omezeno na modelování pravděpodobnosti defaultu. Pravděpodobnost defaultu je ovšem pouze jednou složkou současných modelů používaných při řízení úvěrového rizika, neboť pro jeho

celkový odhad je nutné znát i ztrátu danou defaultem a expozici při defaultu (Papoušková a Hájek 2019).

Parametr ztráta při defaultu měří ztrátu vyjádřenou v procentech expozice při defaultu (Loterman et al. 2012) a ve skutečnosti je to klíčový parametr, který vstupuje do výpočtu kapitálového požadavku. V praxi jsou modely využívané pro odhad pravděpodobnosti defaultu založeny na binárních klasifikátorech, zatímco modely pro odhad ztráty při defaultu jsou považovány za regresní problémy. To znamená, že první přístup klasifikuje úvěry do dvou tříd (dobrý a špatný úvěr) a druhý přístup odhaduje ztrátu danou defaultem (Papoušková a Hájek 2019).

Tradiční metody, které se k odhadu ztráty při defaultu využívají jsou lineární regrese (metoda nejmenších čtverců) (Caselli et al. 2008) nebo regresní stromy (Bastos 2010). Nelineární regresní modely jako podpůrné vektorové stroje ovšem významně převyšovaly přesnost lineární regrese při predikování ztráty při defaultu firemních dluhopisů (Yao et al. 2015). Podobně jako u modelování pravděpodobnosti defaultu lze výkonnost jednotlivých prediktorů zlepšit kombinováním různých metod. Například Nazemi a kol. ve své studii využili fuzzy regresi, logistickou regresi, podpůrnou vektorovou regresi (Support vector regression) a regresní stromy (Nazemi et al. 2017). Ještě lepšího výsledku lze dosáhnout pomocí dvoustupňové metodiky, která využívá skutečnosti, že datové soubory pro ztrátu při defaultu jsou silně nevyvážené ve prospěch nulového, anebo plně navráceného úvěru ve vymáhacím procesu. Například metodu podpůrných vektorových strojů použili Yao a kol. pro klasifikaci nulové vs. nenulové návratnosti v první fázi a ve druhé fázi využili podpůrnou vektorovou regresi pro predikci ztráty při defaultu (Yao et al. 2017).

Parametr expozice při defaultu je definován jako očekávaná expozice z prodlení. Odhad expozice při defaultu závisí na vlastnostech úvěru. V případě osobních úvěrů či hypotečních úvěrů se jedná o nesplacenou částku v době výpočtu. U revolvingových expozic je expozice při defaultu rozdělena na čerpaný a nevyčerpaný závazek. Predikční modely odhadují expozici při defaultu buď přímo nebo nepřímo pomocí tzv. kreditního konverzního faktoru (credit conversion factor). Kreditní konverzní faktor je definován jako procentní podíl současných nevyčerpaných závazků při selhání (Tong et al. 2016). Počáteční studie upřednostňovaly nepřímé přístupy (Yang a Tkachenko 2012). Nedávno však bylo prokázáno, že oba přístupy fungují podobně (Tong et al. 2016). Pro predikci expozice při defaultu kdykoli během období splácení úvěru byl navržen smíšený model (mixture model), který kombinuje dva

modely s náhodnými efekty pro odhad očekávaného zůstatku a očekávaného úvěrového limitu. Všechny tyto přístupy byly omezeny na jednotlivé prediktory expozice při defaultu (Papoušková a Hájek 2019). Přehled studií, které se zabývaly modelováním ztráty a expozice při defaultu, je uveden v tabulce 3.

Tabulka 3: Přehled studií modelování ztráty při defaultu (LGD) a expozice při defaultu (EAD)

Studie v chronologickém řazení	Parametr modelování	Metoda
(Caselli et al. 2008)	LGD	Lineární regrese
(Bastos 2010)	LGD	Regresní stromy
(Loterman et al. 2012)	LGD	Lineární regrese (logistická regrese) + podpůrná vektorová regrese (neuronové sítě)
(Yao et al. 2015)	LGD	Podpůrná vektorová regrese
(Nazemi et al. 2017)	LGD	Fuzzy regrese
(Yao et al. 2017)	LGD	Podpůrné vektorové stroje + podpůrná vektorová regrese
(Yang a Tkachenko 2012)	EAD	Neuronové sítě, boosting, smíšený model
(Tong et al. 2016)	EAD	Smíšený model
(Leow a Crook 2016)	EAD	Smíšený model
(Gürtler et al. 2018)	EAD	Lineární regrese

Zdroj: vlastní

1.4 Současné problémy při výstavbě skórovací karty

V této podkapitole jsou uvedeny hlavní problémy, se kterými je možné se setkat v úvěrovém skórování. Příkladem těchto problémů může být definice defaultu, velikost vzorku dat, chybějící či odlehlé hodnoty aj.

Hlavními problémy, které řeší tato podkapitola, jsou:

- Běžné problémy při výstavbě skórovacího modelu, a to jak po praktické, tak technické stránce.
- Opatření, která lze přijmout s cílem zlepšení skórovací karty.

Podle R. Raesida a J. Walkera (Raeside a Walker 2001) jsou hlavními praktickými problémy souvisejícími s dolováním dat při výstavbě skórovacího modelu tyto:

- Generování falešných vzorů a vztahů, kdy jsou proměnné a koeficienty modelu v rozporu.
- Model nemůže být postaven bez dostatku dat a dostatečně dlouhého období.
- Čištění dat je jedním z nejvíce časově náročných úkolů.
- Je obtížné určit účinnost a spolehlivost modelů, zatímco probíhá jejich implementace.
- Chování zákazníků a podmínky splácení úvěru se mohou v průběhu času měnit, proto modelování může být sporné v případě potřeby vybudování dlouhodobého modelu.

Výše uvedený seznam poskytuje informaci o výzvách, kterým čelí analytici, jejichž cílem je vytvoření skórovacího modelu.

Mezi technické problémy patří (Siddiqi 2012):

- Definice defaultu.
- Výběr vzorku.
- Chybějící data.
- Odlehlá pozorování.
- Validace.
- Analýza zamítnutých žádostí (reject inference).
- Multikolinearita.

Podle Basilejského výboru pro bankovní dohled je default definován takto (García 2017):

"Za default se považuje nesplnění povinnosti konkrétního dlužníka, pokud došlo k některé z těchto skutečností:

(a) instituce se domnívá, že je nepravděpodobné, aby dlužník splatil své úvěrové závazky v plné výši, aniž by udělala nějaké opatření;

(b) dlužník má za posledních více než 90 dní jakýkoli významný úvěrový závazek vůči instituci. 90 dní může být nahrazeno 180 dny pro expozice zajištěné nemovitostmi."

V oblasti kreditních karet se místo 90 dní po splatnosti využívá často 60 dní po splatnosti. Vzhledem k údajům o spotřebitelských úvěrech, se kterými se pracuje v této práci, bude za default považováno 60 dní po splatnosti.

Další otázkou je velikost vzorku. Skórovací modely jsou často vyvíjeny s nedostatečně velkým vzorkem dat pro dosažení spolehlivosti při úvěrovém skórování. Podle autorů

W. Henley a D. Hand (Henley a Hand 1996) se v praxi podíl špatných a dobrých klientů v populaci mění v závislosti na daném úvěrovém produktu.

Na kvalitě informací o klientovi ať už interních, nebo z úvěrových registrů závisí, kolik bude chybějících hodnot ve sledovaném vzorku. Při poskytování spotřebitelských úvěrů může dojít k tomu, že některé marketingové kampaně deklarují minimální dokladovost při sepsání úvěrové smlouvy. Žadatel o úvěr tedy poskytne pouze nutné informace o jeho osobě (jméno, rodné číslo, adresu) a doplňující informace, které by se hodily při výstavbě skórovací karty, nejsou k dispozici (vzdělání, typ bydlení, počet dětí, rodinný stav aj.). Ne všechny žadatele je možné nalézt v úvěrovém registru ať už negativním, nebo pozitivním. Chybějící informace představují obtíže. V ideálním případě by měli všichni žadatelé vyplnit dotazník a úvěrová společnost by akceptovala pouze ty, kteří dotazník opravdu vyplní.

Dalším problémem, který se vyskytuje při tvorbě skórovací karty, jsou odlehlá pozorování**,** tj. ty, která jsou daleko od většiny ostatních. Tyto hodnoty bývají označovány jako „extrémní“ nebo „nepravděpodobné“. Odlehlé hodnoty mohou být detekovány automaticky nebo manuálně např. pomocí jednoduchých histogramů.

Pro validaci modelů se běžně používá Giniho koeficient nebo Kolmogorovova-Smirnovova statistika. Některé společnosti či software využívají také informační hodnotu. Giniho koeficient je založen na kumulativní distribuci dobrých a špatných klientů. Kolmogorovova-Smirnovova statistika vyjadřuje odchylku mezi celkovým rozdělením dobrých a kumulativním rozdělením špatných klientů. Informační hodnota se získá vynásobením váhy evidence (weight of evidence) pro každou skupinu/pozorování s rozdílem procent dobrých a špatných klientů pro téže skupinu/pozorování (García et al. 2010; Siddiqi 2012). V literatuře není zmíněno, který ukazatel je lepší než ostatní. Často se používají všechny najednou.

K dalším problémům se řadí skutečnost, že při výstavbě skórovací karty vychází historie chování z již schválených klientů, tzn., do modelu nevstupují informace o klientech, kteří byli v minulosti zamítnuti. V případě, že míra schvalování je vysoká, tak podle N. Siddiqi (Siddiqi 2012) nebudou mít zamítnutí klienti význam, protože celé portfolio klientů bude podobné jako portfolio žadatelů. J. Banasik a kol. (Banasik et al. 2003) viděl prostor pro zlepšení na základě nasazení informací o zamítnutých žádostech, které by byly schváleny s cílem sledovat schopnost splácet. Výsledkem toho bylo velmi malé zlepšení v kvalitě modelu.

Problémem je také multikolinearita (Capon 1982), která nastává v případě vzájemné korelace mezi nezávislými proměnnými. Logistická regrese se tomuto problému vyhýbá, neboť proměnné, které do modelu přináší stejnou informaci, jsou z modelu automaticky vyloučeny. Kromě toho je model vystavěn na trénovacím vzorku tzn., že riziko nedetekované multikolinearity se sníží na testovacím vzorku. Důležitým prvkem je také porozumění datům (viz metodika CRISP-DM v kapitole 3), což je způsob, jak multikolinearitu zjistit.

Vedle problémů, se kterými se lze setkat při výstavbě skórovací karty existují i možnosti, jak skórovací karty zlepšit. V literatuře se nejčastěji uvádějí možnosti jako segmentace, kombinace dvou skóre či zahrnutí makroekonomických proměnných. Další část textu tyto možnosti popisuje.

Segmentace se v úvěrovém skórování zavádí s cílem ohodnotit různé segmenty pomocí jiného přístupu, skórovací karty. Segmentace vychází z kvality segmentu, nákladů na daný segment a popř. z výnosů, které bude daný segment generovat v budoucnu.

Kombinace dvou skóre znamená použití dvou různých skórovacích karet, z nichž každá je sestavena z různých proměnných.

Zahrnutím makroekonomických veličin se zabýval například B. Stine a W. Lang (Stine a Lang 2007), kteří testovali schopnost predikce skórovacího modelu po přidání makroekonomických proměnných. Zabývali se například mírou nezaměstnanosti. Zakomponování makroekonomických údajů do modelů může napomoci odhalit budoucí finanční problémy žadatelů.

1.5 Úvěrové informace

Úvěrové skóre je statistickým nástrojem, které shrnuje několik prediktivních vlastností do jediného modelu. Pomocí tohoto modelu je usnadněna implementace strategie, obchodní změny či monitorování. Rozhodující pro model je kvalita charakteristik použitých při výstavbě skórovací karty. S cílem získat co nejvíce informací o žadatelích se využívají informace z různých zdrojů. Hlavními zdroji jsou buď interní, nebo externí údaje. Interními zdroji se myslí údaje nashromážděné uvnitř společnosti. Tyto údaje obsahují například informace o žadateli či jeho platbách. Externími údaji bývají informace z různých registrů, ve kterých se sdílí informace mezi jednotlivými finančními institucemi či státními orgány (Siddiqi 2012).

Ve většině literatury se zanedbává výměna informací s ostatními věřiteli (finančními institucemi) jako alternativní způsob získávání informací. Sdílení informací o platební morálce klientů a jejich závazcích rozšiřuje nástroje, které je možné využít při řízení úvěrového rizika. Existují dva přístupy ohledně sdílení informací. Jedním jsou úvěrové registry (credit bureau), které shromažďují, ukládají a distribuují informace dobrovolně poskytnuté svými členy. Druhým přístupem jsou veřejné úvěrové registry, které bývají řízeny centrálními bankami. Jedná se o povinné oznamování údajů o dlužnících.

Úvěrové registry distribuují informace ze svých databází svým členům pomocí úvěrové zprávy. Jestliže finanční instituce poskytující úvěry odešle zašifrovaný požadavek pro jednoho konkrétního žadatele, úvěrový registr zpracuje žádost v dávkovém režimu nebo pomocí on-line systému a vrátí zašifrovanou úvěrovou zprávu žadatele, která obsahuje skóre (credit bureau's score) a detail o bankovní příp. nebankovní historii žadatele.

V České republice existuje několik úvěrových registrů, jsou jimi:

1) Veřejný úvěrový registr
    a) Centrální registr úvěrů (CRÚ);
2) Úvěrové registry
    a) Zájmové sdružení právnických osob (SOLUS);
    b) Bankovní registr klientských informací (BRKI, Czech Banking Credit Bureau);
    c) Nebankovní registr klientských informací (NRKI, Czech Non-Banking Credit Bureau).

Centrální registr úvěrů shromažďuje informace o úvěrech právnících a fyzických osob podnikatelů, tzn., že zde nejsou evidovány spotřebitelské či hypoteční úvěry fyzických osob ani jiné úvěrové produkty poskytnuté fyzickým osobám (ČNB 2019).

SOLUS je zájmové sdružení právnických osob, které sdružuje společnosti z různých odvětví (finanční instituce, telekomunikační operátoři, distributoři energií). Jednotlivé společnosti mezi sebou prostřednictvím SOLUS sdílí informace o svých klientech (SOLUS 2019).

Prostřednictvím bankovního a nebankovního registru klientských informací si jejich účastníci vyměňují informace o svých žadatelích. Databáze obou registrů obsahují pozitivní i negativní informace, které vypovídají o důvěryhodnosti a platební morálce klientů. Účastníci těchto registrů mají přístup jak k informacím o současném stavu zadlužení žadatele, tak o jejich historii (CBCB 2019; CNCB 2019).

2 Cíle disertační práce

V předcházející části práce byl uveden současný stav v oblasti úvěrového skórování a jeho modelování. Na základě uvedených nedostatků předchozích modelů byl stanoven následující cíl doktorské disertační práce: navrhnout model skórovacích karet, který umožní modelování pravděpodobnosti defaultu a expozice při defaultu pomocí heterogenních kombinací metod strojového učení. K dosažení tohoto cíle jsou definovány následující dílčí cíle práce:

1. Návrh modelu pravděpodobnosti defaultu využívající podvzorkování majoritní třídy nedefaultních úvěrů a heterogenní kombinaci klasifikátorů. Podvzorkování má za cíl vybalancování tříd pro klasifikaci. Jako základní klasifikátory jsou použity v současné době nejefektivnější metody klasifikace defaultních / nedefaultních úvěrů. Jejich heterogenní kombinací je dosaženo vyšší přesnosti a odolnosti proti přeučení. Jako meta-klasifikátor je použita jak tradiční logistická regrese, tak nelineární klasifikátor náhodný les. Ten brání přeučení a dále provádí vnořenou selekci výstupů základních metod. Tím se odlišuje od v současnosti používaných lineárních metod používaných v kombinaci modelů jako meta-algoritmy.
2. Návrh metriky pro hodnocení výsledku klasifikace modelu pravděpodobnosti defaultu. Kromě tradičních metrik hodnocení výsledků klasifikace navržená metrika bere v úvahu rozdílné náklady způsobené chybnou klasifikací defaultních a nedefaultních úvěrů. K tomu jsou použity náklady obětované příležitosti v případě dobrého (nedefaultního) úvěru chybně klasifikovaného jako špatného (defaultního) a ztráta investice v případě špatného úvěru chybně klasifikovaného jako dobrého. K výpočtu této metriky je použit třetí klíčový parametr úvěrového rizika, tj. ztráta při defaultu. Tato metrika umožňuje výpočet relativních nákladů pro porovnání ekonomických dopadů klasifikačních modelů.
3. Návrh modelu expozice při defaultu využívající heterogenní kombinaci regresorů. Jako základní regresory jsou použity v současné době nejefektivnější metody predikce expozice při defaultu. Jako meta-regresory jsou použity tradiční lineární regrese a náhodný les. V porovnání se současným stavem řešení tato kombinace modelů umožní dosažení vyšší přesnosti a modelování různorodých úvěrových profilů klientů.
4. Návrh dvoufázového modelu úvěrového rizika spotřebitelských úvěrů skládající se z modelu pravděpodobnosti defaultu a modelu expozice při defaultu. V první fázi je

použit model pravděpodobnosti defaultu využívající heterogenní kombinaci klasifikátorů k identifikaci defaultních úvěrů. Pro tyto úvěry je ve druhé fázi použit model expozice při defaultu využívající heterogenní kombinaci regresorů. Do druhé fáze tedy nevstupují všechny úvěry, ale pouze ty, u nichž se předpokládá default. Tím je zvýšena efektivnosti modelování ve druhé fázi. V porovnání se současným stavem řešení umožní tento model nejen predikci pravděpodobnosti defaultu, ale také predikci ekonomických dopadů defaultu každého úvěru.

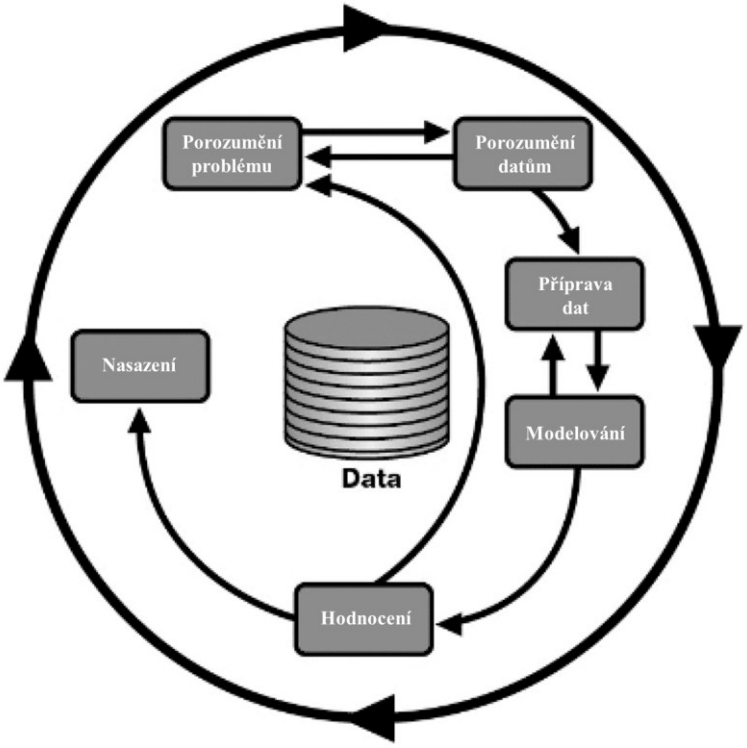
5. Verifikace navrženého modelu na reálných datech české nebankovní finanční instituce. Získaná data jsou předzpracována a vzhledem k nevyváženosti tříd v datech je provedeno náhodné podvzorkování většinové třídy.
6. Provést srovnávací analýzu výsledků navrženého modelu se: (a) v současnosti používanými modely pro predikci pravděpodobnosti defaultu a expozice při defaultu; (b) jednofázovými a dvoufázovými modely predikce úvěrového rizika. Toto porovnání je provedeno jednak z hlediska přesnosti predikce, jednak z hlediska nákladů chybné klasifikace a dále z hlediska celkového zisku dosaženého pomocí navrženého modelu.
7. Provést srovnávací analýzu výsledků navrženého modelu se současným modelem skórovací karty používaným ve vybrané nebankovní finanční instituci a kalibrace současné skórovací karty pomocí navrženého modelu.

3 Metodologie vytváření skórovacích karet

Velké banky mívají svou metodologii a možnost mezistátního srovnání v rámci jedné korporace. Částečně jsou banky ovlivňovány předpisy Basel II. Nicméně budování skórovací karty je typickou komerční datamingovou úlohou, pro kterou lze využít metodologii CRISP-DM (CRoss-Industry Standard Proces for Data Mining).

Metodologie CRISP-DM rozdělí celý proces data miningového projektu do šesti fází (Witten et al. 2011):

1. porozumění problému (business understanding),
2. porozumění datům (data understanding),
3. příprava dat (data preparation),
4. modelování (modeling),
5. hodnocení (evaluation),
6. nasazení (deployment).



Obrázek 1: Metodologie CRISP-DM

Zdroj: (Chapman et al. 2000)

3.1 Porozumění problému

Proces výstavby tradičních skórovacích karet začíná definováním obchodních cílů. Vstupní fáze se zaměřuje na pochopení problematiky z obchodního hlediska. Je nutné porozumět cíli projektu a požadavkům na řešení, která jsou formulována na manažerské úrovni.

V případě aplikačního skórování je cílem posouzení nově vznikajících úvěrových expozic s účelem optimalizovat proces schvalování úvěrů. Pro toto posouzení jsou typickými vstupy (Siddiqi 2012):

- socio-demografické údaje,
- bonita žadatele,
- účetní a ekonomická data (v případě podnikatelských úvěrů),
- úvěrové registry,
- úvěrová historie žadatele,
- parametry úvěru,
- zajištění.

Při behaviorálním skórování se posuzují stávající úvěrové expozice na základě jejich dosavadního vývoje a chování za účelem marketingových kampaní, předcházení vymáhání a optimalizace kapitálové přiměřenosti dle Basel II. Vstupními informacemi jsou:

- aplikační údaje a jejich vývoj,
- splácení úvěru (včasnost, prodlení),
- nakládání s přidělenými prostředky.

Optimalizací procesů řízení rizika nebo marketingových procesů lze (García 2017; Siddiqi 2012):

- zrychlit posouzení žádosti o úvěr,
- snížit náklady na posouzení žádosti,
- standardizovat postup posuzování,
- posuzovat žádosti objektivně,
- rychle aktualizovat kritéria,
- snížit rizika interního podvodu,
- automaticky a objektivně posoudit úvěrové expozice,
- vybírat klienty pro marketingové kampaně,

- dělat včasnou prevenci,
- stabilizovat ratingové třídy.

Pro vytvoření skórovací karty ve fázi „porozumění obchodu“ je nutné znát zdroje, kterých se výstavba bude týkat (Witten et al. 2011):

- Datové zdroje
    - Interní datové zdroje vč. historie
    - Externí úvěrové registry
- Lidské zdroje
    - Data mineři
    - Procesní specialisté
    - Management
    - Uživatelé skórovacích karet
    - IT specialisté
    - Bankovní dohled
- Software
    - Dataminignový
    - Databázový
    - Provozní
    - Implementační
- Hardware
    - Vývojový
    - Provozní

Požadavky, které je nutné definovat před vytvořením skóre karty, jsou tyto (Siddiqi 2012):

- Rychlost přístupu a vyhodnocení
    - On-line
    - Off-line dávkové
- Systémové požadavky prostřední
    - Integrace s provozním systémem
- Možnosti aktualizace a auditu
    - Transparentní datamingový projekt
    - Ukládání datových snímků

- Plán monitorování a aktualizace
    - Automatizace
    - Kritéria
- Implementace
    - Automatický export karty do produkce
    - Programátorská práce

Mezi potencionální rizika, která souvisí s touto fází, lze zařadit přístup k datům, spolupráci s IT specialisty, kvalitu dat a metadat, dostupnost historických snímků, podíl selhání (defaultů), přístup do úvěrových registrů, cenu implementace a intervenci regulátora.

Dataminingovým cílem je vytvoření algoritmu, který přiřadí skóre nebo ratingovou třídu jednotlivým žadatelům a zároveň bude celý postup transparentní a auditovatelný, včetně nezbytných manipulací se vstupními daty včetně výpočtu a využití skóre.

Kritérii úspěšnosti řešení, ať už obchodními, nebo datamingovými je bezesporu zkvalitnění dosavadního postupu, snížení kapitálové přiměřenosti, snížení podílu defaultujících úvěrů, zrychlení úvěrového procesu, snížení nákladů na posouzení žádosti, optimalizace nákladů na marketingovou kampaň, optimální nahrazení všech vstupních charakteristik jediným číselným skóre, maximalizace citlivosti postupu na rozpoznání defaultních úvěrů, stabilizace pravděpodobnosti defaultu v ratingových třídách nebo maximalilzace zacílení v marketingové kampani (García 2017).

3.2 Porozumění datům

Podstatnou částí projektu je přístup k datům nebo jejich získání. Nutné jsou informace o (García 2017; Thomas et al. 2017):

- žadateli či dlužníkovi (socio-demografické či firemní údaje, bonita, úvěrové registry, historie úvěrového chování na jiných úvěrech aj.),
- úvěru (typ produktu, výše úvěru, způsob splácení, splatnost aj.),
- zajištění úvěru (další osoby, movité a nemovité věci, vinkulace),
- splácení (obraty na účtech, prodlení se splácením v minulosti),
- skóre (aplikační u behaviorálního příp. naopak).

Veškeré informace lze získat napojením na různé datové zdroje, kterými jsou (García 2017):

- transakční databáze (pohyby na účtech, prodejní data, vymáhání, účetnictví),
- datové sklady a datová tržiště (časové snímky úvěrů),
- datové soubory (extrakty z databází, SPSS, SAS, MS Excel, texty – přímý přístup nebo přes převodníky),
- volné texty (posudky, maily, záznamy z call centra – přímý přístup nebo extrakce strukturovaných atributů),
- neelektronická data (tištěné žádosti, složky s papírovými dokumenty – nutné pořízení do elektronické podoby),
- data z externích registrů (bankovní, nebankovní registry, pojišťovny, Česká správa sociálního zabezpečení – nutnost stažení mnoha záznamů včetně historie).

3.3 Příprava dat

Příprava dat bývá obvykle nejnáročnější fází, neboť data bývají v různých formátech, v různých tabulkách, obsahují chybějící hodnoty či jiné atributy, než je potřeba pro analýzu nebo data chybějí úplně apod.

Systém propojených tabulek (relační databáze) je sice efektivní pro uložení dat a umožňuje zefektivnění vyhledávání, ale je nevhodný pro analýzy a modelování. Pro využití při modelování je nutné data denormalizovat, což zjednodušeně znamená spojit různé datové zdroje a převést je do jedné tabulky pomocí restrukturalizace, agregace, spojování tabulek aj. Jediná tabulka obsahuje řádky, které reprezentují žadatele/smlouvy a sloupce, které reprezentují

historické vlastnosti žadatelů/smluv. Některé sloupce mohou být pouze identifikátory, které se neúčastní modelování ani hodnocení (např. ID žadatele/smlouvy/produktu). Před modelovací fází je nutné některé proměnné kategorizovat, vyřešit chybějící a odlehlé hodnoty, vybrat proměnné a případy, které budou vstupovat do fáze modelování.

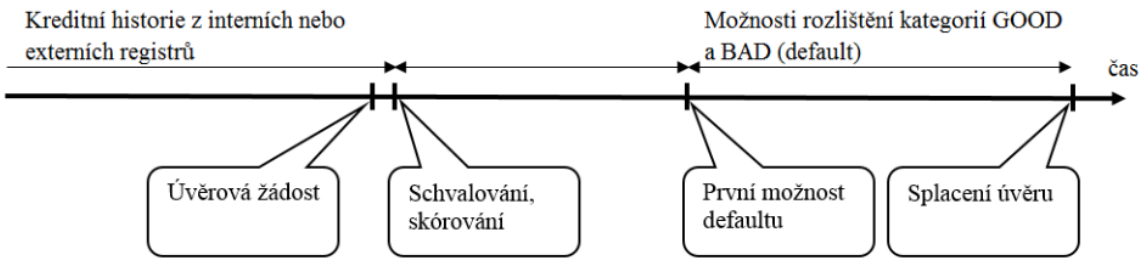
Kategorizace vstupů přináší své výhody, kterými jsou (Thomas et al. 2017):

- Vynechané hodnoty lze považovat za jednu z kategorií.
- Odlehlá pozorování jsou zařazena do krajních intervalů a nemají tak silný efekt.

Nevýhodou lze označit zvýšení počtu vstupů a závislost vstupů mezi sebou.

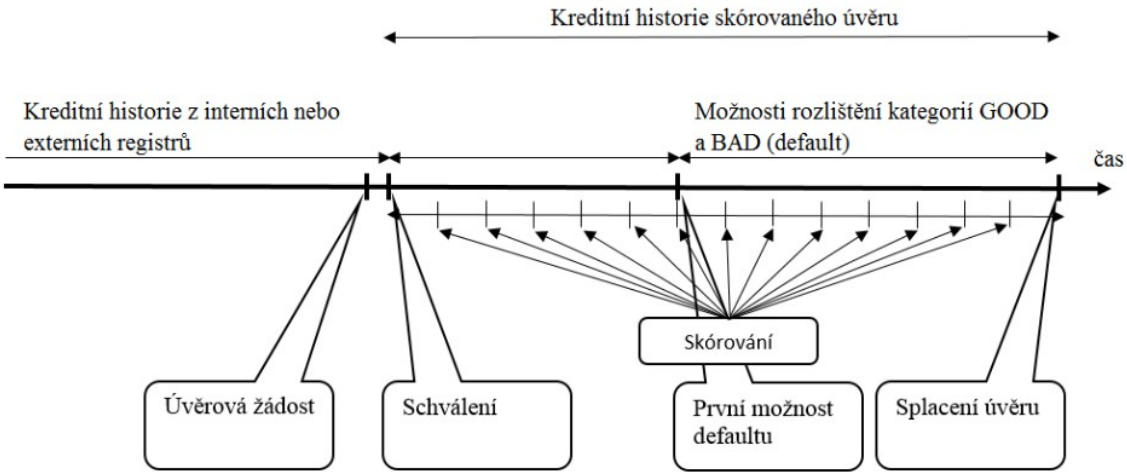
Vstupy lze kategorizovat expertně, a to na základě zvyklostí nebo podle předepsané metodiky (García 2017). Kategorizaci lze provést graficky pomocí histogramu rozdělení nebo výpočtem informační hodnoty, kdy se hledají hranice tak, aby informační hodnota vzniklých intervalů byla co nejvyšší. Vstupy se mohou kategorizovat také za pomoci rozhodovacích stromů, kdy se hledají hranice tak, aby se co nejvíce lišily podíly špatných klientů mezi stanovenými intervaly.

Do aplikačního skórování vstupují údaje o žadateli a údaje úvěrové historie z interních či externích registrů (obrázek 2). Naopak u behaviorálního skórování vstupují do modelování jednak údaje z aplikačního skórování a dále informace, které souvisí s chováním klienta v průběhu splácení (obrázek 3).



Obrázek 2: Časový vývoj úvěru z hlediska aplikačního skórování

Zdroj: vlastní



Obrázek 3: Časový vývoj úvěru z hlediska behaviorálního skórování

Zdroj: vlastní

Příkladem vstupních proměnných mohou být (Siddiqi 2012):

- Socio-demografické údaje (fyzické osoby nepodnikající):
  - Věk, pohlaví, vzdělání, region, velikost bydliště, rodinný stav, zaměstnání, shoda trvalé a korespondenční adresy, doba na současné adrese, doba v zaměstnání, průkaz totožnosti.
- Firemní údaje (fyzické osoby podnikající a právnické osoby):
  - Právní forma, počet zaměstnanců, obrat, odvětví, stáří, zisk, region.
- Bonita:
  - Příjem domácnosti, frekvence výplaty, příjem na hlavu, výdaje domácnosti, podíl příjmů a výdajů, počet úvěrů, počet dětí, počet členů domácnosti.
- Interní úvěrová historie:
  - Aplikační skóre, behaviorální skóre, počet aktivních úvěrů, počet splacených úvěrů, delikvence, dlužná částka, počet upomínek, doba do splatnosti, podíl splacené části dluhu.
- Interní platební morálka:
  - Současná delikvence, maximální delikvence, doba od poslední delikvence, počet delikvencí, počty upomínek, počet splátek.
- Registry:
  - Skóre, riziková skupina, počet aktivních úvěrů, počet splacených úvěrů, doba delikvence, dlužná částka.

- Nákupní košík:
    - Příznaky produktů nebo produktových skupin.
- Kontaktní historie:
    - Počet oslovení, počet stížností, příznaky kontaktních kanálů.
- Úvěrový produkt:
    - Produkt, výše úvěru, splatnost, frekvence a výše splátek, úrok, způsob výplaty, účel.
- Zajištění:
    - Výše, způsob, vlastnosti osob, vlastnosti nemovitosti, zůstatky na účtech.

Co se týče cílové proměnné, ta je predikována modelem. Nejvhodnější cílovou proměnnou je dichotomická proměnná (binární), tj. kategorie GOOD (dobrý) a BAD (špatný = default). Někdy se také používá střední kategorie, která ovšem nevstupuje do modelování. Cílová proměnná se odvozuje od délky delikvence, tj. od počtu dlužných splátek, běžná delikvence je například 30 a více dní prodlení od splatnosti splátky. Sledovaným obdobím je doba od skórování po nastání delikvence. Pokud je skórovací karta vytvářena dle požadavků Basel II definuje se délka delikvence jako 90 a více dní prodlení od splatnosti splátky během roční sledovací periody (García 2017). Doba sledování je úměrná době splatnosti úvěru (krátkodobý spotřební úvěr vs. hypotéka). Běžným rozmezím je 3 až 36 měsíců nebo do splacení úvěru (Siddiqi 2012).

3.4 Modelování

Cílem modelovací fáze je vytvoření skórovací karty, která vychází z modelování dichotomické proměnné. Pro vytvoření skórovací karty je třeba stanovit algoritmus, kterým se ze vstupních proměnných vypočítá skóre. Skóre znamená číslo, které co nejlépe předpovídá budoucí stav GOOD nebo BAD.

V bankovním prostředí je skórovací model tradičně založen na logistické regresi, která je velmi transparentní. Česká národní banka její využití očekává, nicméně Basel II její použití nenařizuje. Možnosti modelování skórovací karty jsou zmíněny v kapitole 4.

3.5 Hodnocení

Tato fáze hodnotí, zda model dosahuje obchodních cílů a snaží se zjistit, jestli neexistuje nějaký důvod, kvůli kterému by byl model nedostatečný. Vytvořený model se hodnotí na testovací množině, která nebyla použita ve fázi modelování. Sestavuje se matice záměn, kdy se porovnává dichotomická predikce karty se skutečností. Graficky se reprezentují odhady hustoty pravděpodobnosti pro GOOD a BAD skupinu. Měří se citlivost skóre, které by mělo optimálně sdružovat vstupní proměnné za účelem predikce selhání. Dále se využívá Kolmogor-Smirnovovův graf, který porovnává distribuční funkce a Kolmogorov-Smirnovova statistika, která vychází z maximálního rozdílu sledovaných distribucí (rozdělení skóre v kategorii dobrý a špatný). K hodnocení se také používá Giniho graf a statistika, která bude stejně jako ostatní hodnotící metody popsány v dalších kapitolách.

3.6 Nasazení

Fáze nasazení má za úkol vzít model a informace z fáze hodnocení a vyvodit z nich strategii pro implementaci. V oblasti poskytování úvěrů lze pro implementaci zvolit vlastní aplikaci nebo integraci do provozního softwaru. Vlastní aplikace je závislá na programátorech, v případě nové skóre karty znamená náročnou implementaci, je třeba doprogramovat logování uživatelů a zpravidla chybí nástroj pro správu verzí. Při začlenění do provozního softwaru je podpora logování, nezávislost na IT při běžném provozu, standardizace zavádění nové verze aj.

Samotné skórování může probíhat on-line nebo off-line. On-line skórování probíhá v případě aplikačního skórování a jedná se o okamžitý výpočet skóre pro jednu žádost. Off-line verze se používá v behaviorálním skórování a u aplikačního skórování jen pokud není potřeba rozhodnout hned. V off-line verzi se provádí výpočet pro všechny úvěrové expozice dávkou k určitému datu.

Aplikační karta nemůže nahradit schvalovací proces, je pouze jeho součástí, většinou se skórují žádosti až po vyhodnocení KO kritérií. KO kritérii se myslí kritéria, která v případě nesplnění znamenají okamžité zamítnutí žádosti o úvěr. Jedná se o některé špatné příznaky v úvěrových registrech, nedostatečný příjem, exekuce aj.

Pro vývoj a provoz skórovací karty je možné využít různý software (Siddiqi 2012; Thomas et al. 2017):

- Databázový software:
    - Nepřehledné programování (obtížná modifikace a aktualizace, obtížný audit);
    - Řešitel se soustředí více na programování než na vývoj karty;
    - Nesnadná dokumentace;
    - Velké riziko chyby.
- Statistický software:
    - Podpora přístupu k datům a datové manipulace;
    - Široké možnosti při budování modelu;
    - Postup vývoje lze zpravidla zaznamenat interním jazykem;
    - Nepodporuje provoz karty uvnitř statistického softwaru.
- Dataminigový software:
    - Podporuje všechny fáze CRISP-DM;
    - Řešitel se soustředí více na vývoj karty než na programování;
    - Vizuální programování usnadňuje dokumentaci;
    - Dávkové skórování je možné provádět přímo v dataminingovém software;
    - Umožňuje automatizovaný export celého řešení do produkčního systému.

Na vývoji karty se zpravidla podílí více profesí (datový specialista, dataminer, programátor, risk manager, auditor …) a většinou nestačí pouze jedna karta pro všechno, společnosti nabízí více produktů. Skórovací karty jsou vytvářeny za různým účelem (aplikační, behaviorální, podvod, marketingové kampaně aj.).

Důležité je zabezpečení skórovacích karet, aby nedošlo ke zneužití osobních údajů při předávání citlivých dat mezi kartou a koncovou aplikací. Algoritmus výpočtu a použití skóre by se měl chránit z důvodu úvěrových podvodů.

4 Modelování skórovací karty

Tato část je zaměřena na teoretický aspekt predikce pravděpodobnosti defaultu (PD) a expozice při defaultu (EAD). Přehled relevantních metod pro úvěrové skórování, které jsou použity v této práci je uveden v tabulce 4. Metody lze rozčlenit do několika skupin, konkrétně na individuální metody (klasifikátory / regresory) a kombinace modelů. Kombinace modelů mohou být založeny na kombinaci stejných individuálních metod (kombinace homogenních modelů) nebo různých individuálních metod (kombinace heterogenních modelů).

Tabulka 4: Přehled použitých metod pro modelování pravděpodobnosti defaultu a expozice při defaultu

Kategorie metod		Metoda
Individuální klasifikátory / regresory	Tradiční statistické metody	Diskriminační analýza Logistická regrese Bayesovská síť Lineární regrese
	Rozhodovací stromy	ForestPA (rozhodovací les penalizující proměnné) Credal rozhodovací strom C4.5 rozhodovací strom REPTree Alternující modelovací stromy (Alternating model tree) Náhodný strom (Random tree) M5P rozhodovací strom
	Podpůrné vektorové stroje	Podpůrné vektorové stroje / podpůrná vektorová regrese
	Neuronové sítě	Vícevrstvá neuronová síť typu perceptron
Homogenní kombinace modelů (homogeneous ensembles)		Bagging AdaBoostM1 MultiBoostAB LogitBoost Decorate Random SubSpace Rotační les (Rotation forest) Náhodný les (Random forest)
Heterogenní kombinace modelů (heterogeneous ensembles)		Stacking Voting

Zdroj: vlastní

Jako individuální metody strojového učení jsou v této práci použity čtyř skupiny metod: (1) tradiční statistické metody, (2) rozhodovací stromy, (3) podpůrné vektorové stroje a (4) neuronové sítě. Tyto metody pak byly použity jako základní algoritmy (klasifikátory / regresory)

v homogenních a heterogenních kombinacích modelů. V této kapitole jsou dále všechny metody popsány.

4.1 Tradiční statistické metody

4.1.1 Diskriminační analýza

Podstatou diskriminační analýzy je zjednodušeně zkoumání jedné kvalitativní proměnné v závislosti na několika kvantitativních proměnných. Podle počtu obměn kvalitativní proměnné se rozlišuje diskriminační analýza pro dvě nebo pro více tříd. Diskriminační analýza hledá pravidla, která umožní zařadit nový objekt do konkrétní třídy podle hodnoty zvolených proměnných. Možnosti aplikace diskriminační analýzy lze nalézt v různých oblastech života. Příkladem může být oblast medicíny a odlišení dvou tříd pacientů, vhodnost zaměstnat uchazeče v dané profesi podle psychologických testů nebo využití ve finančních institucích, které potřebují identifikovat své potenciální klienty (žadatele) z hlediska splácení úvěrů (Stankovičová a Vojtková 2007). Na základě různých proměnných lze vytvořit pravidla na zařazení klienta do třídy dobrých, rizikových (špatných) a případně neurčitých klientů (Stankovičová a Vojtková 2007).

Diskriminační funkce hledá takovou kombinaci proměnných, která maximalizuje rozdíl mezi dvěma populacemi a minimalizuje pravděpodobnost chybné klasifikace (Lane 1972). Model diskriminační analýzy lze definovat takto (Lee a Chiu 2002):

$$Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{2}$$

kde:

Z	diskriminační skóre,
$x_1, \ldots, x_k$	nezávislé proměnné,
β_0	konstanta,
$\beta_1, \ldots \beta_k$	koeficienty.

Pro klasifikaci finančních dat byla zavedena diskriminační analýza poprvé v roce 1941 D. Durandem (Durand 1941). Dalším autorem, který aplikoval diskriminační analýzu v oblasti finančních ukazatelů pro identifikaci finančního selhání byl W. Beaver (Beaver 1966). Vedle mnoha publikací, které se zabývají tématem predikce schopnosti splácet úvěr lze uvést například tyto: (Altman 1968; Sung et al. 1999; Mileris 2010).

4.1.2 Logistická regrese

Logistická regrese je metoda, která je vhodná na modelování jednostranné závislosti mezi proměnnými. Závislou proměnnou je v modelu kategoriální proměnná a vysvětlujícími proměnnými může být jak spojitá, tak kategoriální proměnná (Stankovičová a Vojtková, 2007).

Přístup výstavby modelu je založen na myšlence, že každá jednotlivá proměnná by měla být testována před tím, než je zařazena do modelu. Logistická regrese může být rozdělena do několika kategorií (Stankovičová a Vojtková, 2007):

- Binomická logistická regrese (Binomial logistic regression).
- Multinomická logistická regrese (Multinomial logistic regression).
- Ordinální logistická regrese – uspořádané kategorie (Ordinal logistic regression).

Obecný vzorec má následující tvar (Pacáková a Rublíková 2000):

$$Logit(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_{k,} \tag{3}$$

kde:

p	pravděpodobnost defaultu na základě několika charakteristik,
$x_1, \dots, x_k$	nezávislé proměnné,
β_0	konstanta,
$\beta_1, \dots, \beta_k$	koeficienty,
Logit	$ln\left(\frac{p}{1-p}\right)$

Poprvé použil regresní analýzu pro předpověď, zda zákazník bude v defaultu nebo ne Y. Orgler (Orgler 1970). Mezi mnoho dalších publikací, které se zabývají využitím logistické regrese v oblasti úvěrového skórování, lze uvést například tyto: (Steenackers a Goovaerts 1989; Berkowitz a Hynes 1999; Pacáková et al. 2005).

4.1.3 Lineární regrese

Metoda, která slouží k proložení číselných hodnot přímkou, se nazývá lineární regrese. Lineární regrese přijímá několik předpokladů. Jsou jimi předpoklad normálního rozdělení, homoskedasticity, lineárnosti a nezávislosti (Terek et al. 2010).

Obecný vzorec je tento (Terek et al. 2010):

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_{k,} \tag{4}$$

kde:

Y_i	závislá číselná proměnná,
$x_1, \ldots, x_k$	nezávislé proměnné,
β_0	konstanta,
$\beta_1, \ldots, \beta_k$	koeficienty.

Lineární regrese byla použita pro modelování úvěrového rizika v několika studiích. Jsou jimi např.: (Caselli et al. 2008; Witten a Frank 2005; Loterman et al. 2012).

4.1.4 Bayesovská síť

Bayesovská síť (Bayesian network) je model, který se využívá pro určení pravděpodobnosti. Je to acyklicky orientovaný graf, který zachycuje pravděpodobností závislosti mezi náhodnými veličinami pomocí hran (Witten a Frank 2005).

Bayesovská síť využívá obecný vzorec Bayesovy podmíněné pravděpodobnosti (Bayesův teorém). Ten je definován takto (Witten a Frank 2005):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{5}$$

kde:

$P(A)$	je apriorní pravděpodobnost jevu A,
$P(A\|B)$	je aposteriorní (podmíněná) pravděpodobnost jevu A za předpokladu, že nastal jev B,
$P(B\|A)$	je aposteriorní (podmíněná) pravděpodobnost jevu B za předpokladu, že nastal jev A,
$P(B)$	je pravděpodobnost jevu B.

K naučení podmíněných rozdělení pravděpodobnosti a struktury Bayesovké sítě se používají různé optimalizační algoritmy, včetně genetického nebo horolezeckého algoritmu. Bayesovská síť v úvěrovém skórování byla použita ve studiích autorů Witten a kol. (Witten a Frank 2005) nebo Louzada a kol.(Louzada et al. 2016).

4.2 Rozhodovací stromy a rozhodovací lesy

Rozhodovací stromy jsou tvořeny soustavou na sebe navazujících pravidel, jejichž cílem je rozdělení objektů na menší podmnožiny takovým způsobem, aby v jednotlivých podmnožinách převládaly objekty jedné třídy.

Mezi algoritmy pro tvorbu rozhodovacích stromů patří např. algoritmus C4.5, Credal rozhodovací strom, M5P, náhodný strom či alternující modelovací stromy. Algoritmus C4.5 uvedl Quinlan v roce 1993 (Quinlan 1993). Tento algoritmus staví rozhodovací strom z trénovací množiny výběrem atributů tak, aby se po rozdělení množiny získala co nejlepší informace.

Algoritmus C4.5 (Quinlan 1993)

Vstup: Trénovací množina *D*

1) Jestliže *D* je prázdná množina nebo algoritmus splňuje kritéria ukončení, potom algoritmus končí;
2) Výpočet informačního zisku pro všechny atributy;
3) Výběr nejlepšího atributu na základě informačního zisku;
4) Vytvoření rozhodovacího uzlu založeného na nejlepších atributu z kroku 3;
5) Rozdělení datové sady na základě nově vytvořeného rozhodovacího uzlu z kroku 4;
6) Pro všechny podmnožiny vzniklé v kroku 5 se opakují kroky výše;
7) Připojení uzlů získaných v kroku 6 k uzlu z kroku 4;

Výstup: Výsledný rozhodovací strom.

Metoda tvorby Credal rozhodovacího stromu (Credal decision tree) (Mantas a Abellán 2014) je podobná Quinlanovu algoritmu C4.5 (Quinlan 1993). Hlavním rozdílem je to, že Credal rozhodovacího strom předpokládá, že trénovací data nejsou zcela spolehlivá, odhaduje proto jejich pravděpodobnosti a na základě nich dělí proměnné v rozhodovacích uzlech. V oblasti úvěrového skórování byl algoritmus C4.5 použit např. ve (Witten a Frank 2005) a algoritmus Credal rozhodovacího strom ve studii (Abellán a Castellano 2017).

Algoritmus tvorby stromu M5P, představený Quinlanem v roce 1992 (Quinlan 1992), kombinuje rozhodovací strom s možností lineárních regresních funkcí v uzlech. Tento typ rozhodovacího stromu byl použit jako reprezentant strojového učení pro porovnání s ručně vytvářenými modely úvěrového skórování (Ben-David a Frank 2009).

Alternující modelovací stromy, které byly představeny autory Frank, Mayo a Kramer v roce 2015 (Frank et al. 2015), se skládají ze dvou typů uzlů, a to rozdělovacích a predikčních. Rozdělovací uzel je stejný jako uzel v klasickém rozhodovacím stromu a rozděluje data na základě zvolené proměnné. Predikční uzly obsahují číselné skóre, které se použije k předpovědi číselné závislé proměnné.

REP strom je rychlý algoritmus učení, který je založený na principu výpočtu informačního zisku (pro klasifikační úlohy) a minimalizaci chyby, která vyplývá z rozptylu (pro regresní úlohy). Používá logiku regresního stromu a generuje více stromů v různých iteracích. Poté vybírá nejlepší ze všech vytvořených stromů (Witten a Frank 2005).

Rozhodovací lesy vznikají vhodným kombinováním klasifikačních či regresních stromů. Získat různé podmnožiny z jedné trénovací množiny lze několika způsoby – viz algoritmy bagging či boosting níže.

ForestPA (rozhodovací lesy penalizující proměnné) byl uveden autory Adnan a Islam (Adnan a Islam 2017). Tito autoři navrhli, aby se při tvorbě rozhodovacího lesu váhy přiřadily pouze k proměnným, které se objevují v posledním stromu. Jednotlivé váhy se zvýší, pokud se proměnné nezobrazí v následujícím stromu či stromech. Algoritmus tvorby rozhodovacího lesu s penalizací proměnných je možné najít v (Adnan a Islam 2017).

4.3 Podpůrné vektorové stroje

Podpůrné vektorové stroje (Support vector machines) představili autoři Vapnik a Chervonenkis (Vapnik a Chervonenkis 1974). Podpůrné vektorové stroje tvoří kategorii jádrových algoritmů (kernel machines). Tyto algoritmy využívají jak algoritmy pro nalezení lineární oddělovací hranice, tak jsou schopné reprezentovat složité nelineární funkce. Předpokladem je převedení původního vstupního prostoru do vícedimenzionálního prostoru, ve kterém lze jednotlivé třídy lineárně oddělit. K této transformaci slouží jádrové funkce. Přeučení modelu brání parametr komplexnosti modelu C, jenž dovoluje dosažení určité úrovně chyby na trénovacích datech. V účelové funkci jsou pak zastoupeny jak chyba na trénovacích datech, tak šířka oddělovací hranice (ta je maximalizovaná). Tím dochází k optimalizaci strukturálního rizika.

Podpůrné vektorové stroje mohou být použity i jako regresní metody (Support vector regression). Princip je stejný s tím rozdílem, že výstupem u podpůrné vektorové regrese je

konkrétní číslo, což je pro predikci obtížnější. Hlavní myšlenkou je minimalizovat chybu a maximalizovat okraj modelu při tolerování určité chyby.

Publikace, ve kterých bylo využito podpůrných vektorových strojů v oblasti úvěrového skórování jsou například tyto: (Yu et al. 2010; Zhou et al. 2010; Hens a Tiwari 2012; Harris 2015; Maldonado et al. 2017).

4.4 Vícevrstvá neuronová síť typu perceptron

Neuronové sítě poskytují alternativu ke klasickým statistickým metodám, a to zejména tehdy, kdy závislé a nezávislé proměnné vykazují složité nelineární vztahy (Lee a Chiu 2002). Pro aplikaci neuronových sítí existuje několik možností, ať už jde o regresi, klasifikaci či časové řady.

Pro řešení klasifikačního problému, zda žadatel o úvěr selže se splácením nebo ne, lze použít několik typů modelů neuronových sítí. Mnozí autoři vyvinuli skórovací kartu pomocí dopředné vícevrstvé neuronové sítě typu perceptron (Multi-layer perceptron) (Bishop 1995; Desai et al. 1996; Dimla a Lister 2000; Erbas a Stefanou 2009; Reed a Marks 1998; Trippi a Turban 1992; West 2000).

Neuronová síť typu perceptron navržená Frankem Rosenblattem v roce 1957 je nejjednodušší jednovrstvá neuronová síť s dopředným šířením a učením s učitelem, tzn., že kromě znalosti vektoru vstupů je nutné znát i odpovídající výstupní vektor. Neuronová síť porovnává aktuální výstup s požadovaným výstupem, tj. učitelem a snaží se změnit váhy (synapse) tak, by se snížil rozdíl mezi skutečným a požadovaným výstupem.

Vícevrstvá neuronová síť typu perceptron je tvořena jednotlivými neurony (perceptrony) a skládá se ze vstupní, skryté (nebo více skrytých) a výstupní vrstvy. Počty neuronů v jednotlivých vrstvách jsou parametry struktury neuronové sítě. Počet vstupních neuronů závisí na počtu vstupů modelu. Počet neuronů skryté vrstvy se volí podle složitosti úlohy. Počet neuronů ve výstupní vrstvě závisí na počtu tříd u klasifikačních úloh, zatímco u regresních úloh je použitý jeden neuron. K naučení této neuronové sítě se většinou používají gradientní algoritmy, které zaručují konvergenci k lokálnímu minimu chybové funkce. Problémem může být dosažení globálního minima, respektive uvíznutí v lokálním minimu. K omezení tohoto rizika se používají různé techniky, např. zavedením momentu do učícího algoritmu.

Mezi publikace, které se zabývaly použitím vícevrstvé neuronové sítě typu perceptron v oblasti úvěrového skórování v nedávné době jsou například tyto: (Yu et al. 2016; Luo et al. 2017).

4.5 Homogenní kombinace modelů

4.5.1 Bagging

Metoda bagging je zkratkou slov „**b**ootstrap **agg**regat**ing**“. Bagging používá několik trénovacích množin, kdy je každá množina tvořena výběrem *n* objektů ze základní trénovací množiny. Takto vytvořené podmnožiny jsou použity k naučení jednotlivých klasifikátorů. Výsledná klasifikace se provádí pomocí metody průměrování (averaging) nebo hlasování (voting). Objekt je zařazen do třídy, kterou určila většina klasifikátorů (Breiman 1996).

Algoritmus Bagging (Breiman 1996)

Vstup: trénovací množina D, počet bootstrap vzorků T

Inicializace: klasifikátor C

Pro $t = 1, \ldots, T$:

- Vytvoření bootstrap vzorku D_t velikosti n z D
- Učení základního klasifikátoru C_t na bootstrap vzorku D_t

$C^*(x) = argmax_{y \in Y} \sum_{t:C_t(x)=y} 1$

Výstup: klasifikátor C^*

Mnoho autorů využilo metodu bagging ve svých článcích o úvěrovém skórování například tito: (Wang a Yao 2009; Zhang et al. 2010; Finlay 2011; Marqués et al. 2012b; Wang et al. 2011, 2012; Galar et al. 2013; Abellán a Mantas 2014; Wang et al. 2015; Yu et al. 2016; Sun et al. 2018).

4.5.2 Boosting

Metoda boosting stejně jako metoda bagging používá několik trénovacích množin, kdy se do každé této množiny vybírá *n* objektů. Metoda boosting kombinuje několik špatných klasifikátorů do jednoho silného. Nejprve se vytvoří první klasifikátor, který má přesnost klasifikace více než 0,5 a poté jsou přidávány další klasifikátory (Freund a Schapire 1997).

Algoritmus AdaBoost, jehož název je zkratkou slov „**ada**ptive **boost**ing“ byl představen autory Freund a Schapire v roce 1997 (Freund a Schapire 1997). Tento algoritmus vylepšil dřívější algoritmus boosting. AdaBoost minimalizuje chybu na trénovacích datech a tuto chybu je schopen snižovat exponenciálně (limita se blíží k nule), jestliže se mu daří hledat stále další slabé klasifikátory (tj. s chybou menší než 0,5). Výsledný klasifikátor je lineární kombinací slabých klasifikátorů (Freund a Schapire 1997).

Algoritmus AdaBoostM1 (Freund a Schapire 1999)

Vstup: $(x_1, y_1) \dots (x_m, y_m), x_i \in X, y_i \in Y, Y = \{-1, +1\}$

Inicializace: $C_t(i) = 1/m$

Pro $t = 1, \dots, T$:

- Trénování slabého klasifikátoru použitím distribuce C_t

  Získání slabé hypotézy $h_t: X \to \{-1, +1\}$ s chybou ε_t pro distribuci C_t
- Výpočet váhy

$$\alpha_t = 1/2 ln\big((1 - \varepsilon_t)/\varepsilon_t\big)$$

- Aktualizace modelu

$$C_{t+1}(i) = C_t(i)/Z_t \times \begin{cases} e^{-\alpha_t} & jestliže\ h_t(x_i) = y_i \\ e^{\alpha_t} & jestliže\ h_t(x_i) \neq y_i \end{cases}$$

$$= \Big(C_t(i)\ exp\big(-\alpha_t y_i h_t(x_i)\big)\Big)/Z_t$$

  kde Z_t je normalizační faktor vybraný tak, aby C_{t+1} zůstalo distribucí

Výsledný klasifikátor:

$$C^*(x) = sign \sum_{t=1}^{T} \alpha_t h_t(x)$$

Mezi publikace, ve kterých byl využit algoritmus AdaBoost v oblasti úvěrového skórování lze řadit například (Paleologo et al. 2010; Finlay 2011; Marqués et al. 2012b; Koutanaei et al. 2015).

MultiBoost je rozšířením algoritmu AdaBoost a v úvěrovém skórování byl použit v těchto publikacích: (Zhu et al. 2017; Ekinci a Erdal 2017). MultiBoost kombinuje algoritmus AdaBoost s algoritmem Wagging, což je algoritmus bagging s vahami. Wagging tedy modifikuje vliv každého vzorku pomocí vah. V algoritmu MultiBoost váží AdaBoost jednotlivé prvky souboru klasifikátorů.

Algoritmus MultiBoost AB (Webb 2011)

Vstup:

S_0 vstupní soubor $(x_1, y_1) \dots (x_m, y_m), x_i \in X, y_i \in Y$

$BaseLearn$ – základní učící algoritmus

T – maximální počet iterací k sestavení souboru

Vektor I_i, který specifikuje iteraci t k ukončení sestavení souboru

$S_1 = S_0$

$k = 1$

Pro $t = 1, \dots, T$

jestliže $I_k = t$

převážení S_t

$k = k + 1$

$C_t = \text{BaseLearn}(S')$

$\epsilon_t = \frac{\sum_{x_j \epsilon S_t : C_t(x_j) \neq y_j} váhy(x_j)}{m}$

jestliže $\epsilon_t > 0{,}5$

převážení S_t

$k = k + 1$

jinak jestliže $\epsilon_t = 0$

$\beta_t = 10^{-10}$

převážení S_t

$k = k + 1$

jinak

$\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$

$S_{t+1} = S_t$

pro každé $x_j \epsilon S_{t+1}$

jestliže $C_t(x_j) \neq y_j$ váhy (x_j) vyděl $2\epsilon_t$

jinak vyděl $2(1 - \epsilon_t)$

Výstupní klasifikátor:

$$C^*(x) = \text{a}rgmax_{y \epsilon Y} \sum_{t : C_t(x) = y} log \frac{1}{\beta_t}$$

Další variantou algoritmu Boosting je LogitBoost, který byl pro úvěrové skórování využit v pracích například těchto autorů: (Ben-David a Frank 2009; Finlay 2011).

Algoritmus LogitBoost (Friedman et al. 2000; Webb 2000)

Vstup: váhy $w_{ij} = 1/N, i = 1, \dots, N, j = 1, \dots, J, F_j(x) = 0$ a $p_j(x) = 1/J \; \forall \, j$.

Pro $m = 1, \dots, M$

- Pro $j = 1, \dots, J$
  - Výpočet vah v $j-t$é třídě

$$z_{ij} = \frac{y_{ij}^* - p_j(x_i)}{p_j(x_i)\left(1 - p_j(x_i)\right)}$$

$$w_{ij} = p_j(x_i)\left(1 - p_j(x_i)\right)$$

  - Přizpůsobení funkce $f_{mj}(x)$ pomocí metody nejmenších čtverců s vahami w_{ij}
- Nastavení $f_{mj}(x) \leftarrow \frac{J-1}{J}\left(f_{mj}(x) - \frac{1}{J}\sum_{k=1}^{J} f_{mk}(x)\right)$

$$F_j(x) \leftarrow F_j(x) + f_{mj}(x)$$

Výsledný klasifikátor:

$$\text{argmax}_j F_j(x)$$

4.5.3 Decorate

Algoritmus Decorate (Diverse Ensemble Creation by Oppositional Relabelling of Artificial Training Examples) byl představen autory Melville a Mooney v roce 2003 (Melville a Mooney 2003). Tento algoritmus iterativně generuje soubor klasifikátorů tak, že u každé iterace učí nový klasifikátor. V první iteraci je základní klasifikátor vybudován z daného souboru trénovacích dat a každý další klasifikátor je vytvořen z uměle vytvořené sady trénovacích dat, která je výsledkem spojení originálních trénovacích dat s umělými trénovacími příklady. Klasifikátor postavený z nové sady trénovacích dat se přidá do souboru pouze tehdy, pokud snižuje chybu při trénování souboru, jinak se odmítá. Algoritmus se snaží maximalizovat rozmanitost základních klasifikátorů, tj., aby se se co nejvíce lišily od současného souboru. Tento proces se opakuje, dokud se nedosáhne požadované velikosti nebo se překročí maximální počet iterací (Melville a Mooney 2003).

Algoritmus Decorate (Melville a Mooney 2003)

Vstup:

$BaseLearn$ – základní učící algoritmus

T – sada m trénovacích příkladů $\langle(\mathrm{x}_1, \mathrm{y}_1), \ldots, (x_m, y_m)\rangle$, kdy $y_i \in Y$

C_{size} – požadovaná velikost souboru klasifikátorů

I_{max} – maximální počet iterací k sestavení souboru klasifikátorů

R_{size} – faktor určující počet umělých příkladů, které se mají vygenerovat

$i = 1$

$trials = 1$

$C_i = BaseLearn(T)$

Inicializace souboru $C^* = \{C_i\}$

Výpočet chyby souboru:

$$\varepsilon = \frac{\sum_{x_j \in T: C^*(x_j) \neq y_j} 1}{m}$$

Dokud $i < C_{\text{size}}$ a $trials < I_{\max}$

Generuj $R_{\text{size}} \times |T|$ trénovací příklady, R je založeno na distribuci trénovacích dat

Příklady v R jsou nepřímo úměrné predikci C^*

$T = T \cup R$

$C' = BaseLearn(T)$

$C^* = C^* \cup \{C'\}$

$T = T - R$, odstranění umělých dat

Výpočet trénovací chyby ε' z C^* (viz výpočet chyby souboru výše)

Jestliže $\varepsilon' \leq \varepsilon$

- $i = i + 1$
- $\varepsilon = \varepsilon'$

jinak

- $C^* = C^* - \{C'\}$

$trials = trials + 1$

Ve spojení s úvěrovým skórováním byl algoritmus Decorate využit v několika studiích, jsou jimi například (Abellán a Castellano 2017; Marqués et al. 2012a).

4.5.4 Random SubSpace

Algoritmus Random SubSpace byl navržen tak, aby vyřešil problém kompromisu mezi přeučením a dosažením nejvyšší přesnosti. Tento algoritmus je podobný algoritmu bagging s výjimkou toho, že prvky (proměnné, prediktory) se náhodně odebírají a nahrazují pro každý základní klasifikátor, což způsobí, že se individuální klasifikátory nesoustředí na prvky, které se zdají být v trénovací množině vysoce prediktivní, ale nejsou prediktivní mimo tuto množinu. Algoritmus Random SubSpace je uveden např. v (Barushka a Hajek 2018). V oblasti úvěrového skórování je tato metoda využita v článku autorů Wang a kol. (Wang et al. 2012).

4.5.5 Rotační les

Rotační les je metoda pro generování klasifikačních souborů založených na extrakci prvků trénovací množiny. Trénovací množina je náhodně rozdělena do K podmnožin a na každou podmnožinu zvlášť je spuštěna analýza hlavních komponent. Nová trénovací množina je zkonstruována spojením všech hlavních komponent. Data jsou lineárně transformována do nového prostoru funkcí při zachování všech informací. Klasifikátor se učí na této množině. Různé dělení množiny vede k různým extrahovaným funkcím, které přispívají k větší rozmanitosti souboru (Kuncheva a Rodríguez 2007; Rodriguez et al. 2006).

Metodou rotačních lesů v úvěrovém skórování se zabývali např. autoři Abedini a kol. (Abedini et al. 2016).

4.5.6 Náhodný les

Náhodný les je algoritmus, který byl publikován Breimanem v roce 1996 (Breiman 1996). Tento algoritmus kombinuje algoritmus bagging s bootstrap výběrem, kdy některá pozorování jsou vybrána opakovaně a některá vůbec. V úvěrovém skórování byl tento algoritmus použit např. ve studii autorů Lessmann a kol. (Lessmann et al. 2015).

Algoritmus Náhodný les (Breiman 1996)

Vstup: Trénovací množina *D*, počet bootstrap vzorků T

Pro $t = 1, \ldots, T$

- Z trénovací množiny D se vytvoří náhodným výběrem s vracením bootstrap vzorek s n trénovacími daty a náhodně vybranými k proměnnými
- Každý bootstrap vzorek se použije k sestrojení jednoho stromu

Výstup: Výsledný les, který je dán většinovým hlasováním se stejnými vahami

4.6 Heterogenní kombinace modelů

Metody Voting a Stacking jsou heterogenní metody pro kombinace modelů, kdy obě metody kombinují individuální klasifikátory nebo homogenní metody kombinace modelů.

Metoda Voting je používána ke kombinování základních prediktorů pomocí většinového, váženého nebo pravděpodobnostního hlasování. Metoda Voting v úvěrovém skórování byla použita například v publikaci (Ala'raj a Abbod 2016b).

Přesnost heterogenních kombinací modelů může být zlepšena pomocí komplikovanější metody Stacking (Stacked Generalization).

Algoritmus Stacking (Wolpert 1992)

Vstup: trénovací soubor $D = \{x_i, y_i\}, i = 1, \ldots, N$, základní klasifikátor $C_j, j = 1, \ldots, h,$ kde h je počet základních klasifikátorů

Pro $j = 1, \ldots, h$

Učení C_j na základě souboru D

Pro $i = 1, \ldots, N$

Konstrukce nových dat D_c z prediktorů $D_c = \{x'_i, y_i\}$, kde $x'_i = \{c_1(x_i), \ldots, c_h(x_i)\}$

Učení meta-klasifikátoru C na množině D

Výstup: Klasifikátor C

Metoda Stacking kombinuje předpovědi několika základních algoritmů. Nejprve jsou všechny algoritmy naučeny s využitím dostupných dat a potom je metoda Stacking kombinuje. Na základě předpovědí ostatních algoritmů jako dodatečných vstupů je provedena

konečná predikce. Ke kombinování může být použit libovolný algoritmus, často se používá logistická regrese (Wolpert 1992).

Metoda Stacking byla úspěšně použita při predikci pravděpodobnosti defaultu v publikacích autorů Lessmann a kol. či Wang a kol. (Lessmann et al. 2015; Wang et al. 2011). Dále byla tato metoda použita pro úvěrové skórování (Dzeroski a Zenko 2004; Ala'raj a Abbod 2016b).

5 Hodnocení modelů

Cílem této kapitoly je popsat různé nástroje, které se mohou použít pro hodnocení kvality modelu v oblasti řízení úvěrového rizika. Tyto nástroje (ukazatele výkonnosti) slouží k rozhodnutí, který skórovací model je nejlepší, a který tedy vybrat. V literatuře není jednoznačně uvedeno, že by jeden ukazatel byl lepší než jiný. Následující část práce se bude věnovat těmto ukazatelům pro hodnocení klasifikačních modelů (Siddiqi 2012):

- Giniho koeficient (Gini), ROC (Receiver Operating Characteristic) křivka, resp. AUC (Area Under Curve, plocha pod ROC křivkou);
- Kolmogorovova-Smirnovova (KS) statistika;
- Přesnost (Accuracy, Acc);

a regresních modelů (Yao et al. 2015, 2017):

- Odmocnina ze střední kvadratické chyby (Root Mean Square Error, RMSE);
- Průměrná absolutní chyba (Mean Absolute Error, MAE);
- Koeficient determinace (Correlation Coefficient Square, R^2).

5.1 Giniho koeficient a ROC křivka

Giniho koeficient vyjadřuje číselně predikční sílu modelu, tj. měří stupeň zlepšení separace na základě skórovací karty oproti náhodnému rozhodování. Konkrétně měří podíl části B, která reprezentuje účinnost rozhodování na základě skóre, k části $A + B$, které reprezentují náhodné rozhodování. Hodnota 0 značí, že model nedokáže rozlišit mezi dobrými a špatnými klienty. Naopak hodnota 1 říká, že model rozdělí dobré a špatné klienty perfektně. Giniho koeficient je dán tímto vzorcem (Siddiqi 2012):

$$\frac{B}{A+B} = 100 * \left\{ 1 - 2 * \sum_{i}^{n} \left[\left(1 - \sum_{j}^{i} \frac{b(x_j)}{T_b} \right) * \frac{g(x_j)}{T_g} \right] \right\}, \tag{6}$$

kde:

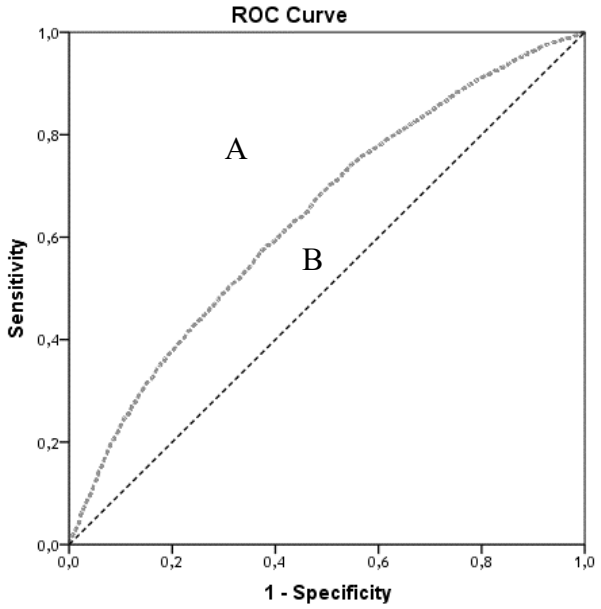
$b(x_j)$ počet špatných případů v rámci j-té skóre skupiny,

$g(x_j)$ počet dobrých případů v rámci j-té skóre skupiny,

T_b celkový počet špatných případů,

T_g celkový počet dobrých případů.

Giniho koeficient je roven dvojnásobku obsahu plochy mezi ROC křivkou a diagonálou. ROC křivka znázorňuje predikční sílu a umožňuje ji posoudit v závislosti na senzitivitě (sensitivity) a specificitě (specificity). Senzitivitou se myslí relativní četnost správně zařazených dobrých případů (True Positive Rate). Specificita znamená relativní úspěšnost při klasifikaci negativních případů (True Negative Rate). Na obrázku 4 je znázorněna ilustrativní ROC křivka. Kvalitu modelu pomocí ROC křivky lze vyjádřit i číselně, a to na základě plochy pod křivkou (AUC). Vztah mezi Giniho koeficientem (GI) a plochou pod ROC křivkou (AUC) je tento: $GI = 2 \times AUC - 1$.

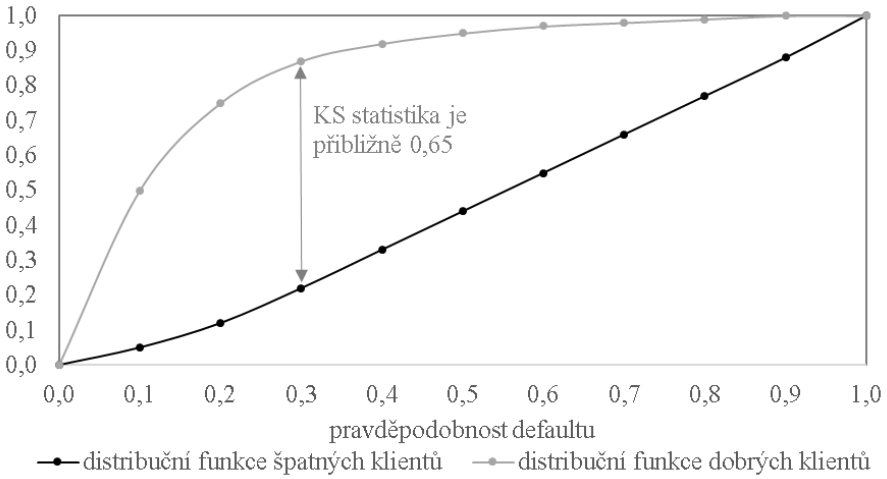


Obrázek 4: ROC křivka

Zdroj: vlastní

5.2 Kolmogorovova-Smirnovova statistika

Kolmogorovova-Smirnovova (KS) statistika měří maximální rozdíl mezi kumulativním rozdělením špatných (default) a kumulativním rozdělením dobrých (ne-default) klientů. KS statistika se pohybuje v rozmezí od 0 do 1. Hodnota 0 naznačuje, že model není schopen rozlišit mezi dobrými a špatnými klienty, zatímco hodnota 1 znamená, že model je schopen rozlišit dobré a špatné klienty dokonale. Na obrázku 5 je uveden příklad KS statistiky.



Obrázek 5: KS Statistika

Zdroj: vlastní

5.3 Přesnost klasifikace

Mezi standardní měřítka výkonnosti modelů patří i přesnost (Accuracy, Acc), která je na základě matice záměn (viz tabulka 5) vypočtena jako procento správně klasifikovaných úvěrů (Siddiqi 2012):

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}, \tag{7}$$

kde TP, TN, FP a FN jsou počty objektů klasifikované jako skutečně pozitivní (TP), skutečně negativní (TN), chybně pozitivní (FP) a chybně negativní (FN).

Tabulka 5: Matice záměn

Skutečnost	Predikce	
	Špatný úvěr	Dobrý úvěr
Špatný (defaultní) úvěr	TP	FN (chyba II. typu)
Dobrý (nedefaultní) úvěr	FP (chyba I. typu)	TN

Zdroj: vlastní

5.4 Hodnocení regresních modelů

Dalšími metrikami výkonnosti u regresních modelů jsou průměrná absolutní chyba (Mean Absolute Error, MAE), odmocnina ze střední kvadratické chyby (Root Mean Square Error, RMSE) nebo koeficient determinace R^2 které jsou dány těmito vzorci (Yao et al. 2015, 2017):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|, \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}, \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

kde y_i je predikovaná hodnota pro *i*-tý úvěr, $\tilde{y}_i$ je skutečná hodnota pro *i*-tý úvěr a $\bar{y}$ je průměrná hodnota dané proměnné.

6 Modelování skórovacích karet spotřebitelských úvěrů

Modelování úvěrového rizika spotřebitelských úvěrů je fundamentální úlohou bankovních či nebankovních finančních institucí, jež podporuje rozhodování o poskytnutí úvěru. Pro modelování celkového úvěrového rizika spotřebitelských úvěrů z hlediska očekávané ztráty je klíčové odhadnout parametry úvěrového rizika jako pravděpodobnost defaultu, ztrátu při defaultu a expozici při defaultu. Dosavadní výzkum měl tendenci modelovat tyto parametry zvlášť. Navíc opomíjenou oblastí na poli modelování ztráty a expozice při defaultu je aplikace kombinace modelů (ensemble learning), které díky výhodám odlišných základních metod snižuje problém přeučení a umožňuje modelování různorodých rizikových profilů defaultních úvěrů. K řešení těchto problémů je dále navržen dvoustupňový model úvěrového rizika, který

1) propojuje heterogenní kombinace klasifikačních modelů s nevyrovnanými třídami (tzv. class-imbalanced ensemble learning) pro predikci pravděpodobnosti defaultu;
2) a predikci expozice při defaultu použitím heterogenní kombinace regresních modelů (regression ensemble).

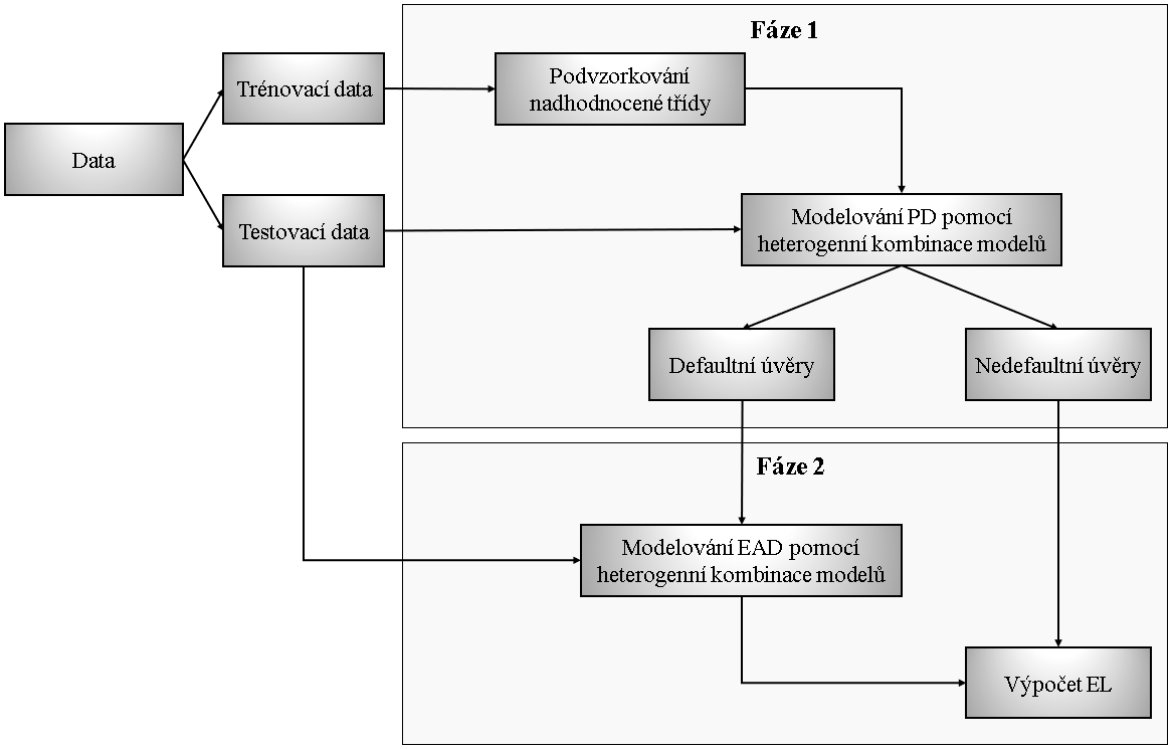
Dále je navržena metrika pro výpočet nákladů chybné klasifikace vhodná pro spotřebitelské úvěry s fixní expozicí, která kombinuje náklady obětované příležitosti a ztráty při defaultu.

Je ukázáno, že navrhovaný model úvěrového rizika je nejen efektivnější než jednofázové modely, ale také překonává nejmodernější metody používané při modelování úvěrového rizika z hlediska predikce a ekonomické výkonnosti.

6.1 Dvoufázový model úvěrového rizika

Jak bylo uvedeno v kapitole 1, modelování úvěrového rizika pomocí metod strojového učení se do nedávné doby omezovalo pouze na modelování pravděpodobnosti defaultu, což je pouze jeden nástroj používaný pro řízení úvěrového rizika. Pro celkový odhad tohoto rizika je potřeba znát také očekávanou ztrátu (expected loss, EL). Jak bylo uvedeno výše, očekávaná ztráta EL je dána vztahem $\mathrm{PD} \times \mathrm{EAD} \times \mathrm{LGD}$, kde PD je pravděpodobnost defaultu, EAD je expozice při defaultu a LGD je ztráta při defaultu.

Jednofázové statistické modely předpokládají, že všechny úvěry mají stejné pravděpodobnostní rozdělení, avšak hodnoty ztráty při defaultu se obvykle soustředí kolem 0 (plně splacený úvěr) nebo kolem 1 (nevymožení dlužné částky). Tento problém lze řešit tak, že se vytvoří klasifikátor k oddělení úvěrů na hranicích 0 nebo 1 a teprve potom se vytvoří regresní model pro odhad ztráty při defaultu mezi těmito hodnotami. Tímto je motivován dvoufázový model navržený v této práci, tzn., že dojde k oddělení defaultních úvěru s očekávanou ztrátou $EL > 0$ v první fázi a ve druhé fázi se predikuje expozici při defaultu jednotlivých úvěrů. Koncepční rámec je uveden na obrázku 6.



Obrázek 6: Koncepční rámec pro modelování úvěrového rizika

Zdroj: (Papoušková a Hájek 2019)

Nejdříve byla data rozdělena na trénovací a testovací množinu použitím pětinásobné křížové validace. Problém s vysoce nevybalancovanými daty byl vyřešen v první fázi pomocí algoritmu EUSBoost (Galar et al. 2013), což je evoluční algoritmus, který snižuje množství potřebných dat nadhodnocené třídy pomocí jejího podvzorkování (undersampling). V další části práce je ukázáno, že tento přístup je efektivnější než SMOTEBagging algoritmus (Wang a Yao 2009), který je založen na zvýšení množství dat podhodnocené třídy nadvzorkováním (oversampling).

Vzhledem k tomu, že heterogenní kombinace modelů (heterogeneous ensemble learning) fungovala v předchozích studiích na modelování pravděpodobnosti defaultu nejlépe

(Lessmann et al. 2015), je v této práci tento přístup aplikován v obou fázích modelování úvěrového rizika. Konkrétně je využita metoda Stacking, která kombinuje více prediktivních modelů pomocí algoritmů meta-učení. Přesněji řečeno, nejprve jsou vytvořeny základní modely z dat trénovací množiny a na základě jejich výstupů je vytvořen konečný model pomocí meta-učícího algoritmu. Tento model je heterogenní, protože základní modely jsou reprezentovány různými učícími algoritmy. Pro zlepšení výkonnosti modelu v obou fázích je využit namísto standardního lineárního algoritmu (logistická či lineární regrese) nelineární meta-klasifikátor a meta-regresor.

Nevybalancovaný počet objektů v jednotlivých třídách je považován za náročný problém, tzn., jedna třída v datech je reprezentována velmi odlišným počtem objektů. V úvěrovém skórování obvykle převažují nedefaultní úvěry nad úvěry defaultními, což může výrazně zhoršit výkonnost klasifikačních algoritmů, které jinak očekávají vyvážený soubor dat. Změna vzorkovací frekvence (resampling), ať už podvzorkování většinové třídy nebo nadvzorkování menšinové třídy bylo využito v několika studiích (Marqués et al. 2013; Brown a Mues 2012; Crone a Finlay 2012). V rozsáhlé studii (Brown a Mues 2012) o modelování pravděpodobnosti defaultu ve spojitosti s kombinací modelů byl efektivnější přístup podvzorkování. Proto je v této práci použit přístup založený na snížení množství potřebných dat ve většinové třídě. Konkrétně je použit algoritmus EUSBoost (Galar et al. 2013), který je založený na algoritmu RUSBoost, který propojuje náhodné podvzorkování a boosting. Hlavní nevýhodou náhodného podvzorkování je to, že může odmítnout potenciálně užitečné objekty většinové třídy. Tento problém je v EUSBoost vyřešen pomocí evolučního podvzorkování, které podporuje rozmanitost vybraných objektů většinové třídy.

Heterogenní klasifikační, resp. regresní kombinace modelů jsou sadou různých klasifikátorů, resp. regresorů, jejichž jednotlivé predikce jsou kombinovány za účelem zlepšení přesnosti a zvýšení rozmanitosti a komplementarity základních prediktorů. Pro kombinování základních prediktorů je obvykle používána metoda Voting. Přesnost heterogenních kombinací modelů může být výrazně vylepšena za pomoci složitější metody kombinování jako například algoritmus Stacking (Stacked Generalization) (Todorovski a Džeroski 2003). Tento algoritmus zahrnuje dva kroky, kdy nejprve je vygenerován soubor základních prediktorů a poté jsou předpovědi z prvního kroku použity k naučení meta-klasifikátoru či meta-regresoru.

Jako algoritmus meta-učení se u malého počtu základních prediktorů obvykle používá logistická regrese (Wang et al. 2011). U většího počtu základních prediktorů zvládají lépe meta-učení nelineární klasifikátory (Xia et al. 2018). Algoritmus meta-učení by měl být odolný proti multikolinearitě základních předpovědí (Lessmann et al. 2015). V další části textu jsou zkoumány dva meta-učící algoritmy: logistická regrese a náhodný les. Jako základní prediktory jsou využity nejmodernější algoritmy strojového učení, které jsou využívány při modelování úvěrového rizika. Modelování pravděpodobnosti defaultu je provedeno pomocí heterogenní kombinace klasifikačních modelů, zatímco k modelování expozice při defaultu je použita heterogenní kombinace regresních modelů.

Pro modelování pravděpodobnosti defaultu, resp. expozice při defaultu byly použity různé základní klasifikátory, resp. regresory jako rozhodovací či regresní stromy, homogenní kombinace modelů (např. bagging, boosting), neuronové sítě a podpůrné vektorové stroje.

Ve srovnávací studii autorů Abellán a Castellano se jako nejlepší základní prediktory pro modelování pravděpodobnosti defaultu jevila logistická regrese, vícevrstvá neuronová síť typu perceptron, podpůrné vektorové stroje či rozhodovací stromy C4.5 a Credal rozhodovací stromy (Abellán a Castellano 2017). V jiných studiích byly efektivně použity na modelování úvěrového skóre i další algoritmy homogenní kombinace modelů. Jde o algoritmus MultiBoostAB (Zhu et al. 2017), AdaBoostM1 (Koutanaei et al. 2015), LogitBoost (Finlay 2011), bagging (Wang et al. 2012), náhodný les (Lessmann et al. 2015), Decorate (Abellán a Castellano 2017), Random Subspace (Wang et al. 2012) a rotační les (Abellán a Castellano 2017). Rozhodovací strom s penalizovanými proměnnými (ForestPA) je algoritmus aplikovaný v oblasti úvěrového skórování teprve v nedávné době (Adnan a Islam 2017).

V související literatuře byly jako meta-učící algoritmy použity hluboké neuronové sítě (Deep neural networks) (Yu et al. 2016), podpůrné vektorové stroje (Lessmann et al. 2015), rotační les (Abedini et al. 2016) a boosting (Xia et al. 2018). V této práci je dále využit náhodný les, neboť je považován za v současnosti nejefektivnější algoritmus (benchmark) v oblasti úvěrového skórování (Lessmann et al. 2015). Pro účely srovnání byla však použita jako meta-klasifikátor také logistická regrese.

Při modelování expozice při defaultu byly zvoleny jako základní a meta-regresory ty metody, které byly již dříve použity při modelování ztráty a expozice při defaultu včetně lineární regrese, regresních stromů (M5P (Frank et al. 1998), REP Tree (Witten a Frank

2005), náhodný strom (Witten a Frank 2005) a alternující modelovací stromy (Witten a Frank 2005)), podpůrná vektorová regrese (Yao et al. 2015) a vícevrstvá neuronová síť typu perceptron (Yang a Tkachenko 2012). Navíc, vzhledem k pozorované účinnosti v předchozích studiích, které se zabývaly modelováním pravděpodobnosti defaultu, byly pro modelování expozice při defaultu použity homogenní kombinace modelů jako rotační les (Abellán a Castellano 2017), bagging (Wang et al. 2012), Random Subspace (Wang et al. 2012) a náhodný les (Lessmann et al. 2015). Jako meta-učící algoritmy při modelování expozice při defaultu byly použity regresní varianty algoritmů použité při modelování pravděpodobnosti defaultu, tj. náhodný les a lineární regrese.

Ve srovnávací studii z oblasti úvěrového skórování (Lessmann et al. 2015) se strategie, které využívaly selekci základních prediktorů jevily jako lepší z hlediska přesnosti než ty bez selekce. Nicméně na datové sadě, jež je představena níže, ke zvýšení přesnosti nedošlo, a proto se selekce základních prediktorů v dalším textu neuvažuje.

Nastavení metod, které byly použity pro modelování pravděpodobnosti defaultu a expozice při defaultu je uvedeno v tabulce 6. Pro realizaci experimentů bylo pro všechny metody použito programové prostředí Weka 3.8.

Tabulka 6: Nastavení metod použitých při modelování pravděpodobnosti defaultu a expozice při defaultu

Metoda	Parametry a jejich hodnoty
Forest PA	Počet stromů = 10
Credal rozhodovací strom	Min. celková váha případů na listu = 2, maximální hloubka stromů neomezena, min. poměr rozptylu v uzlu = 0,001
C4.5	Verze J48, min. počet případů na listu 2, faktor spolehlivosti pro prořezání = 0,25
REP Tree	Min. celková váha případů na listu = 2, maximální hloubka stromů neomezena, min. poměr rozptylu v uzlu = 0,001
Alternující modelovací strom	Počet iterací = 10, parametr smrštění = 1
M5P strom	Min. počet případů na listu = 4
Náhodný strom	Počet náhodně vybraných atributů = $\log_2$(počet prediktorů) + 1, min. poměr rozptylu v uzlu = 0,001
Log. regrese	Broyden–Fletcher–Goldfarb–Shanno učící algoritmus
Lin. regrese	Metoda nejmenších čtverců
Bayes Network	K2 algoritmus, počet rodičů = 2

SVM	Sekvenční minimalizační algoritmus s parametrem složitosti $C = 2^3$, polynomiální jádrová funkce s exponentem = 2
SVR	Epsilon-SVR, epsilon pro ztrátovou funkci = 0,1, $C = 2^3$, polynomiální jádrová funkce s exponentem = 2
Neuronová síť	Vícevrstvá perceptronová síť, velikost dávek pro mini-batch gradientní algoritmus = 100, počet skrytých vrstev = 2, neurony ve skrytých vrstvách = 2^4 a 2^3, rychlost učení = 0,3, míra redukce neuronů pro vstupní vrstvu = 0,2 a počet iterací = 500
Bagging	Základní prediktor = REP strom, počet iterací = 10, velikost vzorku je dána procentem z trénovací množiny = 50
Rotační les	Základní prediktor = C4.5 pro klasifikaci a REP strom pro regresi, počet iterací = 10, procento případů, které mají být odebrány = 50, projekční filtr = hlavní komponenty
MultiBoostAB	Základní prediktor = Decision Stump, počet iterací = 10 a počet modelů = 3
AdaBoostM1	Základní prediktor = Decision Stump, počet iterací = 10
LogitBoost	Základní prediktor = Decision Stump, počet iterací = 10, parametr smrštění = 1
Decorate	Základní prediktor = C4.5, počet členů klasifikátoru = 15, počet iterací = 50, počet umělých případů během učení = 1
Random SubSpace	Základní prediktor = C4.5 pro klasifikaci a REP strom pro regresi, velikost každého podprostoru = 0,5, počet iterací = 10
Voting	Pravidlo pro kombinování = průměr pravděpodobností nebo většinové hlasování
Náhodný les	Maximální hloubka stromů neomezena, počet stromů, které mají být generovány = 100, počet náhodně vybraných proměnných jako kandidátů na každém rozdělení = $\log_2$(počet prediktorů) + 1
Stacking	Meta-učící algoritmus = {LogR, náhodný les} pro klasifikaci a = {lin. regrese, náhodný les} pro regresi

Zdroj: vlastní

Aby se zabránilo přeučení meta-učícího algoritmu při použití metody Stacking, parametry učení základní prediktorů nebyly v prvním kroku optimalizovány. Zároveň bylo pro učení meta-algoritmu použito stejné rozdělení dat (trénovací a testovací data) jako v případě učení základních prediktorů.

Vhodnost výše uvedených klasifikačních a regresních kombinací modelů je posouzena pomocí standardních klasifikačních a regresních metrik, které byly v oblasti úvěrového skórování použity dříve (Yao et al. 2017; Florez-Lopez a Ramon-Jeronimo 2015; Abdou a Pointon 2011), viz kapitola 5. Dále je navrženo měření nákladů chybné klasifikace, aby

bylo možné zvážit finanční důsledky různých typů chyb při predikci pravděpodobnosti defaultu.

Standardní měření výkonnosti používané k hodnocení modelů pravděpodobnosti defaultu zahrnují (viz kapitola 5):

- přesnost (Acc),
- AUC, tj. plocha pod ROC křivkou,
- Giniho koeficient,
- Kolmogorovova-Smirnovova statistika.

Při modelování pravděpodobnosti defaultu dochází na jedné straně k chybné predikci úvěru, který není splacen (chyba II. typu) a vede ke ztrátě investice. Na druhé straně je předpověď špatného (nevymožitelného) úvěru, který by ve skutečnosti byl plně splacen (chyba I. typu), což může vést ke ztrátě potencionálního úroku (tzv. náklady obětované příležitosti). Několik studií, které se zabývaly úvěrovým skórováním zkombinovaly tyto dvě chyby do vzorce pro výpočet nákladů chybné klasifikace (misclassification cost, MC) (Louzada et al. 2016), které jsou považovány za klíčové kritérium pro hodnocení efektivnosti úvěrového skórování (Abdou a Pointon 2011). Toto kritérium je však zřídka využíváno jako kritérium hodnocení výkonnosti predikčních modelů (Bahnsen et al. 2015). Jinými slovy lze říci, že se předchozí literatura neadekvátně zabývala náklady spojenými s modelováním úvěrového rizika. Článek autorů Bahnsen a kol. (Bahnsen et al. 2015) byl inspirací pro navrhovanou metriku pro hodnocení modelování pravděpodobnosti defaultu, která kombinuje ztrátu při defaultu a náklady obětované příležitosti takto:

$$MC = FNR * LGD + FPR * \left(r + \left(-r * \pi_{good} + LGD * \pi_{bad}\right)\right), \tag{11}$$

$$LGD = 1 - \frac{vymožená\ částka\ - náklady\ na\ vymáhání}{nesplacená\ výše\ úvěru}, \tag{12}$$

kde:

FPR je podíl chybně pozitivních (defaultních) úvěrů,

FNR je podíl chybně negativních (nedefaultních) úvěrů,

π_{good} je procento dobrých úvěrů,

π_{bad} je procento špatných úvěrů,

r je průměrný zisk z úvěru (úrok),

LGD je ztráta při defaultu.

Hlavní rozdíl mezi navrhovaným výpočtem a náklady chybné klasifikace autorů Bahnsen a kol. (Bahnsen et al. 2015) je ten, že navrhované měření je vyjádřeno relativně, tzn., že není závislé na individuálních datech o úvěrech, což má dvě výhody. První výhodou je, že relativní MC je pro investory lépe interpretovatelné a druhou výhodou je, že výsledky různých modelů mohou být mezi sebou snadno porovnány. Další měřítko, které bylo použité v dřívějším souvisejícím výzkumu je očekávaný maximální zisk (Verbraken et al. 2014, 2013), jehož cílem je vybrat klasifikátor s nejvyšším ziskem. Tento alternativní přístup ovšem nezohledňuje náklady spojené s odmítnutím úvěru (náklady obětované příležitosti). Navíc je očekávaný maximální zisk vyjádřen v absolutních jednotkách, a proto nemůže být použit k porovnání výkonnosti modelů pro více datových souborů.

Pro vyhodnocení výsledků modelování expozice při defaultu a celkové očekávané ztráty byly použity tři různé metriky, jimiž jsou index determinace (R^2), střední absolutní chyba (MAE) a odmocnina ze střední kvadratické chyby (RMSE). Všechny metriky jsou definovány v kapitole 5. Tento výběr odpovídá předchozím studiím, které se zabývaly modelováním ztráty a expozice při defaultu (Yao et al. 2017; Nazemi et al. 2017).

6.2 Datový soubor

V datovém souboru jsou pro každý úvěr k dispozici socio-demografické údaje žadatele jako věk, pohlaví, rodinný stav, typ bydlení, kraj trvalého bydliště, stupeň vzdělání, typ klienta (zaměstnanec, důchodce (starobní či invalidní) nebo osoba na mateřské či rodičovské dovolené) a počet let v zaměstnání. Dále je v datovém souboru charakterizována finanční situace žadatelů, která je dána čistým měsíčním příjmem, měsíčními náklady a volnými zdroji. Vedle již zmíněných údajů soubor obsahuje vlastnosti každého úvěru, jako vyplacenou částku, částku, která má být splacena (tj. vyplacená částka navýšená o budoucí úroky), délka úvěru (splatnost), označení existence spoludlužníka a výši měsíční splátky. Zbývající atributy obsahují profil žadatele (nový žadatel, vracející se žadatel, souběžný úvěr či přeúvěrování již existujícího úvěru) a informace dostupné z úvěrových společností. Jednou úvěrovou společností byla společnost NRKI (Nebankovní registr klientských informací), z něhož byl vyvozen atribut Credit Bureau Score (CBS) a druhou úvěrovou společností byla společnost SOLUS, která poskytla informace o existenci jednotlivých příznaků žadatele.

Tabulka 7 popisuje jednotlivé proměnné, které byly pro další využití normalizovány do intervalu [0, 1]. Histogramy jednotlivých proměnných a jejich podrobnější popis je možné nalézt v příloze této práce. Datový soubor neobsahuje žádná chybějící pozorování. Selekce proměnných nebyla uvažována, neboť nevedla k významnému zpřesnění modelu (Papoušková a Hájek 2019). To je dáno tím, že vybraná finanční instituce nesbírá redundantní informace o žadatelích. Kromě toho některé metody (např. náhodný les) používají vnořenou selekci proměnných.

Datový soubor není veřejně dostupný, neboť data o spotřebitelských úvěrech byla poskytnuta nebankovní finanční institucí, která chce zůstat v anonymitě. Datový vzorek zahrnuje období od 1. 1. 2015 do 31. 3. 2016. Pro účely dalšího použití byly úvěry rozděleny na základě toho, jak jsou spláceny po dobu prvních 12 měsíců od poskytnutí. Interval 12 měsíců byl přitom počítán od prvního měsíce následujícího po měsíci poskytnutí.

Za příznak selhání byla považována kterákoli z následujících událostí:

(a) překročení 60 dnů po splatnosti (days past due),
(b) prohlášení insolvence,
(c) zesplatnění nebo žaloba.

Tabulka 7: Popis atributů

Atribut	Škála	Popis
CBS	$\{12, 13, 14, 16, 17, 166+\}$	Skóre z Nebankovního registru klientských informací (NRKI), CBS = Credit Bureau Score
Výše úvěru	$[11\,018, 484\,884]$	Vyplacená částka navýšená o budoucí úroky, tj. částka, kterou má klient splatit (Kč)
Splatnost	$[12, 48]$	Splatnost úvěru v měsících
Délka zaměstnání	$[0, 47]$	Délka zaměstnání v letech
K1 koeficient	$[-3{,}1, 75{,}5]$	Koeficient K1 = volné zdroje / měsíční splátka
LTD	$[-23{,}67, 5\,285{,}71]$	LTD = vyplacená částka / volné zdroje (Kč)
Region	$\{1, 2, \dots, 15\}$	Kraj trvalého bydliště
Měsíční splátka	$[456, 14\,562]$	Výše měsíční splátky (Kč)
Počet příznaků v SOLUS	$[0, 4]$	Počet příznaků v registru SOLUS
Pohlaví	$\{0, 1\}$	Pohlaví klienta
Čisté měsíční příjmy	$[0, 165\,252]$	Čisté měsíční příjmy klienta (Kč)
Ostatní příjmy	$[0, 106\,000]$	Ostatní měsíční příjmy (Kč)
Příjmy celkem	$[0, 165\,252]$	Celkové měsíční příjmy (Kč)
Profil klienta	$\{1, 2, 3, 4\}$	Profil klienta
Rodinný stav	$\{1, 2, \dots, 6\}$	Rodinný stav
Příznaky v SOLUS	$\{0,1\}$	Příznaky A, B, C, D, P, U, Z v registru SOLUS
Spoludlužník	$\{0,1\}$	Existence spoludlužníka
Typ klienta	$\{1, 2, 3\}$	Typ klienta
Typ bydlení	$\{1, 2, \dots, 7\}$	Typ bydlení
Věk klienta	$[18, 86]$	Věk v době žádosti
Volné zdroje	$[-7\,794, 155\,179]$	Volné zdroje (Kč)
Měsíční výdaje	$[6\,300, 41\,473]$	Měsíční výdaje klienta (Kč)
Splátky úvěrů/leasing	$[0, 25\,000]$	Měsíční splátky úvěrů a leasingů u ostatních společností (Kč)
Splátky úvěrů	$[0, 26\,697]$	Současné měsíční splátky úvěrů u nebankovní finanční společnosti (Kč)
Vyplacená částka	$[7\,000, 166\,000]$	Vyplacená částka (Kč)
Vzdělání	$\{1, 2, \dots 8\}$	Vzdělání
Výše splátek z registru SOLUS	$[0, 21\,022]$	Výše splátek z registru SOLUS (Kč)
Default	$\{0,1\}$	Označení, zda došlo k defaultu nebo ne
EAD	$[0,\ 372\,240]$	Expozice při defaultu (Kč)

Zdroj: vlastní

Každý úvěr byl označen jako:

- uspokojivě splácen (dále pouze „dobrý úvěr“), pokud splňoval všechny následující podmínky:
  - jednalo se o poskytnutý úvěr, tj. úvěr nebyl zamítnut,
  - úvěr mohl být pozorován alespoň 12 měsíců, tj. byl poskytnut do 31. 3. 2016,
  - v prvních 12 měsících po poskytnutí nenastala žádná z událostí (a), (b), (c),
- neuspokojivě splácen (dále pouze „špatný úvěr“), pokud splňoval všechny následující podmínky:
  - jednalo se o poskytnutý úvěr, tj. úvěr nebyl zamítnut,
  - úvěr mohl být pozorován alespoň 12 měsíců, tj. byl poskytnut do 31. 3. 2016,
  - v prvních 12 měsících po poskytnutí nastala alespoň jedna z výše uvedených událostí (a), (b), (c),
- nevyhodnocená z hlediska splácení (dále pouze „vyřazený úvěr“), pokud nesplnil podmínky pro dobrý nebo špatný úvěr.

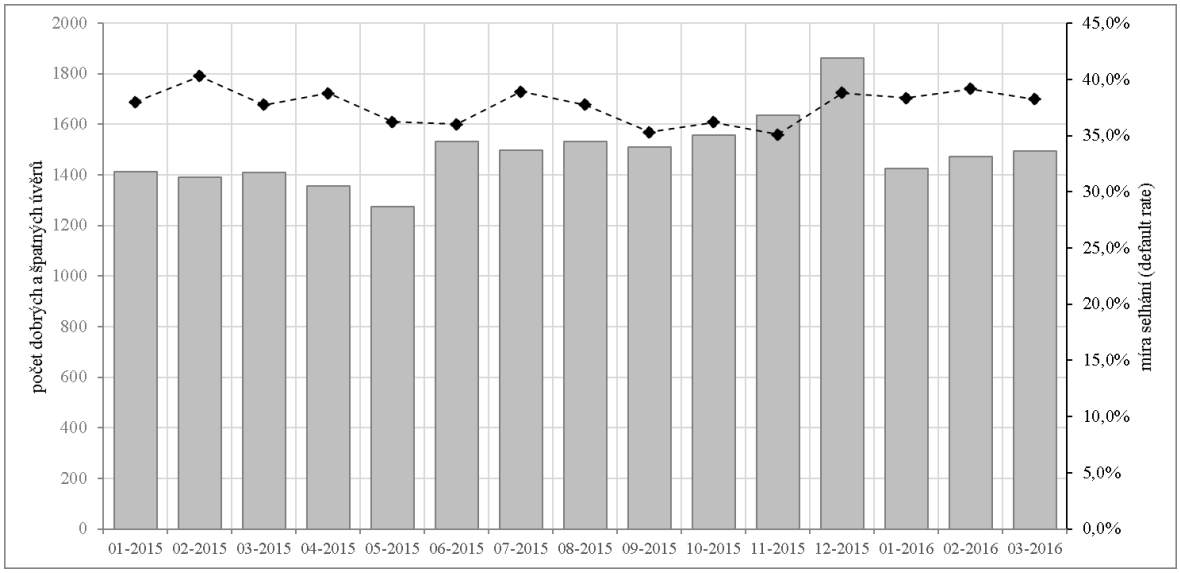
Celkový počet dobrých a špatných úvěrů činí 22 364, z toho špatných 8 610. Podíl nesplacených úvěrů činí 38,50 %, čímž dochází k nerovnováze 1,6 ku 1 ve prospěch dobrých (nedefaultních) úvěrů. Četnost výskytu událostí selhání (a), (b), (c) je uvedena v tabulce 8.

Tabulka 8: Četnost výskytu událostí

Událost selhání			Počet úvěrů	
a) dny po splatnosti > 60	b) insolvence	c) zesplatnění nebo žaloba		
ne	ne	ne	13 754	13 754
ne	ne	ano	5	
ne	ano	ne	76	
ne	ano	ano	0	
ano	ne	ne	311	8 610
ano	ne	ano	5 065	
ano	ano	ne	1 265	
ano	ano	ano	1 888	
Celkem úvěrů a z toho podíl špatných			22 364	38,50 %

Zdroj: vlastní

Časový vývoj počtu dobrých a špatných úvěrů a relativního zastoupení špatných úvěrů je v měsíční řadě znázorněn na obrázku 7.



Obrázek 7: Měsíční vývoj dobrých a špatných úvěrů

Zdroj: vlastní

Očekávaná ztráta (EL) byla vypočítána pomocí tří již zmíněných parametrů úvěrového rizika $PD \times EAD \times LGD$. Hodnota pravděpodobnosti defaultu PD byla odhadnuta pomocí kategorizace klientů do defaultní a nedefaultní třídy.

EAD (expozice při defaultu) představuje poměr aktuálně čerpané částky ku celkové částce úvěru. V průměru činila expozice při defaultu 87,61 %. LGD (ztráta při defaultu) pro individuální spotřebitelské úvěry nebyla k dispozici (z důvodu neukončeného vymáhacího procesu, který obvykle trvá několik měsíců či let). Pro účely odhadu očekávané ztráty (EL) bylo LGD nastaveno jako 74,88 %. Parametr LGD je stanoven pevně a je možné jej použít pro výpočet EL pro všechny úvěry. Odhadované LGD bylo použito k výpočtu nákladů chybné klasifikace. Empirické rozložení hodnot EL je znázorněno na obrázku 8. Je zřejmé, že EL nemá normální rozdělení pravděpodobností, přičemž EL = 0 znamená dobré úvěry (tj. úvěry bez selhání) a EL = 1 (EAD = 1) znamená nesplacený úvěr v plné výši.

Obrázek 8: Rozložení hodnot EL

Zdroj: vlastní

6.3 Výsledky modelování

V této podkapitole jsou uvedeny průměrné hodnoty a směrodatné odchylky z 5 pokusů prováděných na různých datových oddílech (pětinásobná křížová validace). Pro klasifikační úkol (fáze 1, modelování pravděpodobnosti defaultu) jsou ukázány metriky přesnost (Acc), plocha pod ROC křivkou (AUC), Giniho koeficient a náklady chybné klasifikace (MC). Pro regresní úlohy (fáze 2, modelování expozice při defaultu a očekávané ztráty) jsou reportovány metriky RMSE, MAE a R^2.

Mnoho klasifikátorů poskytuje skóre, které udává pravděpodobnost zařazení do dané třídy pro každý úvěr. Hranice, na základě které jsou jednotlivé úvěry zařazeny do tříd (dobrý vs. špatný úvěr), se nazývá cut-off. Pro zjištění optimální hranice (cut-off) byl využit přístup uvedený autory Verbraken a kol. (Verbraken et al. 2014), který zohledňuje náklady pro TNR (podíl správně klasifikovaných dobrých úvěrů) a FNR (podíl chybně klasifikovaných dobrých úvěrů):

$$\lambda = \frac{\left(r + \left(-r * \pi_{good} + LGD * \pi_{bad}\right)\right) * \pi_{good}}{LGD * \pi_{bad}}, \tag{13}$$

kde λ je podíl TNR a FNR.

Hodnoty nákladů chybné klasifikace MC byly vypočteny dle vzorce (11) následovně:

$$MC = FNR * 0{,}749 + FPR * \left(0{,}802 + \left(-0{,}802 * 0{,}615 + 0{,}749 * 0{,}385\right)\right), \tag{14}$$

kde $LGD = 0{,}749$ je vymožená část po defaultu a $r = 0{,}802$ je průměrná výše úrokové sazby (zisk z úvěru). Procento dobrých, resp. špatných úvěrů je $\pi_{good} = 0{,}615$, resp. $\pi_{bad} = 0{,}385$. Cut-off byl vypočten pomocí vzorce (13) $1/(1+\lambda) = 0{,}44$.

V první sadě experimentů byl testován vliv podvzorkování (metoda EUSBoost) a nadvzorkování (metoda SMOTE (Chawla et al. 2002)). V metodě EUSBoost (Galar et al. 2013) bylo provedeno podvzorkování (undersampling) v rámci algoritmu AdaBoost.M2. Při podvzorkování bylo nastaveno procento většinové třídy tak, aby se vyvážil počet případů menšinové třídy. Pro nadvzorkování byla použita metoda SMOTEBagging, která kombinuje nadvzorkování menšinové třídy s metodou bagging, kdy jsou generovány umělé případy.

Porovnání metody EUSBoost a SMOTEBagging bylo provedeno pomocí algoritmu náhodný les, kdy došlo k učení na vyváženém souboru dat. Bylo zjištěno, že metoda EUSBoost převyšovala metodu SMOTEBagging či klasifikaci náhodného lesu na původních datech z hlediska jednotlivých metrik. Algoritmus náhodný les naučený na původních datech fungoval dobře pouze na většinové třídě, zatímco metoda SMOTEBagging byla účinná pro menšinovou třídu, což naznačují hodnoty AUC a MC v tabulce 9. Za účelem statistického srovnání byl proveden Wilcoxonův test. Tyto výsledky naznačují, že podhodnocení většinové třídy efektivně řeší otázku nerovnováhy mezi třídami v datovém souboru. Tabulka 9 ukazuje, že EUSBoost statisticky významně překonal jak náhodný les na původních nevybalancovaných datech, tak nadvzorkování pomocí SMOTEBagging.

Tabulka 9: Efekt vyvážení dat za pomoci učení metodou náhodný les (RF)

Metoda	Acc	AUC	MC
RF	66,13 ± 0,68	68,81 ± 0,62	0,491 ± 0,010
EUSBoost + RF	**78,29 ± 0,97**	**88,16 ± 0,97**	**0,292 ± 0,012**
SMOTEBagging + RF	71,04 ± 0,30	78,44 ± 0,20	0,387 ± 0,004

** Statisticky významně podobné modely (p-hodnota < 0,05) vzhledem k nejlepšímu modelu označenému tučně.*

Zdroj: vlastní

Ve druhé sadě experimentů byla porovnána výkonnost klasifikačních metod použitých při modelování pravděpodobnosti defaultu. Tabulka 10 uvádí výsledky tří typů klasifikátorů. Prvně jde o individuální klasifikátory ForestPA (Adnan a Islam 2017), Credal rozhodovací strom (CDT) (Abellán a Castellano 2017), C4.5 rozhodovací strom (Witten a Frank 2005), logistická regrese (LogR) (Witten a Frank 2005), podpůrné vektorové stroje (SVM) (Lessmann et al. 2015) a neuronovou síť typu perceptron (NN) (Yu et al. 2016). Druhým

uváděným typem jsou homogenní kombinace modelů MultiBoostAB (Zhu et al. 2017), AdaBoostM1 (Koutanaei et al. 2015), LogitBoost (Finlay 2011), rotační les (RotForest) (Abedini et al. 2016), náhodný les (RF) (Lessmann et al. 2015), Decorate (Abellán a Castellano 2017), bagging (Wang et al. 2012) a Random SubSpace (Wang et al. 2012). Za třetí jsou to heterogenní kombinace modelů Voting (Ala'raj a Abbod 2016b) a Stacking (Dzeroski a Zenko 2004). Heterogenní kombinace modelů integrují předpovědi jak individuálních klasifikátorů, tak homogenních kombinací modelů. Nejlepší výsledky jsou označeny tučně. Dále jsou označené statisticky podobné výsledky (získané opět pomocí Wilcoxonova testu).

Tabulka 10: Výkonnost klasifikátorů modelů pravděpodobnosti defaultu

Metoda	Přesnost (Acc)	AUC	MC	Gini koeficient
ForestPA	74,942 ± 1,187	0,828 ± 0,011	0,338 ± 0,016	0,657 ± 0,022
CDT	65,402 ± 0,832	0,691 ± 0,013	0,466 ± 0,009	0,382 ± 0,026
C4.5	70,851 ± 0,970	0,729 ± 0,010	0,394 ± 0,014	0,457 ± 0,020
LogR	64,291 ± 1,410	0,699 ± 0,012	0,479 ± 0,019	0,397 ± 0,024
BayesNet	63,477 ± 0,928	0,685 ± 0,013	0,489 ± 0,012	0,369 ± 0,027
SVM	63,797 ± 1,838	0,638 ± 0,018	0,498 ± 0,036	0,276 ± 0,037
NN	66,209 ± 1,146	0,712 ± 0,008	0,448 ± 0,012	0,423 ± 0,016
MultiBoostAB	61,995 ± 1,233	0,660 ± 0,008	0,502 ± 0,010	0,319 ± 0,015
AdaBoostM1	63,637 ± 0,687	0,681 ± 0,009	0,485 ± 0,010	0,363 ± 0,018
LogitBoost	63,949 ± 0,789	0,686 ± 0,010	0,483 ± 0,011	0,372 ± 0,020
RotForest	77,543 ± 1,113	0,872 ± 0,006	0,302 ± 0,016	0,745 ± 0,012
RF	78,436 ± 1,184*	0,884 ± 0,009	0,290 ± 0,016*	0,767 ± 0,017
Decorate	75,857 ± 0,681	0,851 ± 0,007	0,326 ± 0,009	0,703 ± 0,015
Bagging	71,883 ± 1,496	0,786 ± 0,012	0,378 ± 0,019	0,573 ± 0,024
RandomSubSpace	73,714 ± 1,049	0,812 ± 0,009	0,354 ± 0,013	0,623 ± 0,019
Voting (avg)	73,946 ± 1,042	0,816 ± 0,011	0,351 ± 0,017	0,632 ± 0,022
Voting (maj)	73,707 ± 1,069	0,737 ± 0,011	0,354 ± 0,016	0,474 ± 0,021
Stacking (LogR)	78,189 ± 1,066*	0,886 ± 0,008*	0,292 ± 0,015*	0,771 ± 0,016*
Stacking (RF)	**78,967 ± 1,043**	**0,888 ± 0,010**	**0,278 ± 0,015**	**0,776 ± 0,020**

** Statisticky významně podobné modely (p-hodnota < 0,05) vzhledem k nejlepšímu modelu označenému tučně.*

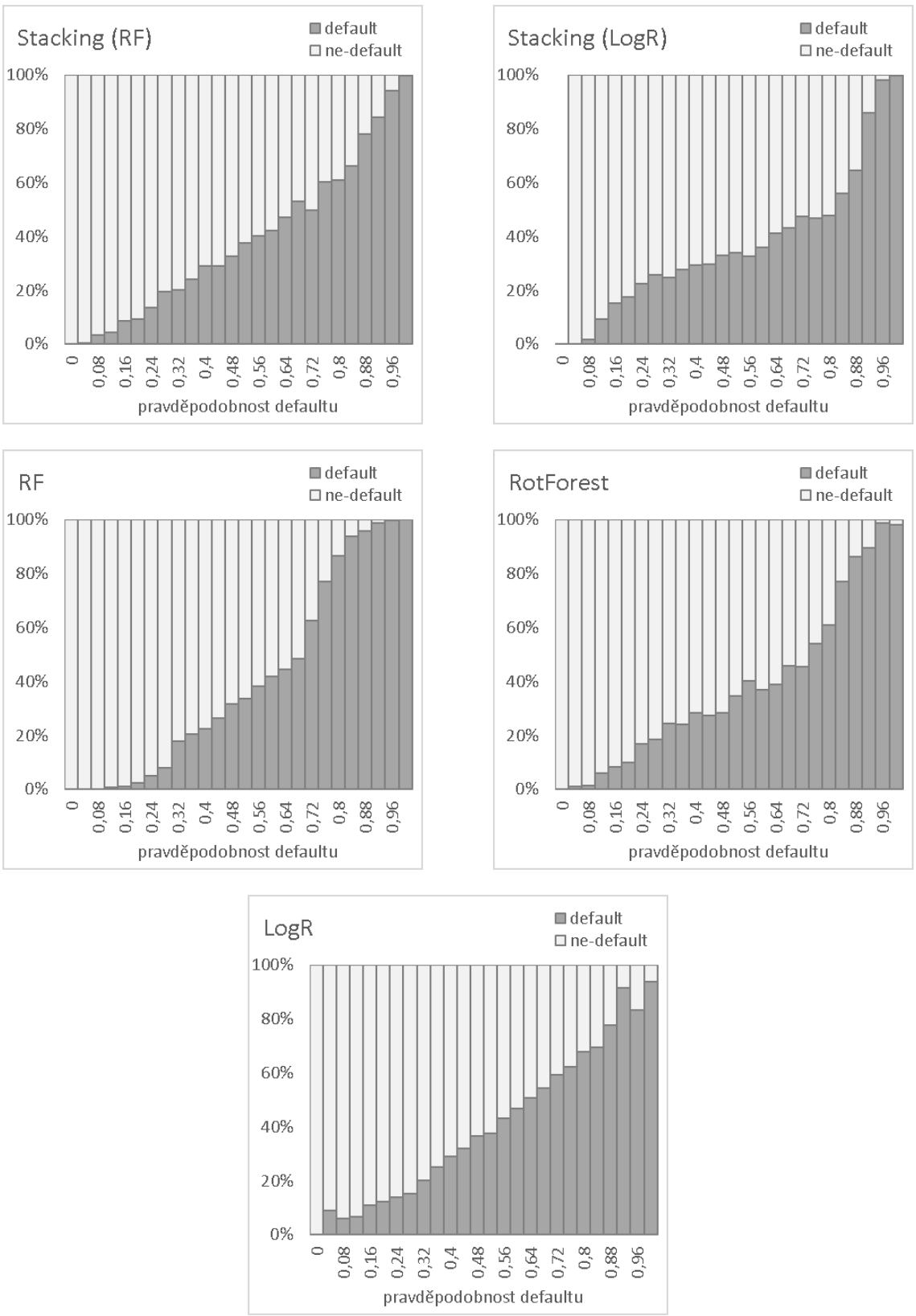
Zdroj: vlastní

Výsledky ukazují, že metoda Stacking s meta-klasifikátorem náhodný les převyšuje ostatní klasifikátory u všech hodnotících metrik, tj. přesnost (Acc), AUC, Giniho koeficient i náklady chybné klasifikace (MC), proto byla tato metoda zvolena jako nejvhodnější pro modelování pravděpodobnosti defaultu ve dvoufázovém modelu.

Na grafech níže (obrázek 9) je možné pozorovat poměr dobrých (ne-default) a špatných (default) klientů z hlediska pravděpodobnosti defaultu. Cílem je mít co nejméně špatných

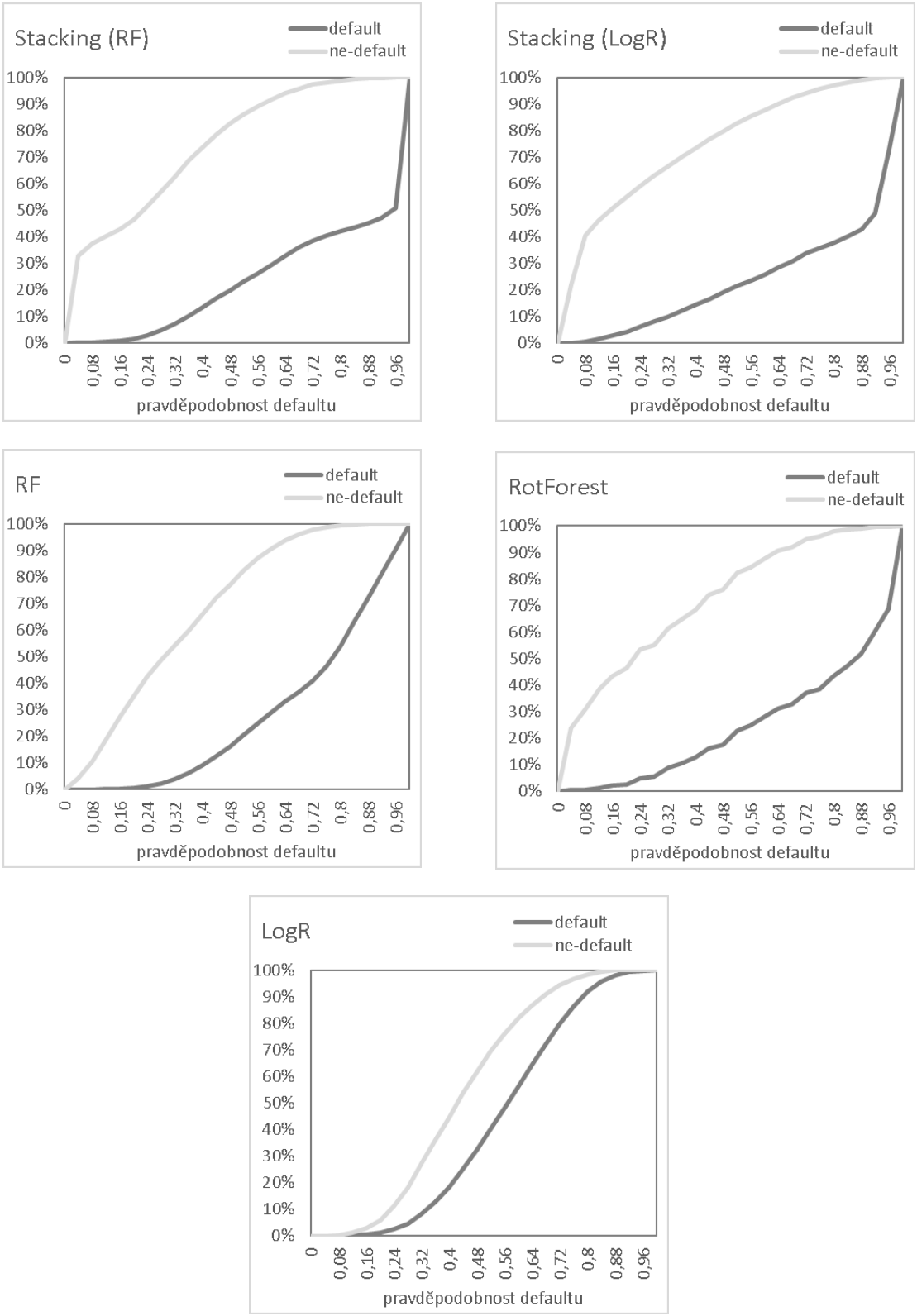
klientů v levé části grafu, kdy je pravděpodobnost defaultu nejnižší a co nejvíce špatných klientů na opačné straně grafu. Skládané sloupcové grafy níže ukazují čtyři nejlepší metody z hlediska jednotlivých metrik (Stacking (náhodný les, RF), náhodný les (RF), Stacking (LogR) a rotační les (RotForest)) a pro porovnání je zde i LogR.

Distribuční funkce dobrých a špatných klientů, na základě kterých lze zjistit Kolmogorov-Smirnovovu statistiku, která je dána jako maximální rozdíl mezi kumulativním rozdělením špatných a kumulativním rozdělením dobrých klientů je možné vidět na obrázku 10, kde jsou zobrazeny stejné metody jako v obrázku 9.



Obrázek 9: Poměr dobrých a špatných klientů dle pravděpodobnosti defaultu

Zdroj: vlastní



Obrázek 10: Distribuční funkce dobrých a špatných klientů dle pravděpodobnosti defaultu

Zdroj: vlastní

Dále byly ze souboru dat vybrány nesplacené úvěry, pro které byla předpovězena expozice při defaultu pomocí regresních modelů. Stejně jako modelování pravděpodobnosti defaultu, regresní modely zahrnovaly jak individuální regresory (REP Tree (Witten a Frank 2005), alternující modelovací stromy (AMT) (Witten a Frank 2005), M5P rozhodovací stromy (Frank et al. 1998), náhodný les (RandomTree) (Witten a Frank 2005), lineární regrese (LR) (Witten a Frank 2005), podpůrná vektorová regrese (SVR) (Yao et al. 2017) a neuronová síť typu perceptron (NN) (Yang a Tkachenko 2012)), tak homogenní kombinace modelů (náhodný les (RF) (Lessmann et al. 2015), rotační les (RotForest) (Abellán a Castellano 2017), bagging (Wang et al. 2012) a Random SubSpace (Wang et al. 2012)) a heterogenní kombinace modelů Stacking (Dzeroski a Zenko 2004). Metoda Stacking s LR překonala metodu Stacking s RF z hlediska R^2 a RMSE (viz tabulka 11). Navíc model vystavěný na metodě Stacking s LR překonal většinu ostatních modelů ve všech hodnotících metrikách kromě AMT, M5P a LR. Ze všech homogenních kombinací modelů má nejlepší výsledky RotForest. Špatný výkon RF lze vysvětlit špatným výkonem náhodných stromů, které byly použity jako základní regresory, což lze připsat skutečnosti, že nebylo prováděno ladění hyperparametrů z důvodu zabránění přeučení ve fázi meta-učení.

Metoda Stacking s LR byla vyhodnocena jako nejlepší, a proto byla použita ve dvou fázovém modelu očekávané ztráty (EL). Nejlepšího výsledku nedosáhla pouze s ohledem na MAE, ale poskytla statisticky podobný výsledek (Wilcoxonův test).

Tabulka 11: Výkonnost regresorů modelů expozice při defaultu

Metoda	R^2	RMSE	MAE
REPTree	0,0731 ± 0,0216	0,1075 ± 0,0027	0,0861 ± 0,0020
AMT	0,1136 ± 0,0111	0,1040 ± 0,0016	0,0836 ± 0,0016*
M5P	0,1229 ± 0,0178	0,1033 ± 0,0020	**0,0832 ± 0,0019**
RandomTree	0,0122 ± 0,0080	0,1477 ± 0,0034	0,1123 ± 0,0029
LR	0,1231 ± 0,0189*	0,1032 ± 0,0021*	0,0833 ± 0,0019
SVR	0,0993 ± 0,0170	0,1067 ± 0,0025	0,0840 ± 0,0020*
NN	0,0723 ± 0,0318	0,1119 ± 0,0049	0,0909 ± 0,0030
RF	0,0890 ± 0,0188	0,1053 ± 0,0019	0,0853 ± 0,0014
RotForest	0,1156 ± 0,0134	0,1039 ± 0,0019	0,0844 ± 0,0013
Bagging	0,0840 ± 0,0172	0,1064 ± 0,0019	0,0856 ± 0,0016
RandomSubSpace	0,1026 ± 0,0239	0,1048 ± 0,0016	0,0851 ± 0,0013
Stacking (LR)	**0,1305 ± 0,0187**	**0,1028 ± 0,0019**	0,0836 ± 0,0016*
Stacking (RF)	0,1068 ± 0,0139	0,1046 ± 0,0014	0,0836 ± 0,0011*

** Statisticky významně podobné modely (p-hodnota < 0,05) vzhledem k nejlepšímu modelu označenému tučně.*

Zdroj: vlastní

Na závěr byly obě fáze spojeny, jak modelování pravděpodobnosti defaultu pomocí metody Stacking s RF, tak modelování expozice při defaultu pomocí metody Stacking s LR, přesněji řečeno, druhá fáze byla aplikována pouze na úvěry, které byly klasifikované jako špatné v první fázi. To znamená, že nesplacené úvěry zahrnovaly jak úvěry klasifikované jako správně špatné úvěry (true positive, TP), tak nesprávně špatné úvěry (false positive, FP). Naproti tomu očekáváná ztráta, EL = 0, byla přiřazena jak úvěrům klasifikovaným v první fází jako správně dobré úvěry (true negative, TN), tak úvěrům klasifikovaným jako nesprávně dobré úvěry (false negative, FN). Jinými slovy, EL úvěrů v kategorii TP a FP koresponduje s predikcí expozice při defaultu získanou pomocí metody Stacking s LR ve druhé fázi.

Tabulka 12 porovnává navrhovaný dvoufázový model s ostatními nejmodernějšími modely používanými v úvěrovém skórování. Navrhovaný model představuje efektivnější nástroj pro modelování EL. Mezi srovnávané metody patří již dříve použité metody pro modelování ztráty a expozice při defaultu. Jsou rozděleny do dvou kategorií, na jednofázové a dvoufázové modely. Jednofázové modely předvídají EL přímo, aniž by kategorizovaly spotřebitelské úvěry do tříd dle pravděpodobnosti defaultu. Pro porovnání jednofázových modelů byl použit model Stacking s LR, který v této práci nejlépe predikoval expozici při defaultu a dále metoda LR (Caselli et al. 2008; Gürtler et al. 2018), SVR (Yao et al. 2015), NN (Yang a Tkachenko 2012) a regresní stromy (Bastos 2010). Metody nepřímého odhadu

expozice při defaultu použité v předchozích studiích (Tong et al. 2016; Leow a Crook 2016) není v této práci možné použít kvůli fixní expozici spotřebitelských úvěrů.

Pro porovnání dvoufázových modelů byla použita stejná metodika, jaká byla použitá pro model navržený v této práci. Byla předpovězena pravděpodobnost defaultu v první fázi a poté bylo předpovězena expozice při defaultu pro ty úvěry, které byly klasifikované jako špatné, tj. defaultní. Byly testovány čtyři modely, konkrétně LogR + SVR (Loterman et al. 2012), LogR + NN (Loterman et al. 2012), SVM + SVR (Yao et al. 2017) a RF + NN. Model RF + NN kombinuje dva srovnávací algoritmy, které byly efektivně použity pro modelování ztráty a expozice při defaultu v předchozích studiích. Tento model je aplikován na datový soubor, který byl vybalancován (podvzorkován), tzn. na stejný datový soubor jako soubor použitý v navrhovaném dvoufázovém modelu. Zbývající srovnávané modely jsou vytvořeny na původním datovém souboru tak, jak byly navrženy v odpovídajících studiích. Nastavení parametrů učení pro všechny modely je uvedeno v tabulce 6. Po porovnání výsledků z obou kategorií není patrná nadřazenost jednofázových ani dvoufázových modelů. Průměrná hodnota R^2 se pohybovala přibližně od 4 % pro NN až po 39 % pro navrhovaný dvoufázový model (Stacking s RF + Stacking s LR). Navrhovaný dvoufázový model EL tak převyšoval všechny ostatní modely ve všech hodnotících metrikách.

Tabulka 12: Porovnání výkonnosti navrhovaného dvoufázového modelu očekávané ztráty EL s nejmodernějšími přístupy modelování úvěrového rizika

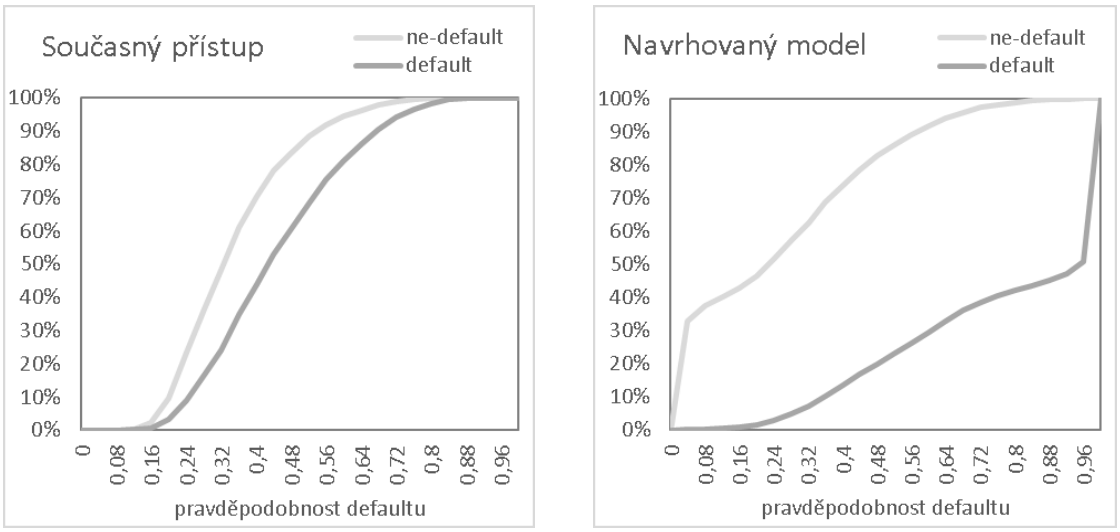
Kategorie	Metoda	R^2	RMSE	MAE
Jedna fáze	NN	0,044 ± 0,015	0,302 ± 0,042	0,281 ± 0,036
	SVR	0,058 ± 0,003	0,277 ± 0,001	0,267 ± 0,001
	RF	0,113 ± 0,006	0,267 ± 0,001	0,242 ± 0,001
	M5P	0,125 ± 0,007	0,265 ± 0,001	0,239 ± 0,001
	LR	0,126 ± 0,006	0,265 ± 0,001	0,239 ± 0,001
	Stacking (LR)	0,134 ± 0,007	0,264 ± 0,001	0,237 ± 0,001
Dvě fáze	RF + NN	0,050 ± 0,009	0,471 ± 0,019	0,422 ± 0,020
	LogR + SVR	0,176 ± 0,006	0,355 ± 0,002	0,307 ± 0,001
	SVM + SVR	0,193 ± 0,007	0,326 ± 0,004	0,269 ± 0,003
	LogR + NN	0,321 ± 0,022	0,356 ± 0,032	0,276 ± 0,070
	Navrhovaný model Stacking (RF) + Stacking (LR)	**0,390 ± 0,019**	**0,221 ± 0,004**	**0,165 ± 0,003**

Zdroj: vlastní

6.4 Efektivnost navrhovaného modelu v porovnání se současným stavem

Efektivnost navrhovaného modelu pro využití v praxi je prezentována porovnáním současného přístupu používaného nebankovní finanční institucí (logistická regrese) a navrhovaného modelu. Z předchozích experimentů v této práci je navržen model na základě heterogenní kombinace modelů pomocí metody Stacking s RF (fáze 1). Pro vytvoření obou modelů byl využit datový soubor zmíněný výše v této práci.

Distribuční funkce dobrých a špatných klientů dle pravděpodobnosti defaultu jsou uvedeny na obrázku 11 pro a) současný přístup společnosti a b) navrhovaný model. Je patrné, že navrhovaný model dokáže lépe oddělit dobré a špatné klienty, což dokazují i výsledky v tabulce 13. Přesnost modelu se současným přístupem společnosti je 66,63 %, kdežto přesnost navrhovaného modelu je 82,28 %.



Obrázek 11: Distribuční funkce dobrých a špatných klientů dle pravděpodobnosti defaultu a) současný přístup společnosti, b) navrhovaný model

Zdroj: vlastní

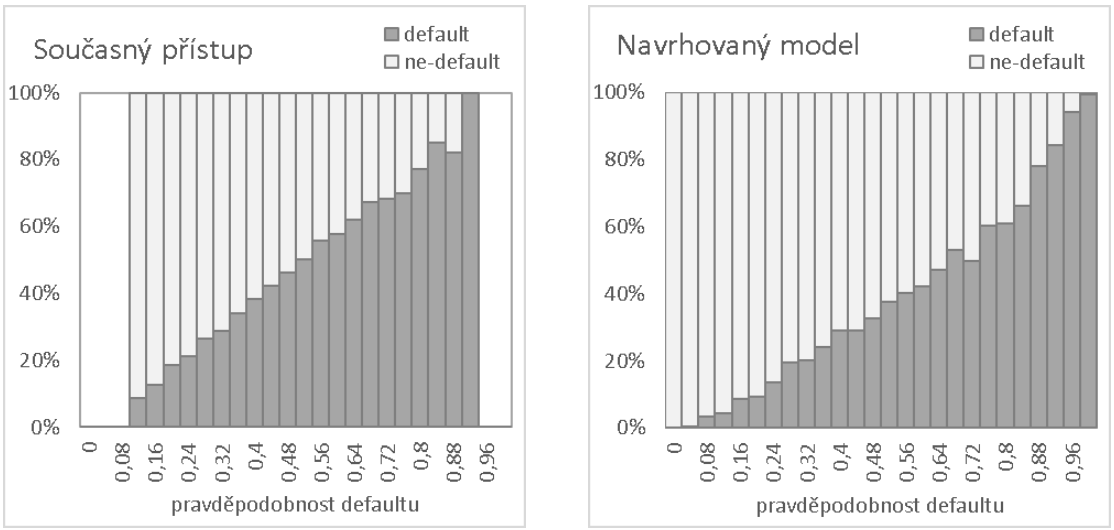
V tabulce 13 jsou kromě přesnosti modelu uvedeny i další hodnoty, které slouží pro porovnání obou modelů. Jmenovitě jde o Giniho koeficient, Kolmogorovovu-Smirnovovu (KS) statistiku, plochu pod ROC křivkou (AUC) a náklady chybné klasifikace (MC), které byly vypočteny pomocí přístupu uvedeného dříve v této práci. Ve všech hodnotících metrikách byl navrhovaný model výrazně lepší.

Tabulka 13: Porovnání výkonnosti a) současný přístup společnosti, b) navrhovaný model

Metoda	Přesnost	AUC	MC	Gini koeficient	KS statistika
Současný přístup společnosti	0,6663	0,6806	0,4878	0,3611	0,2638
Navrhovaný model (z fáze 1)	0,8228	0,9128	0,2446	0,8257	0,6278

Zdroj: vlastní

Na obrázku 12 je pro srovnání znázorněn poměr dobrých a špatných klientů dle pravděpodobnosti defaultu (v procentech) pro a) současný přístup společnosti a b) navrhovaný model.

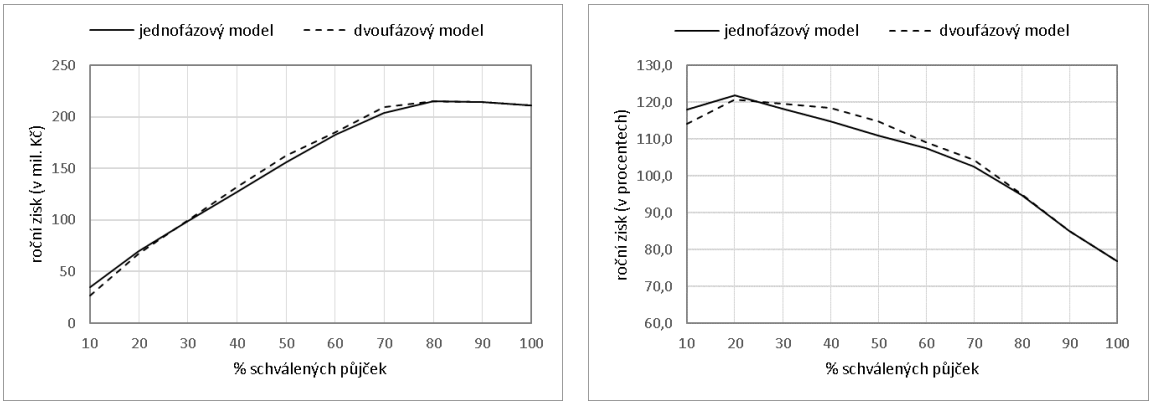


Obrázek 12: Poměr dobrých a špatných klientů dle pravděpodobnosti defaultu a) současný přístup společnosti, b) navrhovaný model

Zdroj: vlastní

Dále byla efektivnost navrhovaného modelu pro využití v praxi demonstrována pomocí vypočteného ročního zisku, kterého věřitel může dosáhnout při výběru 10 %, 20 %, …, 100 % nejlepších úvěrů podle 1) skóre dle pravděpodobnosti defaultu získaného pomocí Stacking s RF a 2) skóre dle očekávané ztráty získané navrhovaným dvoufázovým přístupem. Obrázek 13 znázorňuje dosažený roční zisk v absolutním a relativním vyjádření, což naznačuje, že navrhovaný přístup vede k lepší ekonomické výkonnosti. Roční zisk pro každý úvěr byl vypočten jako rozdíl příjmů z úvěru za rok (tj. součet měsíčních splátek, které klient během roku skutečně uhradil) a vyplacené částky přepočtené na jeden rok. Relativní vyjádření je potom poměr ročního zisku v absolutním vyjádření a očekávaného příjmu z úvěru, který je dán součtem všech měsíčních splátek za rok. Dle očekávání průměrný relativní zisk klesá s vyšším podílem schválených úvěrů, tzn. s vyšším úvěrovým rizikem. Nejvyšší roční zisk

byl získán výběrem 80 % úvěrů. Úvěrové portfolio je ziskové bez ohledu na jeho úvěrové riziko, to lze přičíst vysokým úrokovým sazbám. Jinak řečeno, mnohé úvěry v prodlení jsou stále ziskové kvůli vysokému úroku, který kompenzuje vysoké úvěrové riziko.



Obrázek 13: Roční zisk dosažený pomocí modelu pravděpodobnosti defaultu (jednofázový model) a navrhovaného dvoufázového modelu; a) v mil. Kč, b) v procentech

Zdroj: vlastní

6.5 Kalibrace stávající skórovací karty

V této kapitole je kalibrována skórovací karta, která je v současné době používána v nebankovní finanční instituci, jejíž data byla použita pro experimenty v této práci. Stávající skórovací karta obsahuje 6 proměnných, konkrétně podíl vyplacené částky ku volným zdrojům v korunách (LTD), měsíční splátku, věk klienta, hodnotu skóre z Nebankovního registru klientských informací (CBS), výši splátek z registru SOLUS a typ bydlení. Z těchto 6 proměnných a jedné cílové proměnné (informace, zda došlo k selhání se splácením nebo ne) je pomocí logistické regrese vytvořen model, jehož parametry jsou uvedeny v tabulce 14. Na základě experimentů v předchozí části práce bylo predikováno selhání se splácením úvěru pomocí heterogenní kombinace modelů metodou Stacking s RF (navrhovaný model). Tyto predikce byly použity na naučení logistické regrese namísto skutečného pozorování. Výsledné kalibrované parametry jsou rovněž uvedeny v tabulce 14. Z této tabulky vyplývá, že nejdůležitějšími parametry skórovací karty jsou parametry typ bydlení, věk klienta a LTD. Například lze říci, že s rostoucím věkem klienta klesá pravděpodobnost defaultu (resp. roste skóre). Naopak s rostoucí hodnotou LTD pravděpodobnost defaultu roste (tj. skóre klesá).

Tabulka 14: Porovnání parametrů skórovací karty a) současný model společnosti, b) současný model společnosti po kalibraci parametrů (navrhovaný model)

Proměnná	Parametr (referenční třída = default)	
	Současný model	Navrhovaný model
Konstanta	0,547219	1,560921
LTD	0,006108	0,017134
Měsíční splátka	0,000097	0,000141
Věk klienta	-0,022688	-0,044204
CBS	-0,001443	-0,002734
Výše splátek z registru SOLUS	0,000151	0,000321
Typ bydlení		
1	-0,414730	0,830896
2	0,059960	0,007739
3	0,252385	0,551782
4	0,299562	0,385855
5	0,132478	0,401067
6	0,005541	0,174365
7	-0,287547	-0,505357

Zdroj: vlastní

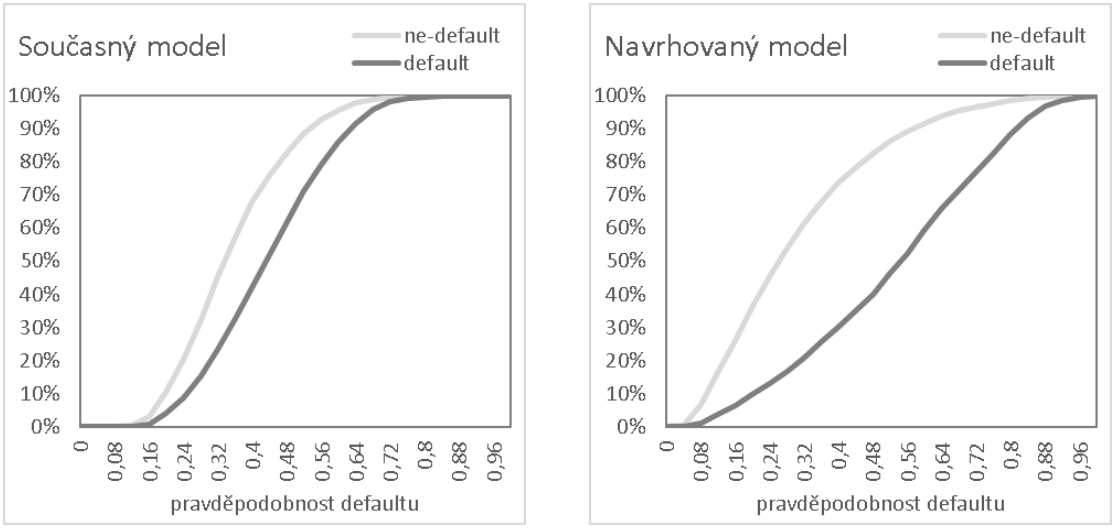
Distribuční funkce dobrých a špatných klientů dle pravděpodobnosti defaultu jsou uvedeny na obrázku 14 pro a) současný model společnosti a b) současný model společnosti po kalibraci parametrů. Je zřejmé, že navrhovaný model dokáže lépe oddělit dobré a špatné klienty, což dokazují i výsledky v tabulce 15. Přesnost současného modelu společnosti je 65,74 %, kdežto přesnost navrhovaného modelu je 73,82 %. Tuto přesnost lze zvýšit navýšením počtu proměnných jako vysvětlujících proměnných, což dokazují výsledky podkapitoly 6.4, kde byly využity všechny dostupné proměnné.

Tabulka 15: Porovnání výkonnosti a) současný model společnosti, b) současný model společnosti po kalibraci parametrů (navrhovaný model)

Model	Přesnost	AUC	MC	Gini koeficient	KS statistika
Současný model společnosti	0,6574	0,6690	0,5031	0,3380	0,2597
Navrhovaný model	0,7382	0,7781	0,3723	0,5562	0,4356

Zdroj: vlastní

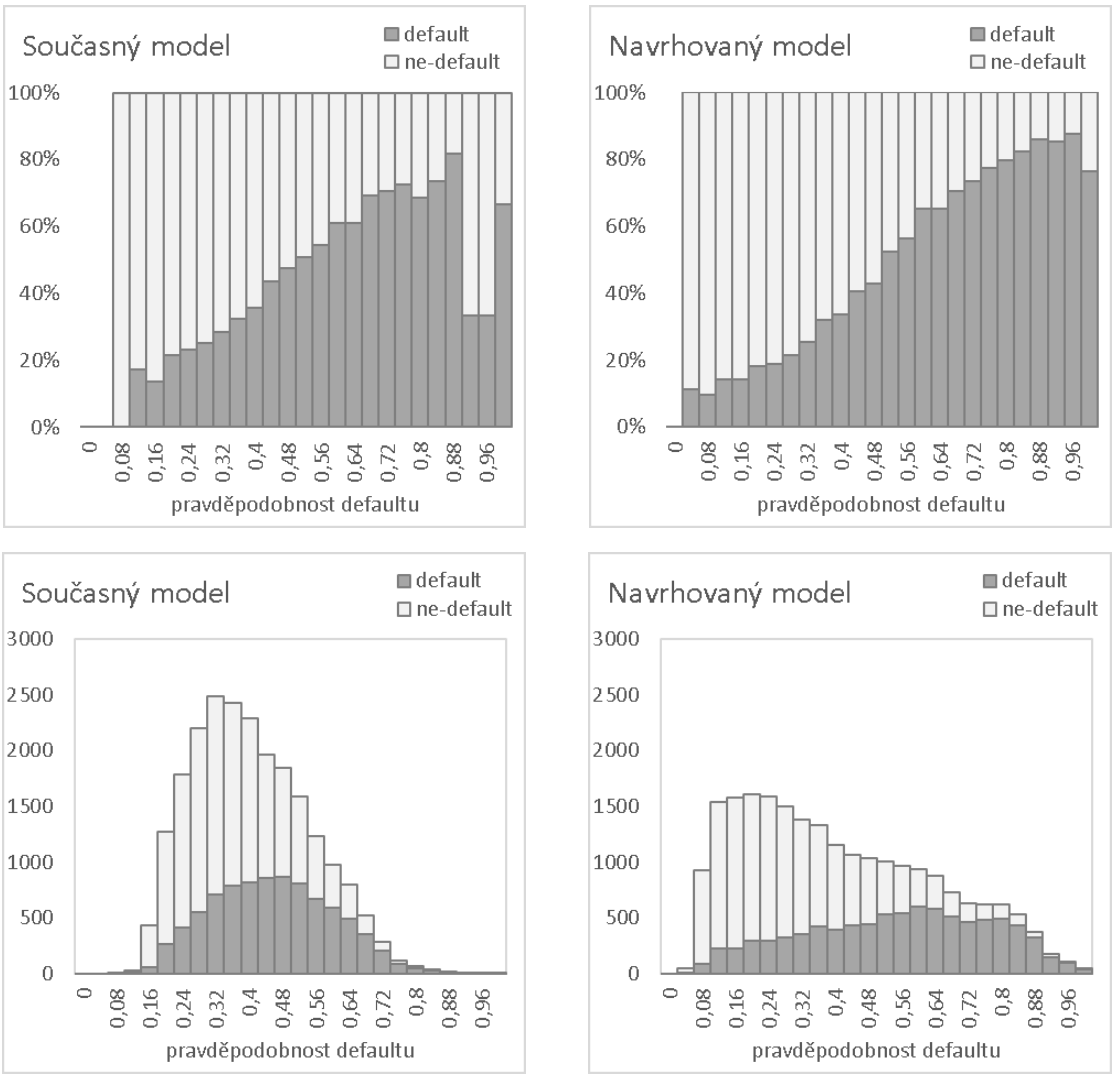
V tabulce 15 jsou kromě přesnosti uvedeny i další hodnoty, které slouží pro porovnání obou modelů. Jmenovitě jde o Giniho koeficient, Kolmogorovovu-Smirnovovu (KS) statistiku, plochu pod ROC křivkou (AUC) a náklady chybné klasifikace (MC), které byly vypočteny pomocí přístupu uvedeného výše.



Obrázek 14: Distribuční funkce dobrých a špatných klientů dle pravděpodobnosti defaultu a) současný model společnosti, b) současný model společnosti po kalibraci parametrů (navrhovaný model)

Zdroj: vlastní

Na obrázku 15 je znázorněn poměr dobrých a špatných klientů dle pravděpodobnosti defaultu (v procentech a kusech) pro a) současný model společnosti a b) současný model společnosti po kalibraci parametrů. Je patrné, že navrhovaný model dokáže lépe oddělit dobré a špatné klienty, což dokazují i výsledky jednotlivých metrik v tabulce 15.



Obrázek 15: Poměr dobrých a špatných klientů dle pravděpodobnosti defaultu a) současný model společnosti, b) současný model společnosti po kalibraci parametrů (navrhovaný model)

Zdroj: vlastní

7 Diskuze

Vzhledem k rostoucí popularitě kombinací modelů použitých při modelování pravděpodobnosti defaultu lze předpokládat, že tyto metody mohou být efektivní i při modelování klíčových ukazatelů úvěrového rizika. Tato práce se tedy zaměřila na modelování úvěrového rizika s využitím dvoufázového přístupu. První fáze se zabývala modelováním pravděpodobnosti defaultu a ve druhé fázi byl modelován ukazatel expozice při defaultu. Byla zkoumána výkonnost homogenních i heterogenních kombinací modelů pro modelování úvěrového rizika při nevybalancovaném datovém souboru. Bylo zjištěno, že heterogenní kombinace modelů překonaly homogenní kombinace modelů při modelování pravděpodobnosti modelů, což odpovídá předchozí srovnávací studii autorů Lessmann a kol. (Lessmann et al. 2015).

Vynikající výkonnost heterogenních kombinací modelů lže přičítat různorodosti defaultních úvěrů v použité datové množině. Tyto výsledky mohou být také způsobeny nenormálním rozdělením dat EL. Dále bylo prokázáno, že výkonnost kombinací modelů při modelování pravděpodobnosti defaultu může být vylepšena pomocí náhodného lesu (RF) jako meta-učící metody tzn., že nelineární klasifikátor zlepšil výkonnost spíše než tradiční lineární klasifikátor logistická regrese. Toto zjištění, které je v souladu s nedávným použitím nelineárních metod meta-učení v modelování pravděpodobnosti defaultu (Yu et al. 2016), má důležité důsledky pro vývoj budoucích modelů pravděpodobnosti defaultu.

Nejzajímavějším zjištěním mezi jednotlivými výsledky bylo, že metoda Stacking s RF převyšovala ostatní metody z hlediska sensitivity (True Positive Rate, TPR). U modelů predikce pravděpodobnosti defaultu tato metrika totiž hraje rozhodující úlohu. Dosažení vysoké hodnoty TPR znamená nižší náklady chybné klasifikace (MC) spojené s nesplacením úvěrů. Pokud jde o TPR je ukázáno, že modely založené na RF obecně fungují dobře. Toto zjištění podporuje použití RF jako referenční metody při modelování pravděpodobnosti defaultu (Lessmann et al. 2015).

Výsledky v této práci potvrzují poznatky několika autorů (Brown a Mues 2012; He et al. 2018), kteří ukázali, že podvzorkování v modelování pravděpodobnosti defaultu zlepšilo přesnost srovnávaných klasifikátorů. Toto zjištění je ovšem v kontrastu s preferencí nadvzorkování autorů Marqués a kol. (Marqués et al. 2013), což může být vysvětleno poměrně velkou datovou sadou použitou v této práci, která umožňuje eliminovat redundantní data ve většinové třídě bez ohrožení velikosti trénovací množiny. Jak bylo uvedeno dříve,

podvzorkování bylo efektivnější než modelování pravděpodobnosti defaultu na originálních datech, tj. na nevyváženém datovém souboru. Lze říci, že převzorkování zvyšuje výkon modelování pravděpodobnosti defaultu.

Heterogenní kombinace modelů doposud nebyly použity pro modelování expozice při defaultu. Bylo zjištěno, že tyto metody mírně převyšují základní metody. Metoda Stacking s RF jako meta-regresorem pro modelování expozice při defaultu nevyšla dobře, což naznačuje, že metoda lineární regrese postačí pro kombinování předpovědí získaných jednotlivými základními metodami. Obecně se zdá, že heterogenní kombinace modelů znamenají větší zlepšení u modelů pravděpodobnosti defaultu než u modelů expozice při defaultu. To může být vysvětleno také větším počtem základních metod u modelu pravděpodobnosti defaultu.

Tato práce dále přispívá k výpočtu nákladů chybné klasifikace. Měření výkonnosti bylo pro modely pravděpodobnosti defaultu považováno za klíčové kvůli vyšším nákladům chybné klasifikace (MC), které se pojí se selháním spotřebitelského úvěru. Navrhované měření MC bylo navrženo pro fixní expozice, jako jsou spotřebitelské úvěry nebo hypoteční úvěry. MC spojené s klasifikací nesprávně špatný úvěr (FP) jsou 0,749, zatímco pro klasifikaci nesprávně dobrý úvěr (FN) jde o 0,597. Poměr je tedy přibližně 1,25:1, což je v rozporu s očekáváním. Nebyl nalezen zásadní rozdíl mezi náklady na chybnou klasifikaci FP a FN pro spotřebitelské úvěry nebankovní finanční instituce. Nízký poměr MC lze připsat vyšší úrokové sazbě, kterou nebankovní finanční instituce účtuje. Poměr tedy značně závisí na typu finanční instituce, resp. na výši úroků. Obecně lze říci, že metrika MC navržená v této práci může být užitečná pro vývoj modelů pravděpodobnosti defaultu bankovních i nebankovních finančních institucí.

Kombinace nejlepších modelů pravděpodobnosti defaultu a expozice při defaultu v rámci dvoufázového přístupu poskytla účinný model pro predikci celkové očekávané ztráty úvěru v porovnání s nejmodernějšími modely. Účinnost lze přičíst jak dvoufázovému přístupu modelování očekávané ztráty, tak i rovnováze trénovacích dat při modelování pravděpodobnosti defaultu. Modely, které kombinují pravděpodobnost defaultu a expozice při defaultu ve skutečnosti vykazují obecně lepší výsledky než tradiční jednofázové modely. Další zvýšení predikční síly lze přičíst i použití náhodných lesů při modelování pravděpodobnosti defaultu na vybalancovaném datovém souboru. Poměrně špatnou výkonnost ostatních dvoufázových modelů (logistické regrese a podpůrných vektorových strojů) lze

vysvětlit špatnou schopností zpracovávat nevyvážené datové soubory. Celkově byla odchylka očekávané ztráty (dáno R^2), kterou lze vysvětlit srovnávací modely, trvale pod 39 %, což znamená, že nemohou vysvětlit většinu rozptylu v datech. To je v souladu s předchozími studiemi a umožňuje budoucí zpřesnění, např. pomocí dalších vstupních proměnných.

Dále bylo prokázáno, že predikce úvěrového rizika spotřebitelských úvěrů pomocí dvoufázového modelu může vést k výběru ziskovějších úvěrových portfolií než těch, které jsou založeny na tradičním modelování pravděpodobnosti defaultu, což lze připsat přesnějšímu odhadu očekávané ztráty vzhledem k expozici při defaultu. Toto zjištění potvrzuje i výsledky získané ze ziskového skórování autorů Serrano-Cinca a Gutiérrez-Nieto (Serrano-Cinca a Gutiérrez-Nieto 2016). V této práci však nebyl systém navržen tak, aby maximalizoval zisk bez ohledu na úvěrové riziko. Namísto toho mohou věřitelé využít přesnější odhad očekávané ztráty tak, aby mohli spravovat úvěrové riziko svých portfolií spotřebitelských úvěrů, včetně podpory rozhodnutí o sekuritizaci portfolia a stanovení cen na úrovni jednotlivých spotřebitelů.

8 Přínosy disertační práce

Na základě analýzy současného stavu řešení v oblasti modelování úvěrového rizika bylo cílem práce navrhnout model skórovacích karet, který umožní modelování pravděpodobnosti defaultu a expozice při defaultu pomocí heterogenních kombinací metod strojového učení. Vědecké a aplikační přínosy tohoto modelu jsou následující.

8.1 Vědecké přínosy

Mezi vědecké přínosy disertační práce lze zařadit:

- Návrh **heterogenní kombinace modelů pravděpodobnosti defaultu**. Tento heterogenní model kombinuje výstupy v současnosti nejefektivnějších individuálních klasifikátorů tak, že je dosaženo vyšší přesnosti klasifikace a nižších nákladů chybné klasifikace v porovnání se současným stavem řešení.
- Porovnání různých přístupů k **vybalancování datového souboru**. K vybalancování tříd na trénovací množině dat byl použit algoritmus podvzorkování většinové třídy, což se ukázalo jako efektivnější v porovnání s v současnosti používanou metodou nadvzorkování dat pomocí uměle vygenerovaných trénovacích případů. Tato práce tedy potvrdila, že problém nevyváženého souboru dat při modelování pravděpodobnosti defaultu může být překonán pomocí podvzorkování většinové třídy.
- Návrh **heterogenní kombinace modelů expozice při defaultu**. Tento přístup je v současnosti v oblasti modelování úvěrového rizika unikátní, neboť kombinuje různorodé individuální regresory (včetně rozhodovacích stromů, neuronových sítí a podpůrné vektorové regrese). Kombinování výstupů těchto metod do jednoho heterogenního modelu přináší možnost modelovat různé úvěrové profily defaultu a tím dosažení nižší chyby predikce očekávané ztráty vzhledem k současným predikčním modelům.
- Koncepční rámec pro modelování úvěrového rizika. V práci je navržen **dvoufázový model úvěrového rizika** spotřebitelských úvěrů**,** který se skládá z modelu pravděpodobnosti defaultu a modelu expozice při defaultu. Ve srovnání se současným stavem řešení umožňuje tento model nejen predikci pravděpodobnosti defaultu, ale také predikci ekonomických dopadů defaultu každého úvěru. V každé fázi je navržena heterogenní kombinace modelů s cílem modelovat různorodé profily klientů.

Tyto heterogenní kombinace modelů integrují základní klasifikátory (regresory), které jsou v současné době nejvýkonnějšími metodami strojového učení v oblasti klasifikace defaultu, resp. predikce expozice při defaultu. Výstupy základních modelů jsou zkombinovány pomocí náhodného lesu, který provádí vnořenou selekci výstupů základních metod, čímž se odlišuje od v současné době používaných lineárních metod.

- Dále byla navržena **metrika klasifikace defaultu** vzhledem k odlišným nákladům způsobených chybnou klasifikací dobrých a špatných úvěrů. Výpočet zahrnuje náklady obětované příležitosti a ztrátu investice v případě špatně klasifikovaných úvěrů. Tato metrika využívá parametr ztrátu při defaultu, která je dalším klíčovým parametrem úvěrového rizika. Tato metrika se od předchozích liší zejména v tom, že bere v úvahu také náklady obětované příležitosti a je vyjádřená v relativních hodnotách, což umožňuje její použití při porovnání různých metod klasifikace na různých datových souborech.
- Navržené modely jsou **ověřeny na reálných datech české nebankovní finanční instituce** a je provedena **srovnávací analýza výsledků navrženého modelu** se: (a) v současnosti používanými modely pro predikci pravděpodobnosti defaultu a expozice při defaultu; (b) jednofázovými a dvoufázovými modely predikce úvěrového rizika. Porovnání je provedeno jak z hlediska přesnosti predikce, nákladů chybné klasifikace, tak celkového dosaženého zisku. Je ukázáno, že navrhovaný model převyšuje výsledky skórovacího modelu, který v současné době používá nebankovní finanční společnost, jež poskytla reálná data.

8.2 Aplikační a ekonomické přínosy

Vzhledem k aplikaci navržených modelů na reálných datech konkrétní nebankovní finanční instituce má práce také několik aplikačních a ekonomických přínosů:

- Ačkoliv byl model skórovací karty navržen pro potřeby zefektivnění rozhodování nebankovní finanční instituce, navržený model je možné **aplikovat také v prostředí bankovní finanční instituce**. Jak bylo naznačeno výše, vše, co je k tomu potřeba, je modifikace parametrů použitých při výpočtu nákladů chybné klasifikace.

- Navržený dvoufázový model má navíc možné uplatnění i **v dalších úlohách**, kde je možné vyjádřit náklady spojené s jednou třídou objektů. Může to být například predikce dopadů bankrotu podniků, ztráty související s výpadky v dodávkách výrobků atd.
- Zvýšení přesnosti modelů skórovacích karet v bankovních i nebankovních institucích může mít pozitivní dopady na **zefektivnění úvěrového trhu** a **snížení asymetrie informací** mezi věřitelem a žadatelem o úvěr. Kromě jiného je možné zpřesnit **rozhodování o poskytnutí úvěru a nastavení úroků** odpovídající výši úvěrového rizika.
- Provedení **kalibrace** v současnosti používané **skórovací karty**. Implementací kalibrované stávající skórovací karty může tato společnost **zvýšit přesnost klasifikace žadatelů o úvěr a snížit tak náklady chybné klasifikace**.
- **Efektivnější správa úvěrového rizika portfolia** spotřebitelských úvěrů u bankovních a nebankovních finančních institucí, včetně efektivnějšího rozhodování o sekuritizaci portfolia a stanovení cen úvěrů na úrovni jednotlivých spotřebitelů. To je umožněno porovnáním úvěrového rizika a očekávaného zisku celého portfolia.

Závěr

Disertační práce se zabývala modelováním úvěrového rizika. Práce se zaměřila na modelování pravděpodobnosti defaultu a expozice při defaultu, kdy byla zkoumána výkonnost homogenních i heterogenních kombinací modelů. Cílem práce bylo navrhnout model skórovacích karet, který umožní modelování pravděpodobnosti defaultu a expozice při defaultu pomocí heterogenních kombinací metod strojového učení. Ke splnění tohoto cíle byl v práci navržen dvoufázový model úvěrového rizika spotřebitelských úvěrů založený na heterogenní kombinaci modelů strojového učení. Výsledky této práce podpořily myšlenku, že heterogenní kombinace modelů jsou efektivní při modelování úvěrového rizika, které zde bylo odhadováno pomocí parametrů pravděpodobnost defaultu a expozice při defaultu. Kromě toho tato práce ukázala, že při modelování pravděpodobnosti defaultu musí být řešen problém nevyváženého souboru dat. V této práci bylo potvrzeno, že účinným postupem k překonání tohoto problému je podvzorkování většinové třídy. K určení výkonu modelů pravděpodobnosti defaultu byla navržena metrika výpočtu nákladů chybné klasifikace, která je vhodná pro předvídání úvěrů s fixní expozicí. V první fázi navrhovaného modelu bylo ukázáno, že náklady chybné klasifikace mohou být sníženy použitím heterogenních kombinací modelů s využitím algoritmu náhodný les jako meta-klasifikátoru. To se zdálo být kritické pro druhou fázi, kdy byla expozice při defaultu předpovězena za použití lineární kombinace základních regresorů. Toto zjištění podpořilo relevantnost dvoufázového modelu.

Nejdůležitější omezení disertační práce spočívá v tom, že v navrhovaném modelu byla stanovena fixní hodnota ztráty při defaultu. Přesněji řečeno, ztráta při defaultu byla zohledněna pouze při výpočtu nákladů chybné klasifikace, proto jsou nutné další studie, které budou brát v úvahu ztrátu při defaultu jako další predikovaný parametr úvěrového rizika. Za druhé, navrhovaný model má i nadále omezenou vysvětlují schopnost (vysvětluje méně než 39 % rozptylu). Tento výsledek potvrzuje výkonnost pozorovanou v článku autorů Yang a Tkachenko (Yang a Tkachenko 2012). Je předpoklad, že stále existuje prostor pro zlepšení. Model byl navržen tak, aby upřednostnil predikční sílu, což vedlo ke kombinaci dvou komplexních kombinací modelů, které způsobily omezenou srozumitelnost výsledného navrhovaného modelu. Doporučují se proto budoucí studie, které se budou zabývat jak výkonností, tak interpretací modelů. Další výzkum by mohl rovněž řešit účinnost navrženého modelu v souvisejících úlohách modelování úvěrového rizika. Jedná se o (1) predikci úvěrového rizika

úvěrů s revolvingovými expozicemi či (2) predikci úvěrového rizika úvěrů s fixními expozicemi u bankovních finančních institucí. V případě (1) je vhodnější nahradit metriku výpočtu nákladů chybné klasifikace z této práce metrikou použitou v článku autorů Bahnsen a kol. (Bahnsen et al. 2015). Pro druhý případ by měl být použit vyšší poměr nákladů chybné klasifikace.

Seznam použité literatury

[1] ABDOU, H. a J. POINTON, 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance & Management*. **18**(2–3), 59–88.

[2] ABEDINI, Mohammadali, Farzaneh AHMADZADEH a Rassoul NOOROSSANA, 2016. Customer credit scoring using a hybrid data mining approach. *Kybernetes*. **45**(10), 1576–1588. ISSN 0368-492X.

[3] ABELLÁN, Joaquín a Javier G. CASTELLANO, 2017. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*. **73**, 1–10. ISSN 09574174.

[4] ABELLÁN, Joaquín a Carlos J. MANTAS, 2014. Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*. B.m.: Pergamon, **41**(8), 3825–3830. ISSN 0957-4174.

[5] ADNAN, Md Nasim a Md Zahidul ISLAM, 2017. Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. *Expert Systems with Applications*. B.m.: Pergamon, **89**, 389–403. ISSN 0957-4174.

[6] AKKOÇ, Soner, 2012. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*. **222**(1), 168–178. ISSN 03772217.

[7] ALA'RAJ, Maher a Maysam F. ABBOD, 2016a. A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **64**, 36–55. ISSN 09574174.

[8] ALA'RAJ, Maher a Maysam F. ABBOD, 2016b. Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*. B.m.: Elsevier B.V., **104**, 89–105. ISSN 09507051.

[9] ALEXANDER, Walter, 1989. What's the score? *American Bankers Association, ABA Banking Journal*. **81**(8), 58.

[10] ALTMAN, Edward I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*. B.m.: Wiley/Blackwell (10.1111), **23**(4), 589–609. ISSN 00221082.

[11] AVERY, Robert B., Raphael W. BOSTIC, Paul S. CALEM a Glenn B. CANNER, 2000. Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files. *Real Estate Economics*. B.m.: Wiley/Blackwell (10.1111), **28**(3), 523–547. ISSN 1080-8620.

[12] BAHNSEN, Alejandro Correa, Djamila AOUADA a Björn OTTERSTEN, 2015. Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*. B.m.: Pergamon, **42**(19), 6609–6619. ISSN 0957-4174.

[13] BANASIAK, MJ a GL KIELY, 2000. Predictive collection score technology. *Business Credit-New York*. **102**(2), 18–34.

[14] BANASIK, J, J CROOK a L THOMAS, 2001. Scoring by usage. *Journal of the Operational Research Society*. **52**(9), 997–1006. ISSN 0160-5682.

[15] BANASIK, J, J CROOK a L THOMAS, 2003. Sample selection bias in credit scoring models. *Journal of the Operational Research Society*. **54**(8), 822–832. ISSN 0160-5682.

[16] BAREFOOT, JAS, 1996. Credits scoring at a crossroads. *American Bankers Association, ABA*. **88**(6), 26.

[17] BARUSHKA, Aliaksandr a Petr HAJEK, 2018. Spam Filtering in Social Networks Using Regularized Deep Neural Networks with Ensemble Learning. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. B.m.: Springer, Cham, s. 38–49.
[18] BASTOS, João A., 2010. Forecasting bank loans loss-given-default. *Journal of Banking & Finance*. B.m.: North-Holland, **34**(10), 2510–2517. ISSN 0378-4266.
[19] BEAVER, WH, 1966. Financial ratios as predictors of failure. *Journal of Accounting*. 71–111.
[20] BEN-DAVID, Arie a Eibe FRANK, 2009. Accuracy of machine learning models versus "hand crafted" expert systems – A credit scoring case study. *Expert Systems with Applications*. B.m.: Pergamon, **36**(3), 5264–5271. ISSN 0957-4174.
[21] BERKOWITZ, Jeremy a Richard HYNES, 1999. Bankruptcy Exemptions and the Market for Mortgage Loans. *The Journal of Law and Economics*. B.m.: The University of Chicago Press, **42**(2), 809–830. ISSN 0022-2186.
[22] BIERMAN, Harold a Warren H. HAUSMAN, 1970. The Credit Granting Decision. *Management Science*. B.m.: INFORMS, **16**(8), B-519-B-532. ISSN 0025-1909.
[23] BILGIN, Zeynep a Ugur YAVAS, 1995. Marketing of consumer credit services in a developing country: a status report. *International Journal of Bank Marketing*. B.m.: MCB UP Ltd, **13**(5), 31–36. ISSN 0265-2323.
[24] BISHOP, Christopher M., 1995. *Neural Networks for Pattern Recognition*. B.m.: Oxford university press.
[25] BOYES, William J., Dennis L. HOFFMAN a Stuart A. LOW, 1989. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*. B.m.: North-Holland, **40**(1), 3–14. ISSN 0304-4076.
[26] BREIMAN, Leo, 1996. Bagging predictors. *Machine Learning*. B.m.: Kluwer Academic Publishers, **24**(2), 123–140. ISSN 0885-6125.
[27] BROWN, Iain a Christophe MUES, 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **39**(3), 3446–3453. ISSN 09574174.
[28] CAPON, Noel, 1982. Credit Scoring Systems: A Critical Analysis. *Journal of Marketing*. B.m.: American Marketing Association, **46**(2), 82. ISSN 00222429.
[29] CASELLI, Stefano, Stefano GATTI a Francesca QUERCI, 2008. The Sensitivity of the Loss Given Default Rate to Systematic Risk: New Empirical Evidence on Bank Loans. *Journal of Financial Services Research*. B.m.: Springer US, **34**(1), 1–34. ISSN 0920-8550.
[30] CBCB, 2019. *CBCB - Czech Banking Credit Bureau, a.s.* [online]. Dostupné z: http://www.cbcb.cz/
[31] CHALOS, Peter, 1985. The superior performance of loan review committees. *Journal of Commercial Bank Lending*. **68**, 60–66.
[32] CHAPMAN, P, J CLINTON, R KERBER a T KHABAZA, 2000. *CRISP-DM 1.0 Step-by-step data mining guide* [online]. Dostupné z: https://www.scribd.com/document/264461662/CRISP-DM-1-0-Step-By-Step-Data-Mining-Guide
[33] 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.
[34] CHEN, Mu-Chen a Shih-Hsien HUANG, 2003. Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*. **24**(4), 433–441.

[35] ČNB, 2019. Centrální registr úvěrů. *Česká národní banka* [online]. Dostupné z: www.cnb.cz/cs/dohled_financni_trh/centralni_registr_uveru/index.html?cnb_css=tru

[36] CNCB, 2019. *CNCB - Czech Non-Banking Credit Bureau, a.s.* [online]. Dostupné z: http://www.cncb.cz/

[37] CRONE, Sven F. a Steven FINLAY, 2012. Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*. B.m.: Elsevier B.V., **28**(1), 224–238. ISSN 01692070.

[38] CYERT, R. M., H. J. DAVIDSON a G. L. THOMPSON, 1962. Estimation of the Allowance for Doubtful Accounts by Markov Chains. *Management Science*. B.m.: INFORMS, **8**(3), 287–303. ISSN 0025-1909.

[39] DESAI, M. A., C. F. FOLEY a J. R. HINES, 2004. Foreign direct investment in a world of multiple taxes. *Journal of Public Economics*. B.m.: North-Holland, **88**(12), 2727–2744. ISSN 0047-2727.

[40] DESAI, V. S., J. N. CROOK a G. A. OVERSTREET, 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*. B.m.: North-Holland, **95**(1), 24–37. ISSN 0377-2217.

[41] DIMLA, D. E. a P. M. LISTER, 2000. On-line metal cutting tool condition monitoring.: II: tool-state classification using multi-layer perceptron neural networks. *International Journal of Machine Tools and Manufacture*. B.m.: Pergamon, **40**(5), 769–781. ISSN 0890-6955.

[42] DURAND, David, 1941. *Risk elements in consumer installment financing*. New York: National Bureau of Economic Research.

[43] DZEROSKI, S. a B. ZENKO, 2004. Is combining classfiers with stacking better than selecting the best one? *Machine Learning*. **54**(3), 255–273.

[44] EDMISTER, R. O., 1988. Combining human credit analysis and numerical credit scoring for business failure prediction. *Akron Business and Economic Review*. **19**(3), 6–14.

[45] EKINCI, Aykut a Halil İbrahim ERDAL, 2017. Forecasting Bank Failure: Base Learners, Ensembles and Hybrid Ensembles. *Computational Economics*. B.m.: Springer US, **49**(4), 677–686. ISSN 0927-7099.

[46] ERBAS, BC a SE STEFANOU, 2009. An application of neural networks in microeconomics: Input–output mapping in a power generation subsector of the US electricity industry. *Expert Systems with Applications*. **36**(2), 2317–2326.

[47] FELDMAN, R., 1997. Small business loans, small banks and big change in technology called credit scoring. *The Region*. 19–25.

[48] FINLAY, Steven, 2011. Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*. B.m.: North-Holland, **210**(2), 368–378. ISSN 0377-2217.

[49] FISHER, R. A., 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. **7**(2), 179–188.

[50] FLOREZ-LOPEZ, Raquel a Juan Manuel RAMON-JERONIMO, 2015. Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **42**(13), 5737–5753. ISSN 09574174.

[51] FRAME, W. Scott, Aruna SRINIVASAN a Lynn WOOSLEY, 2001. The effect of credit scoring on small-business lending. *Journal of Money, Credit and Banking*. 813–825.

[52] FRANK, E, M MAYO a S KRAMER, 2015. Alternating model trees. In: *Proceedings of the 30th Annual Symposium on Applied Computing*. s. 871–878.

[53] FRANK, Eibe, Yong WANG, Stuart INGLIS, Geoffrey HOLMES a Ian H. WITTEN, 1998. Using Model Trees for Classification. *Machine Learning*. B.m.: Kluwer Academic Publishers, **32**(1), 63–76. ISSN 08856125.

[54] FREUND, Yoav a Robert E. SCHAPIRE, 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. B.m.: Academic Press, **55**(1), 119–139. ISSN 0022-0000.

[55] FREUND, Yoav a Robert E. SCHAPIRE, 1999. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*. **14**(5), 771–780.

[56] FRIEDMAN, J, T HASTIE a R TIBSHIRANI, 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*. **28**(2), 337–407.

[57] GALAR, Mikel, Alberto FERNÁNDEZ, Edurne BARRENECHEA a Francisco HERRERA, 2013. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*. B.m.: Elsevier, **46**(12), 3460–3471. ISSN 00313203.

[58] GARCÍA, Francisco Javier Población, 2017. *Financial Risk Management: Identification, Measurement and Management*. B.m.: Springer. ISBN 978-3-319-41365-5.

[59] GARCÍA, Salvador, Alberto FERNÁNDEZ, Julián LUENGO a Francisco HERRERA, 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*. B.m.: Elsevier, **180**(10), 2044–2064. ISSN 0020-0255.

[60] GREENE, Wiliam, 1998. Sample selection in credit-scoring models. *Japan & The World Economy*. **10**(3), 299–316.

[61] GÜRTLER, Marc, Martin Thomas HIBBELN a Piet USSELMANN, 2018. Exposure at default modeling – A theoretical and empirical assessment of estimation approaches and parameter choice. *Journal of Banking and Finance*. **91**, 176–188. ISSN 03784266.

[62] HARRIS, Terry, 2015. Credit scoring using the clustered support vector machine. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **42**(2), 741–750. ISSN 09574174.

[63] HE, Hongliang, Wenyu ZHANG a Shuai ZHANG, 2018. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*. B.m.: Pergamon, **98**, 105–117. ISSN 0957-4174.

[64] HENLEY, W. E. a D. J. HAND, 1996. A k-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk. *The Statistician*. B.m.: WileyRoyal Statistical Society, **45**(1), 77–95. ISSN 00390526.

[65] HENS, Akhil Bandhu a Manoj Kumar TIWARI, 2012. Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications*. B.m.: Pergamon, **39**(8), 6774–6781. ISSN 0957-4174.

[66] HSIEH, Nan Chen a Lun Ping HUNG, 2010. A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **37**(1), 534–545. ISSN 09574174.

[67] KIM, Yoon Seong a So Young SOHN, 2004. Managing loan customers using misclassification patterns of credit scoring model. *Expert Systems with Applications*. B.m.: Pergamon, **26**(4), 567–573. ISSN 0957-4174.

[68] KOUTANAEI, Fatemeh Nemati, Hedieh SAJEDI a Mohammad KHANBABAEI, 2015. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*. B.m.: Elsevier, **27**, 11–23. ISSN 09696989.

[69] KRUPPA, Jochen, Alexandra SCHWARZ, Gerhard ARMINGER a Andreas ZIEGLER, 2013. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*. B.m.: Pergamon, **40**(13), 5125–5131. ISSN 0957-4174.

[70] KUNCHEVA, Ludmila I. a Juan J. RODRÍGUEZ, 2007. An Experimental Study on Rotation Forest Ensembles. In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, s. 459–468.

[71] LANE, Sylvia, 1972. Submarginal Credit Risk Classification. *The Journal of Financial and Quantitative Analysis*. B.m.: Cambridge University Press, **7**(1), 1379–1385. ISSN 00221090.

[72] LEE, T. S. a C. C. CHIU, 2002. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*. **23**(3), 245–254.

[73] LEONARD, Kevin J., 1995. The development of credit scoring quality measures for consumer credit applications. *International Journal of Quality & Reliability Management*. **12**(4), 79–85. ISSN 0265-671X.

[74] LEOW, Mindy a Jonathan CROOK, 2016. A new Mixture model for the estimation of credit card Exposure at Default. *European Journal of Operational Research*. B.m.: Elsevier B.V., **249**(2), 487–497. ISSN 03772217.

[75] LESSMANN, Stefan, Bart BAESENS, Hsin Vonn SEOW a Lyn C. THOMAS, 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*. B.m.: Elsevier Ltd., **247**(1), 124–136. ISSN 03772217.

[76] LI, Jianping, Liwei WEI, Gang LI a Weixuan XU, 2011. An evolution strategy-based multiple kernels multi-criteria programming approach: The case of credit decision making. *Decision Support Systems*. B.m.: North-Holland, **51**(2), 292–298. ISSN 0167-9236.

[77] LI, Shukai, Ivor W. TSANG a Narendra S. CHAUDHARI, 2012. Relevance vector machine based infinite decision agent ensemble learning for credit risk analysis. *Expert Systems with Applications*. B.m.: Pergamon, **39**(5), 4947–4953. ISSN 0957-4174.

[78] LOTERMAN, Gert, Iain BROWN, David MARTENS, Christophe MUES a Bart BAESENS, 2012. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*. B.m.: Elsevier B.V., **28**(1), 161–170. ISSN 01692070.

[79] LOUZADA, Francisco, Anderson ARA a Guilherme B. FERNANDES, 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*. B.m.: Elsevier B.V., **21**(2), 117–134. ISSN 18767354.

[80] LUO, Cuicui, Desheng WU a Dexiang WU, 2017. A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*. B.m.: Elsevier Ltd, **65**(September 2016), 465–470. ISSN 09521976.

[81] MALDONADO, Sebastián, Juan PÉREZ a Cristián BRAVO, 2017. Cost-based feature selection for Support Vector Machines: An application in credit scoring. *European Journal of Operational Research*. B.m.: North-Holland, **261**(2), 656–665. ISSN 0377-2217.

[82] MANTAS, Carlos J. a Joaquín ABELLÁN, 2014. Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*. B.m.: Pergamon, **41**(10), 4625–4637. ISSN 0957-4174.

[83] MARQUÉS, A. I., V. GARCÍA a J. S. SÁNCHEZ, 2012a. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*. **39**(11), 10244–10250. ISSN 09574174.

[84] MARQUÉS, A. I., V. GARCÍA a J. S. SÁNCHEZ, 2012b. Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*. B.m.: Pergamon, **39**(12), 10916–10922. ISSN 0957-4174.

[85] MARQUÉS, A. I., V. GARCÍA a J. S. SÁNCHEZ, 2013. On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*. **64**(7), 1060–1070. ISSN 01605682.

[86] MARTENS, D, T VAN GESTEL, M DE BACKER, R HAESEN, J VANTHIENEN a B BAESENS, 2010. Credit rating prediction using Ant Colony Optimization. *Journal of the Operational Research Society*. **61**(4), 561–573. ISSN 0160-5682.

[87] MEHTA, Dileep, 1968. The Formulation of Credit Policy Models. *Management Science*. **15**(2), B-30-B-50. ISSN 0025-1909.

[88] MELVILLE, P. a R. MOONEY, 2003. Constructing diverse classifier ensembles using artificial training examples. *IJCAI International Joint Conference on Artificial Intelligence*. 505–510.

[89] MILERIS, Ricardas, 2010. Estimation of Loan Applicants Default Probability Applying Discriminant Analysis and Simple Bayesian Classifier. *Economics & Management*.

[90] MYERS, James H. a Warren CORDNER, 1957. Increase Credit Operation Profits. *The Credit World*. 12–13.

[91] NAZEMI, Abdolreza, Farnoosh FATEMI POUR, Konstantin HEIDENREICH a Frank J. FABOZZI, 2017. Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research*. B.m.: Elsevier B.V., **262**(2), 780–791. ISSN 03772217.

[92] NGUYEN, Ha Thu, 2015. Default predictors in credit scoring: evidence from France's retail banking institution. *Journal of Credit Risk*. **11**(2), 41–66.

[93] ORGLER, Yair E., 1970. A credit scoring model for commercial loans. *Credit and Banking*. **2**(4), 435–445.

[94] PACÁKOVÁ, V., I. STANKOVIČOVÁ a B. PACÁK, 2005. Application of Logistic Regression in Financial Sector. In: *12th Polish-Slovak-Ukrainian Scientific Seminar organized by the Department of Demography of the Cracow University of Economics*.

[95] PACÁKOVÁ, Viera a Eva RUBLÍKOVÁ, 2000. Štatistické modely v analýzach trhu práce. *Ekonóm*. **1**.

[96] PALEOLOGO, Giuseppe, André ELISSEEFF a Gianluca ANTONINI, 2010. Subagging for credit scoring models. *European Journal of Operational Research*. B.m.: Elsevier B.V., **201**(2), 490–499. ISSN 03772217.

[97] PAPOUŠKOVÁ, Monika a Petr HÁJEK, 2019. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*. **118**, 33–45. ISSN 0167-9236.

[98] PING, Yao a Lu YONGHENG, 2011. Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Systems with Applications*. B.m.: Pergamon, **38**(9), 11300–11304. ISSN 0957-4174.
[99] PUNCH, Linda, 2000. Shedding Light on Credits Scores. *Credit Card Management*. **13**(5), 78.
[100] QUINLAN, J. R., 1992. Learning with continuous classes. *AI '92 World Scientific*. 343–348.
[101] QUINLAN, J. R., 1993. *C4.5: Programs for Machine Learning*.
[102] RAESIDE, Robert a John WALKER, 2001. Knowledge: the Key to Organisational Survival. *The TQM Magazine*. **13**(3), 156–160.
[103] REED, Russell D. a Robert J. MARKS, 1998. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*.
[104] RODRIGUEZ, J.J., L.I. KUNCHEVA a C.J. ALONSO, 2006. Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **28**(10), 1619–1630. ISSN 0162-8828.
[105] ROSENBERG, Eric a Alan GLEIT, 1994. Quantitative Methods in Credit Management: a Survey. *Operetaions Research*. **42**(4), 589–613.
[106] SAKAI, Yasuhiro, 1998. Sample Selection in Credit Scoring Models. *Japan & The World Economy*. **3**(10), 317–320.
[107] SANDLER, Andrew L., Stacie E. MCGINN a Joseph L. BARLOON, 2000. Fair Lending Scrutiny of Credit Score-based Underwriting Systems. *ABA Bank Compliance*. **21**(3), 37–43.
[108] ŠARLIJA, Nataša, Mirta BENŠIĆ a Zoran BOHAČEK, 2004. Multinomial Model in Consumer Credit Scoring. In: *10th International Conference on Operational Research*.
[109] SERRANO-CINCA, Carlos a Begoña GUTIÉRREZ-NIETO, 2016. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*. B.m.: North-Holland, **89**, 113–122. ISSN 0167-9236.
[110] SIDDIQI, Naeem, 2012. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. B.m.: John Wiley & Sons.
[111] SOLUS, 2019. *SOLUS* [online]. Dostupné z: https://www.solus.cz/
[112] STANKOVIČOVÁ, Iveta a Mária VOJTKOVÁ, 2007. *Viacrozmerné štatistické metódy s aplikáciami*. Iura Editi. Bratislava: Edícia Ekonómia. ISBN 978-80-8078-152-1.
[113] STEENACKERS, A. a M. J. GOOVAERTS, 1989. A Credit Scoring Model for Personal Loans. *Mathematics and Economics*. **8**(1), 31–34.
[114] STIGLITZ, Joseph E. a Andrew WEISS, 1983. Incentive Effects of Terminations: Applications to Credit and Labor Markets. *The American Economic Review*. (912–927).
[115] STINE, B. a W. LANG, 2007. *Space-Time Models for Retail Credit* [online]. Dostupné z: http://www-stat.wharton.upenn.edu/~stine/research/spatial-slides.pdf
[116] SUN, Jie, Jie LANG, Hamido FUJITA a Hui LI, 2018. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*. B.m.: Elsevier Inc., **425**, 76–91. ISSN 00200255.
[117] SUNG, Tae Kyung, Namsik CHANG a Gunhee LEE, 1999. Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction. *Journal of Management Information Systems*. **16**(1), 63–85. ISSN 0742-1222.

[118] TEREK, Milan, Adriana HORNÍKOVÁ a Viera LABUDOVÁ, 2010. *Hĺbková analýza údajov*. 1. vyd. Bratislava: Iura Edition, spol. s r. o. ISBN 978-80-8078-336-5.
[119] THOMAS, Lyn C., David B. EDELMAN a Jonathan N. CROOK, 2002. *Credit Scoring and Its Applications*. Philadelphia: SIAM Monographs on mathematical modeling and computation. ISBN 978-0-898714-83-8.
[120] THOMAS, Lyn, Jonathan CROOK a David EDELMAN, 2017. *Credit scoring and its applications*. 2. vyd.
[121] TODOROVSKI, Ljupčo a Sašo DŽEROSKI, 2003. Combining Classifiers with Meta Decision Trees. *Machine Learning*. B.m.: Kluwer Academic Publishers, **50**(3), 223–249. ISSN 08856125.
[122] TOMCZAK, Jakub M. a Maciej ZIEBA, 2015. Classification Restricted Boltzmann Machine for comprehensible credit scoring model. *Expert Systems with Applications*. **42**(4), 1789–1796. ISSN 09574174.
[123] TONG, Edward N.C., Christophe MUES, Iain BROWN a Lyn C. THOMAS, 2016. Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research*. B.m.: Elsevier B.V., **252**(3), 910–920. ISSN 03772217.
[124] TRIPPI, Robert R. a Efraim TURBAN, 1992. *Neural networks in finance and investing: Using artificial intelligence to improve real world performance*. B.m.: McGraw-Hill, Inc.
[125] TSAI, Chih-Fong, 2014. Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*. B.m.: Elsevier, **16**, 46–58. ISSN 1566-2535.
[126] TSAI, Chih-Fong a Chihli HUNG, 2014. Modeling credit scoring using neural network ensembles. *Kybernetes*. **43**(7), 1114–1123. ISSN 0368-492X.
[127] TWALA, Bhekisipho, 2010. Multiple classifier application to credit risk assessment. *Expert Systems with Applications*. B.m.: Pergamon, **37**(4), 3326–3336. ISSN 0957-4174.
[128] VAPNIK, V. a A. C. CHERVONENKIS, 1974. Theory of Pattern Recognition. *Statistical Learning Problems*.
[129] VERBRAKEN, Thomas, Cristián BRAVO, Richard WEBER a Bart BAESENS, 2014. Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*. B.m.: North-Holland, **238**(2), 505–513. ISSN 0377-2217.
[130] VERBRAKEN, Thomas, Wouter VERBEKE a Bart BAESENS, 2013. A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models. *IEEE Transactions on Knowledge and Data Engineering*. **25**(5), 961–973. ISSN 1041-4347.
[131] WANG, Gang, Jinxing HAO, Jian MA a Hongbing JIANG, 2011. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **38**(1), 223–230. ISSN 09574174.
[132] WANG, Gang, Jian MA, Lihua HUANG a Kaiquan XU, 2012. Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*. B.m.: Elsevier B.V., **26**, 61–68. ISSN 09507051.
[133] WANG, Hong, Qingsong XU a Lifeng ZHOU, 2015. Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLoS ONE*. **10**(2), 1–20. ISSN 19326203.

[134] WANG, Shuo Wang Shuo a Xin Yao Xin YAO, 2009. Diversity analysis on imbalanced data sets by using ensemble models. *2009 IEEE Symposium on Computational Intelligence and Data Mining*. 324–331. ISSN 00200255.
[135] WEBB, Geoffrey I., 2000. MultiBoosting: A Technique for Combining Boosting and Wagging. *Machine Learning*. B.m.: Kluwer Academic Publishers, **40**(2), 159–196. ISSN 08856125.
[136] WEBB, Geoffrey I., 2011. MultiBoosting. *Encyclopedia of Machine Learning. Springer*. 699–701.
[137] WEST, David, 2000. Neural network credit scoring models. *Computers & Operations Research*. **27**(11), 1131–1152.
[138] WITKOWSKA, Dorota, 2006. Discrete choice model application to the credit risk evaluation. *International Advances in Economic Research*. **12**(1), 33–42.
[139] WITTEN, Ian H. a Eibe FRANK, 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2. vyd. United States of America: Morgan Kaufmann. ISBN 978-0-12-80491-5.
[140] WITTEN, Ian H., Eibe FRANK a Mark A. HALL, 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. vyd. B.m.: Morgan Kaufmann. ISBN 978-0-12-374856-0.
[141] WOLPERT, David H., 1992. Stacked generalization. *Neural Networks*. B.m.: Pergamon, **5**(2), 241–259. ISSN 0893-6080.
[142] XIA, Yufei, Chuanzhe LIU, Bowen DA a Fangming XIE, 2018. A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*. B.m.: Pergamon, **93**, 182–199. ISSN 0957-4174.
[143] YANG, Bill Huajian a Mykola TKACHENKO, 2012. Modeling exposure at default and loss given default: empirical approaches and technical implementation. *Journal of Credit Risk*. **8**(2), 22. ISSN 17446619.
[144] YAO, Xiao, Jonathan CROOK a Galina ANDREEVA, 2015. Support vector regression for loss given default modelling. *European Journal of Operational Research*. B.m.: Elsevier B.V., **240**(2), 528–538. ISSN 03772217.
[145] YAO, Xiao, Jonathan CROOK a Galina ANDREEVA, 2017. Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*. B.m.: Elsevier B.V., **263**(2), 679–689. ISSN 03772217.
[146] YAP, Bee Wah, Seng Huat ONG a Nor Huselina Mohamed HUSAIN, 2011. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*. B.m.: Pergamon, **38**(10), 13274–13283. ISSN 0957-4174.
[147] YU, Lean, Zebin YANG a Ling TANG, 2016. A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment. *Flexible Services and Manufacturing Journal*. B.m.: Springer US, **28**(4), 576–592. ISSN 19366590.
[148] YU, Lean, Xiao YAO, Shouyang WANG a K.K. LAI, 2011. Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Systems with Applications*. B.m.: Pergamon, **38**(12), 15392–15399. ISSN 0957-4174.
[149] YU, Lean, Wuyi YUE, Shouyang WANG a K. K. LAI, 2010. Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **37**(2), 1351–1360. ISSN 09574174.

[150] ZHANG, Defu, Xiyue ZHOU, Stephen C.H. LEUNG a Jiemin ZHENG, 2010. Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **37**(12), 7838–7843. ISSN 09574174.

[151] ZHOU, Ligang, Kin Keung LAI a Lean YU, 2010. Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*. B.m.: Elsevier Ltd, **37**(1), 127–133. ISSN 09574174.

[152] ZHU, You, Chi XIE, Gang-Jin WANG a Xin-Guo YAN, 2017. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*. B.m.: Springer London, **28**(S1), 41–50. ISSN 0941-0643.

Seznam publikovaných prací

Časopisy

[1] PAPOUŠKOVÁ, Monika a Viera PACÁKOVÁ, 2015. Application of the logistic regression in non-bank financial institutions. *International Journal of Economics and Statistics*. 3, 178–181. ISSN 2309-0685.

[2] PACÁKOVÁ, Viera a Monika PAPOUŠKOVÁ, 2016. Multidimensional comparisons of health systems functioning in OECD countries. *International Journal of Mathematical Models and Methods in Applied Sciences*. 10(12), 388–394. SJR: 0.122.

[3] PAPOUŠKOVÁ, Monika a Petr HÁJEK, 2019. Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*. 118, 33–45. ISSN 0167-9236. IF: 3.565.

Sborníky z konferencí

[4] PAPOUŠKOVÁ, Monika, 2013. Economic scenario generators and solvency II. In: *Financial Management of Firms and Financial Institutions: The 9th International Scientific Conference Proceedings*, PTS I-III. s. 658–664. ISBN 978-80-248-3172-5.

[5] PAPOUŠKOVÁ, Monika, 2014a. The possibility of using rough sets theory in insurance industry. In: *Scientia iuvenis - Book of Scientific Papers*. s. 407–411. ISBN 978-80-558-0650-1.

[6] PAPOUŠKOVÁ, Monika, 2014b. Využití analýzy přežití při řízení kreditního rizika. In: *Managing and Modelling of Financial Risks: The 7th International Scientific Conference*. s. 606–612. ISBN 978-80-248-3631-7.

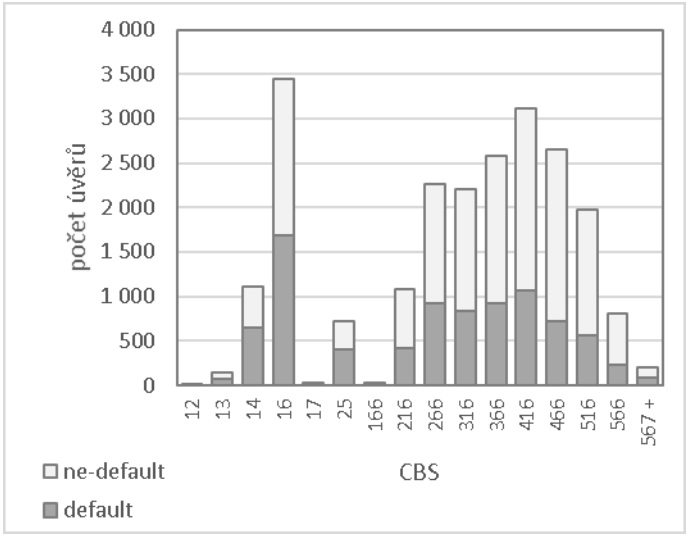
[7] PAPOUŠKOVÁ, Monika, 2015a. Application of the empirical Bayesian credibility models in motor third-party liability insurance. In: *Financial Management of Firms and Financial Institutions: The 10th International Scientific Conference*. s. 942–949. ISBN 978-80-248-3865-6.

[8] PAPOUŠKOVÁ, Monika, 2015b. Survival analysis in portfolio monitoring in non-bank financial institutions. In: *ICTIC - Conference of Informatics and Management Sciences*. s. 25–28. ISBN 978-80-554-1002-9.

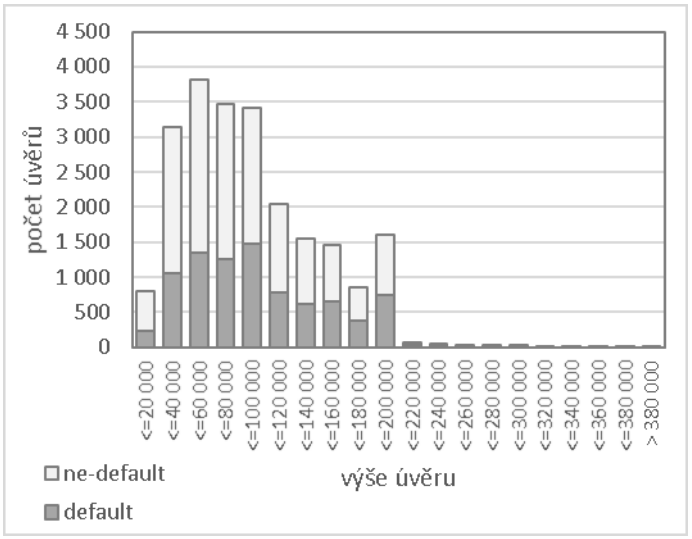
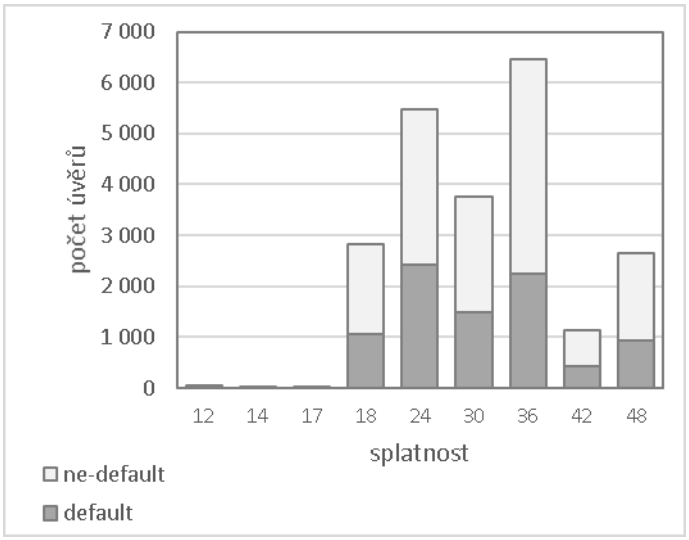
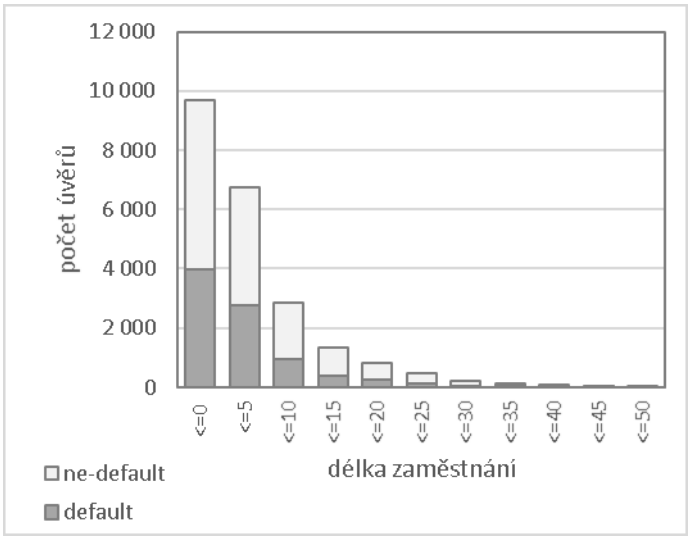
[9] JINDROVÁ, Pavla a Monika PAPOUŠKOVÁ, 2016. Modelling insured catastrophe losses. In: *The 10th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*. s. 53–61. ISBN 978-83-65173-47-8.

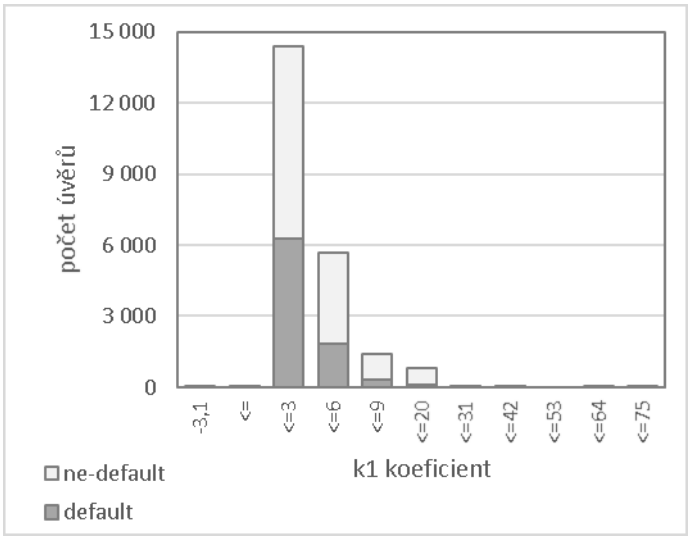
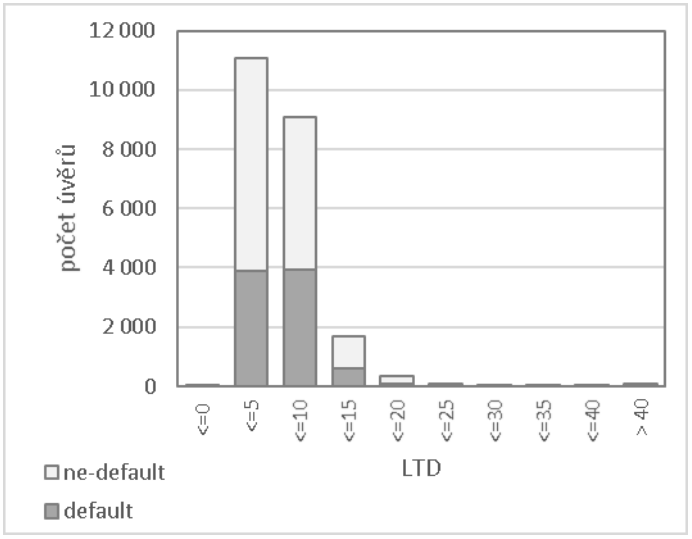
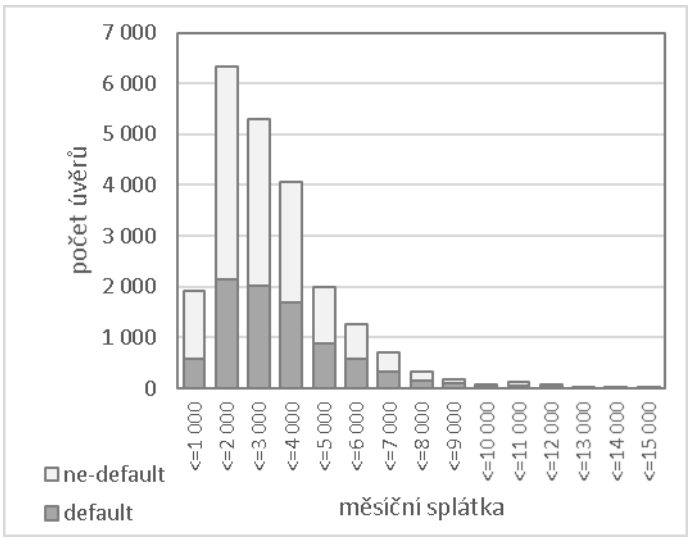
[10] PAPOUŠKOVÁ, Monika a Petr HÁJEK, 2019. Modelling loss given default in peer-to-peer lending using random forest. In: *Intelligent Decision Technologies 2019 – Proceedings of the 11th KES International Conference on Intelligent Decision Technologies (KES-IDT-19).* ISSN 2190-3018. (v tisku)

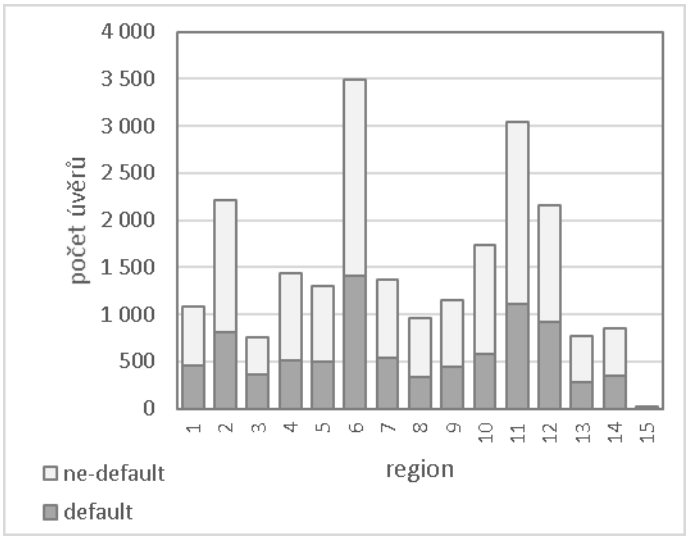
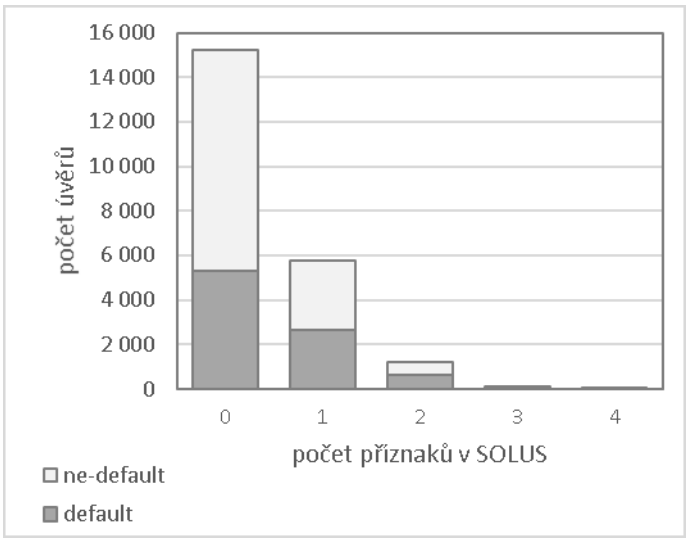
Příloha: Základní informace o souboru dat

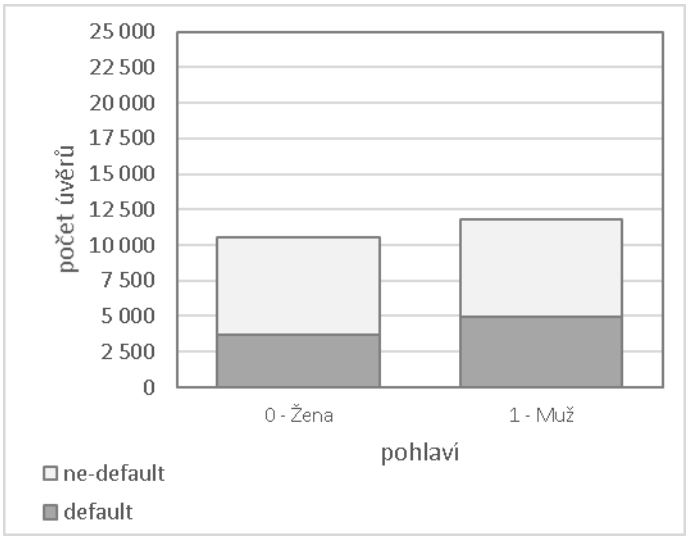
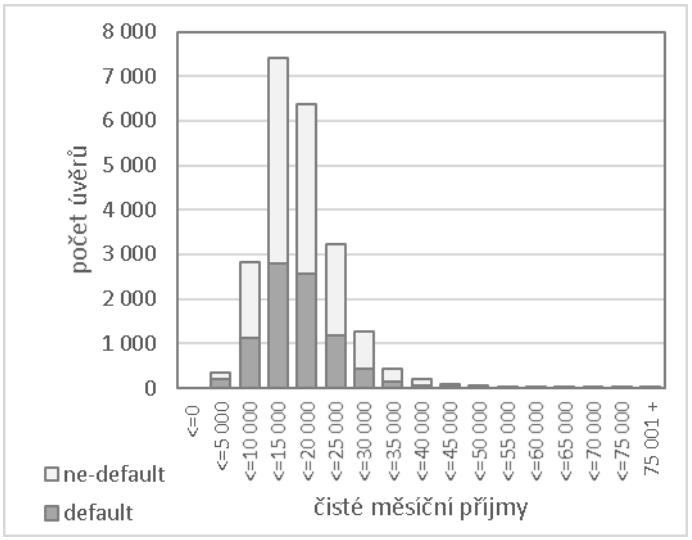
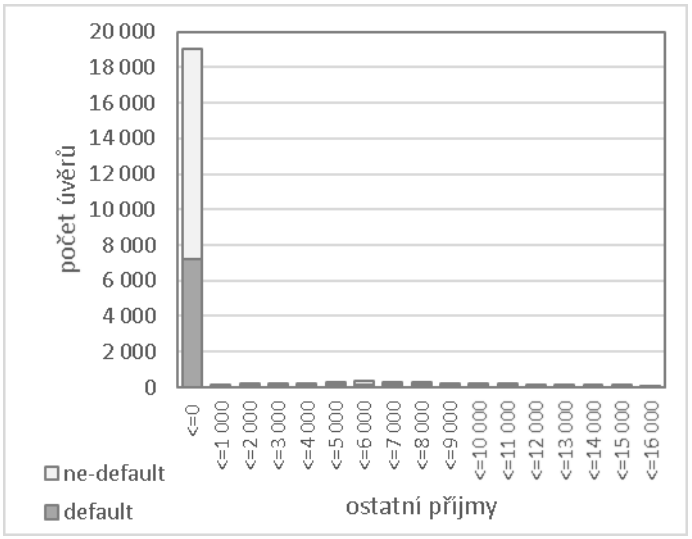
Atribut	Histogram
CBS	

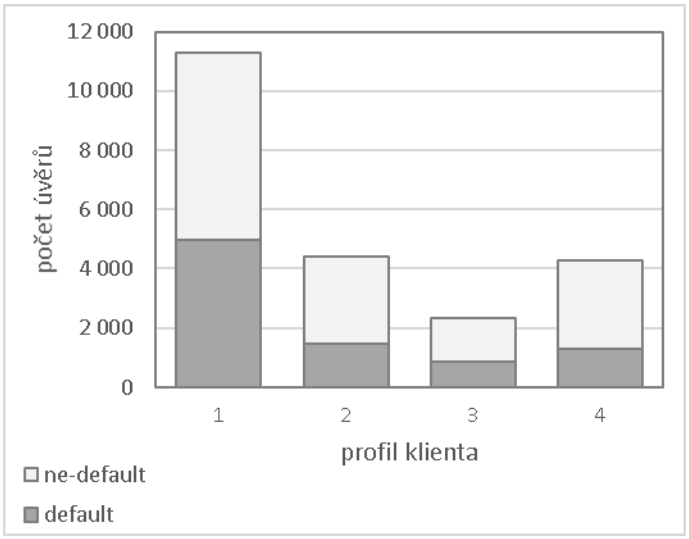
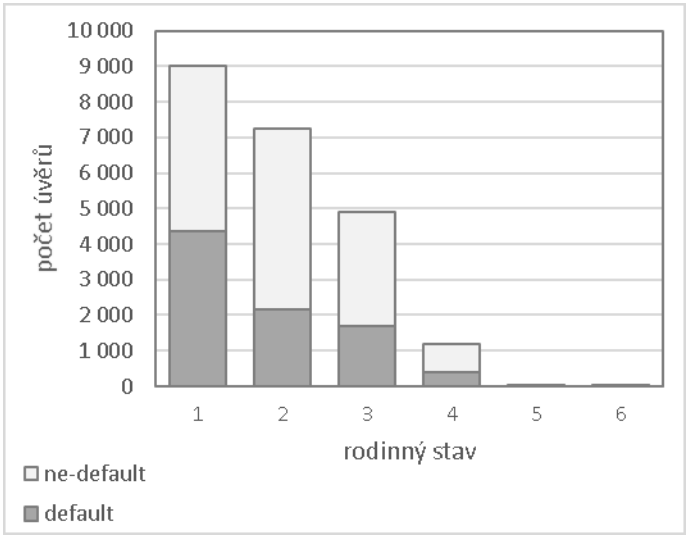
12	Jednotlivec má pouze smlouvy s klientem jako ručitel.
13	Jednotlivec má pouze ukončené smlouvy s klientem (ukončené řádně nebo předčasně) dříve než před 36 měsíci.
14	Příliš nový jednotlivec - jednotlivec má pouze smlouvy s klientem otevřené za poslední 3 měsíce nebo pouze smlouvy, které se zpracovávají; jsou odmítnuté nebo se odvolávají.
16	Smlouva s jednotlivcem je hodnocena jako špatná.
17	Nepovolený debet na běžném účtu - jednotlivec má pouze schválené smlouvy nepovoleného debetu na běžném účtu v roli žadatel nebo spolužadatel.
25	Žádná smlouva - klient nemá žádné úvěrové čerpání uvedené v úvěrové zprávě z bankovního registru klientských informací a/nebo nebankovního registru klientských informací.
166 +	Ohodnocení klienta, jehož bonita se zvyšuje s rostoucím počtem bodů CBS.

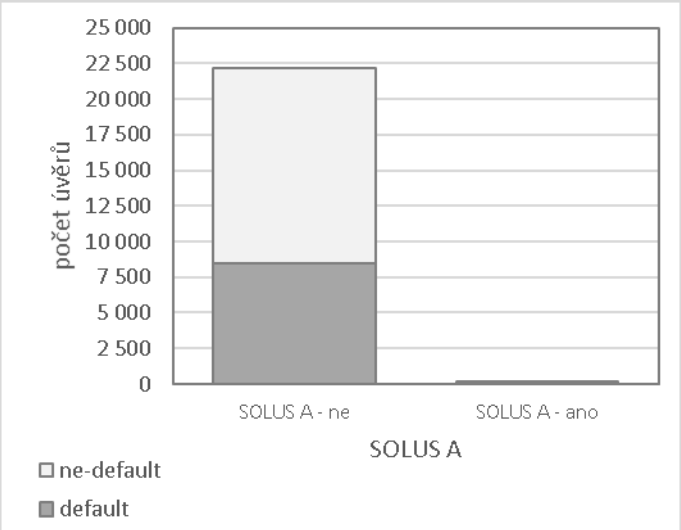
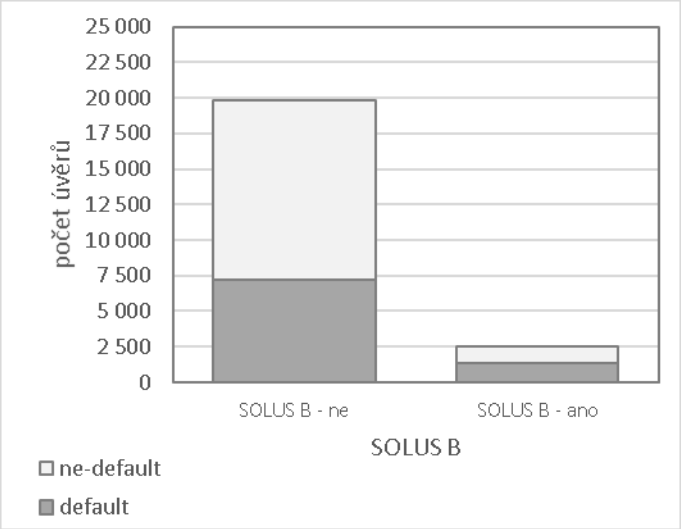
Výše úvěru	
Splatnost (v měsících)	
Délka zaměstnání (v letech)	

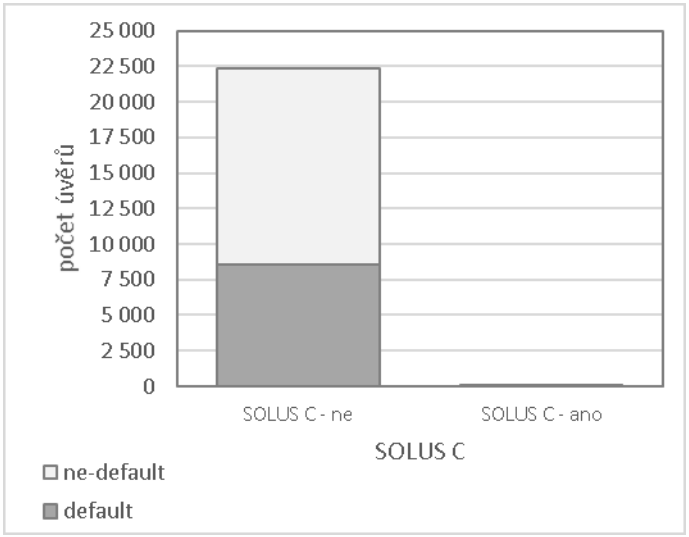
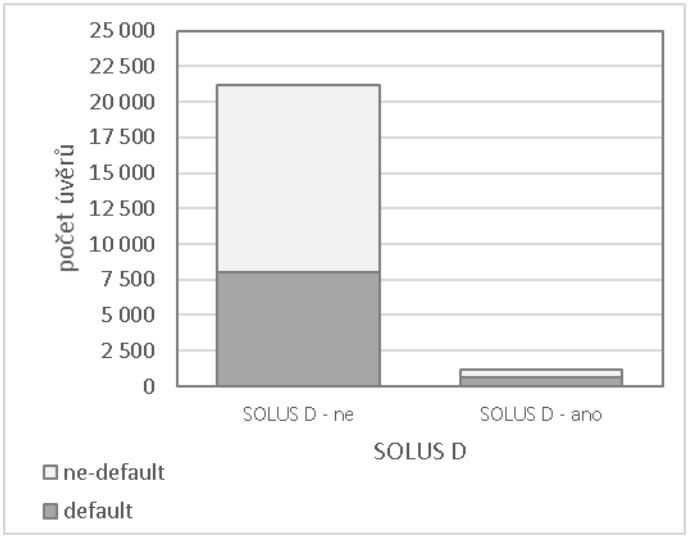
K1 koeficient	
LTD	LTD (loan to disposable income) = vyplacená částka / volné zdroje (Kč) 
Měsíční splátka	

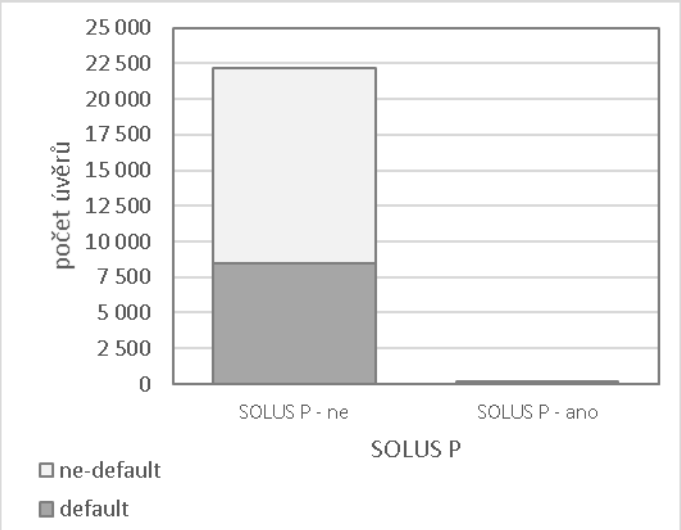
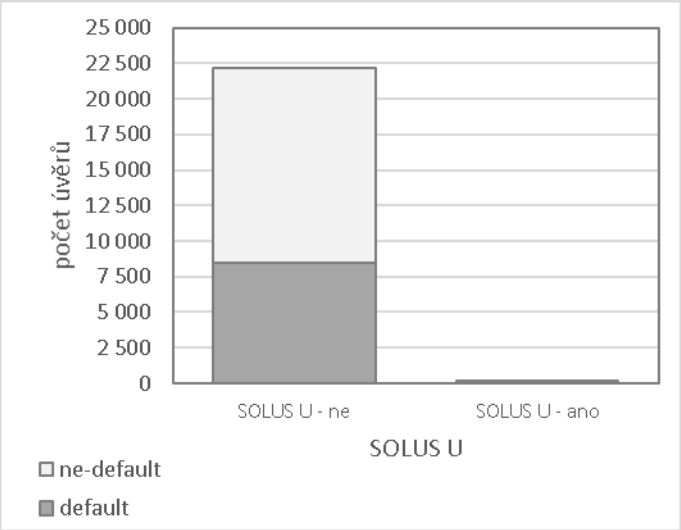
Region	 1 Jihočeský 2 Jihomoravský 3 Karlovarský 4 Královéhradecký 5 Liberecký 6 Moravskoslezský 7 Olomoucký 8 Pardubický 9 Plzeňský 10 Praha 11 Středočeský 12 Ústecký 13 Vysočina 14 Zlínský 15 Slovensko
Počet příznaků v SOLUS	

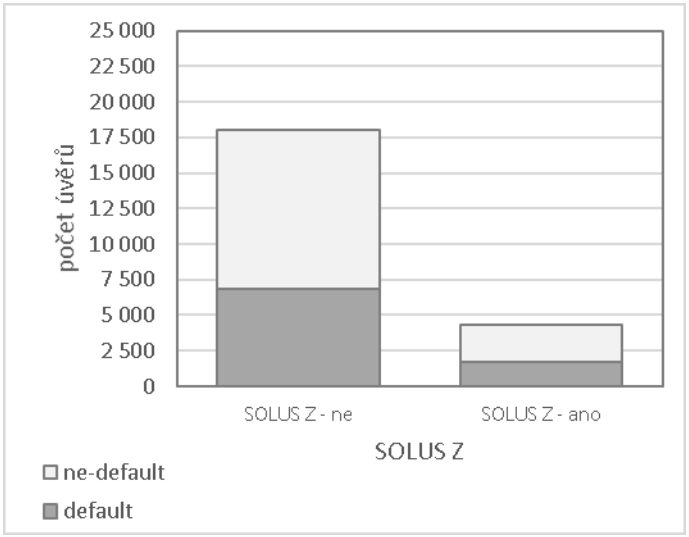
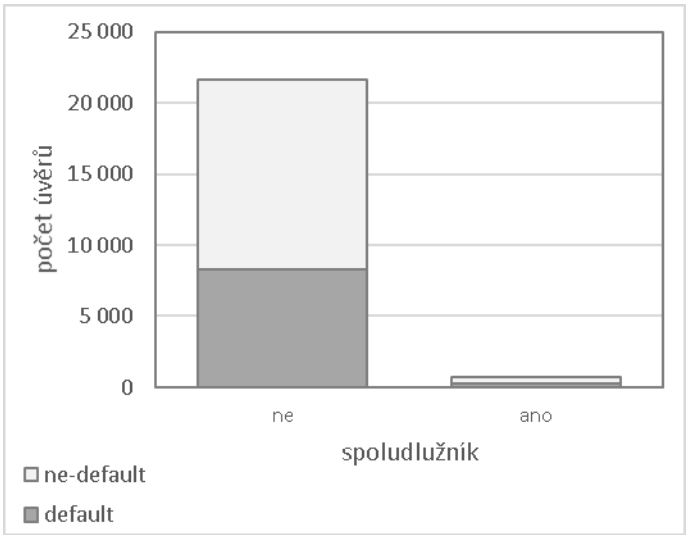
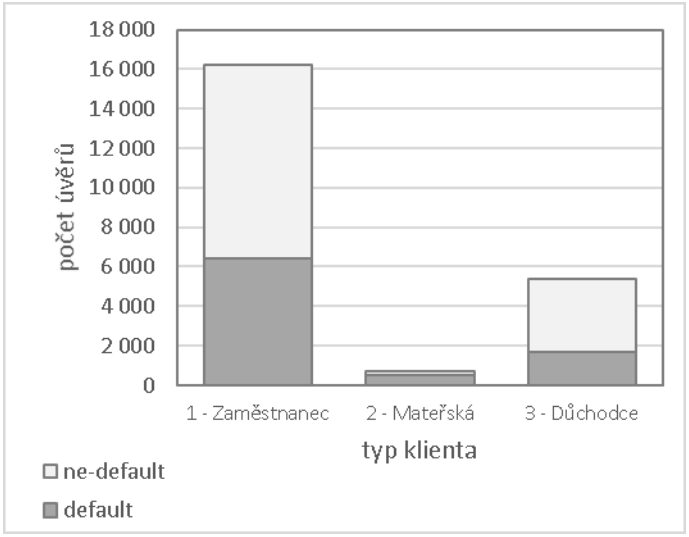
Pohlaví	
Čisté měsíční příjmy	
Ostatní příjmy	

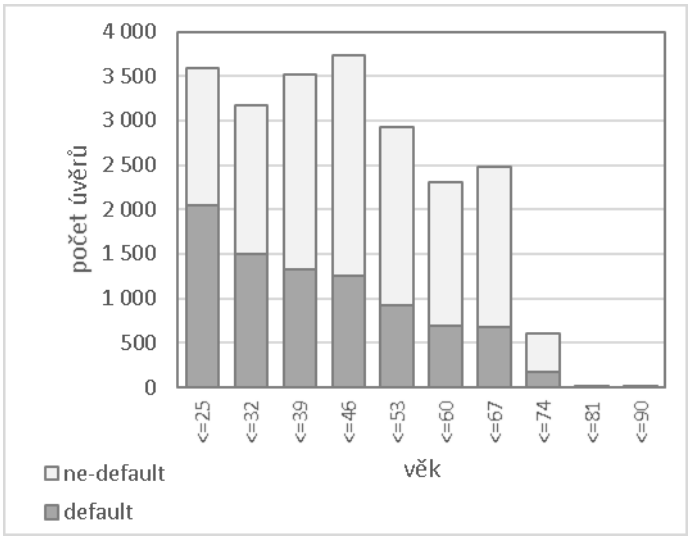
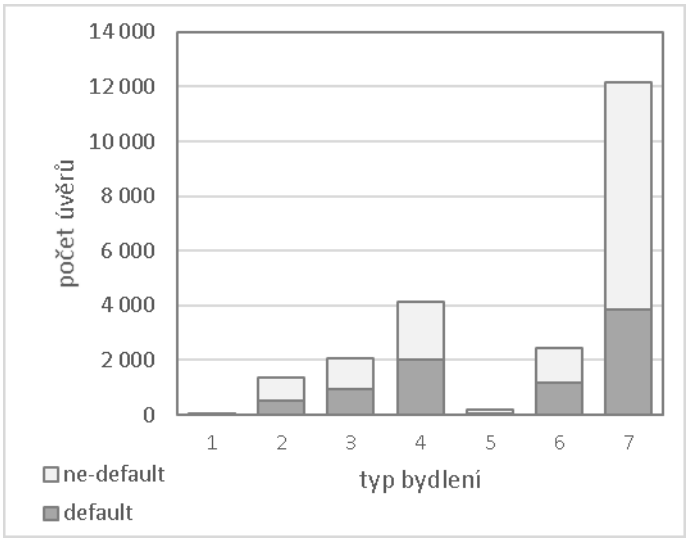
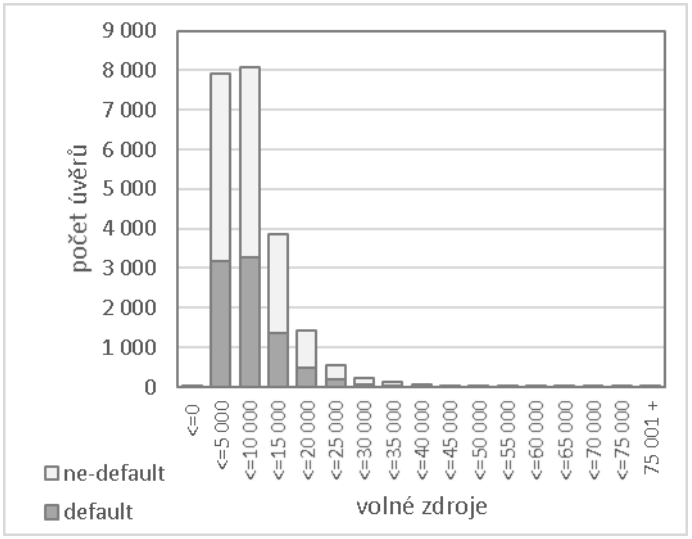
Profil klienta	 1 Nový žadatel 3 Souběžný úvěr 2 Přeúvěrování již existujícího úvěru 4 Vracející se žadatel
Rodinný stav	 1 Svobodný / svobodná 4 Vdovec / vdova 2 Ženatý / vdaná 5 Druh / družka 3 Rozvedený / rozvedená 6 Registrované partnerství

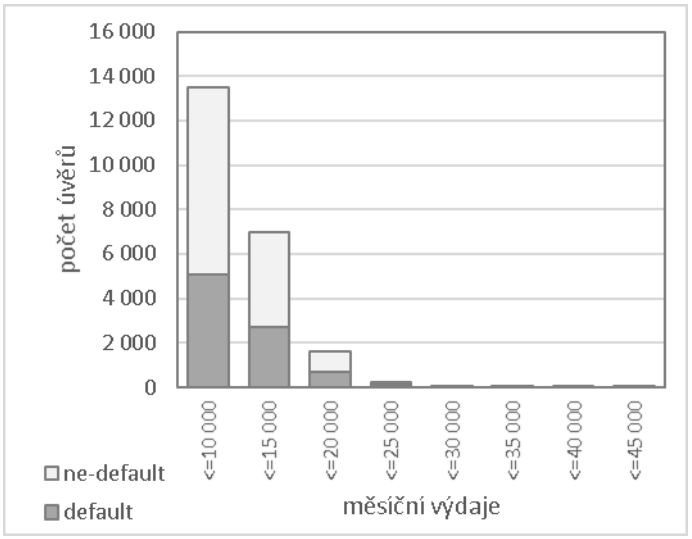
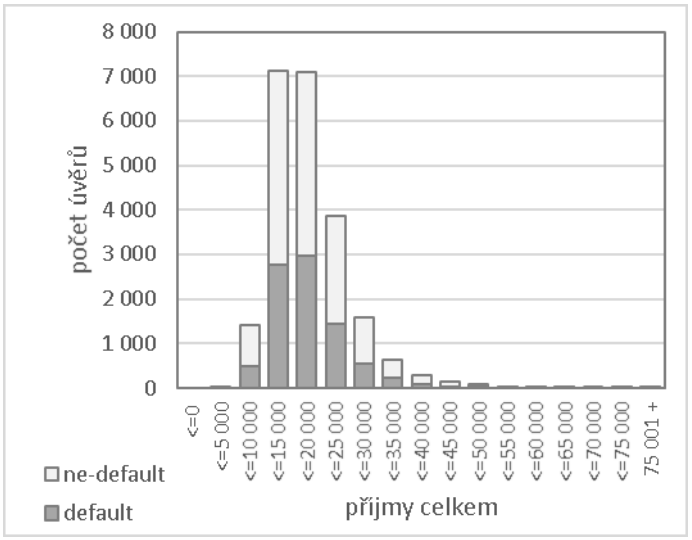
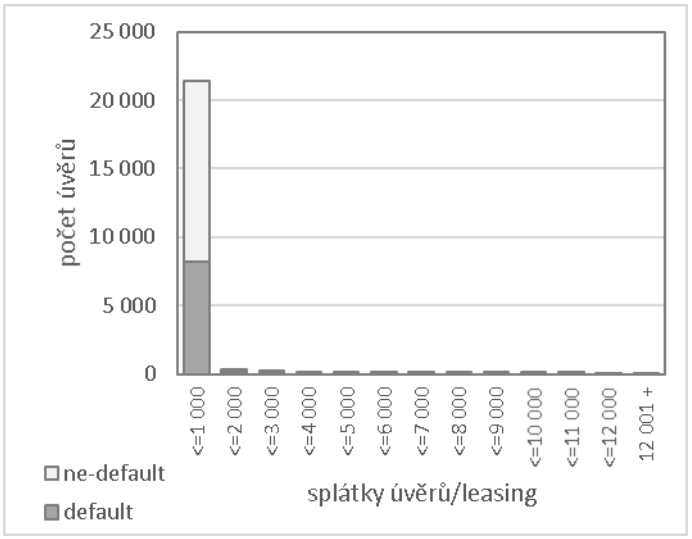
Příznak v SOLUS A	Klient dluží první dvě splátky, resp. první dvě sjednaná finanční plnění. 
Příznak v SOLUS B	Klient dluží někdy v průběhu splácení čtyři splátky za sebou. 

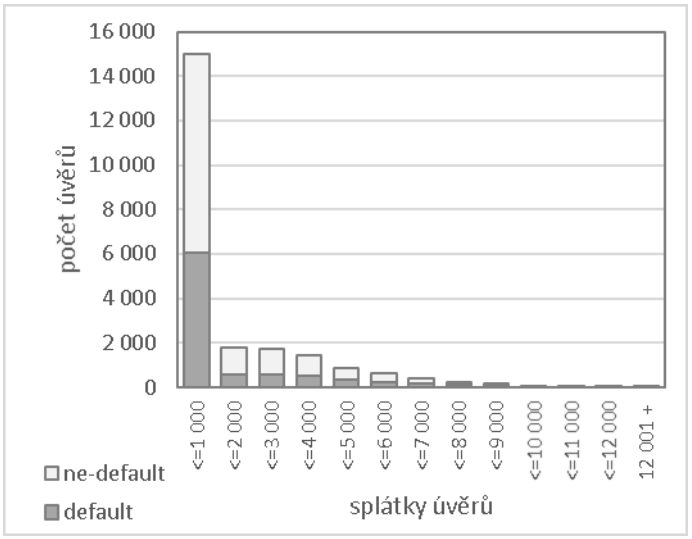
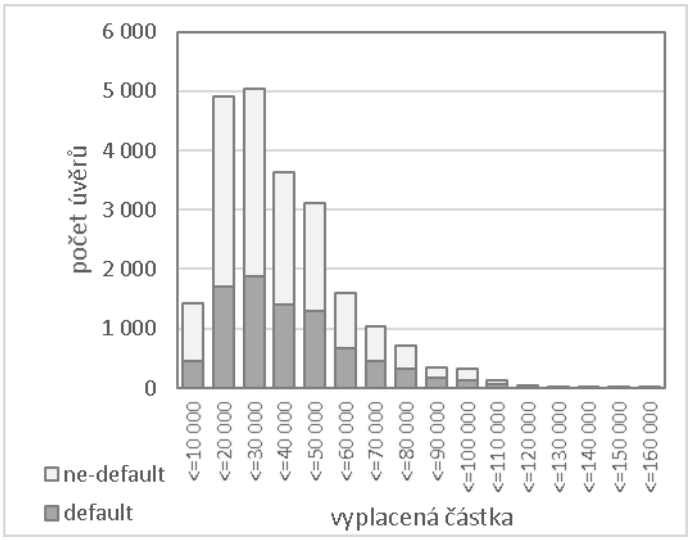
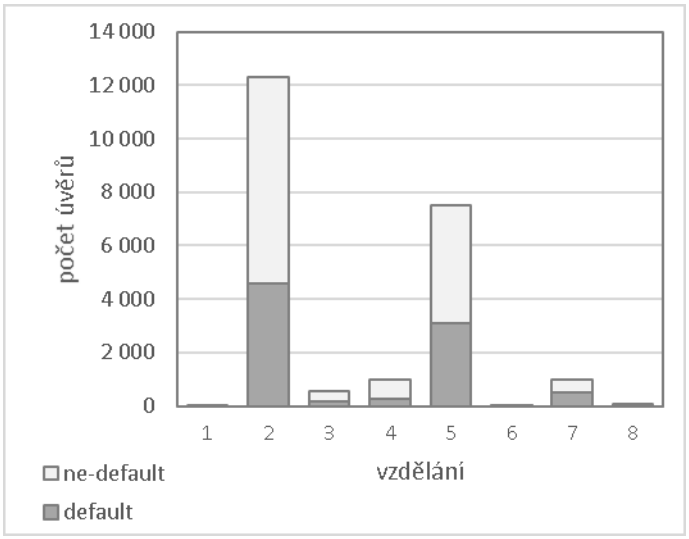
Příznak v SOLUS C	Tento příznak označuje „zaplaceno“ a je použit v případě, že věřitel již nepovažuje zařazený subjekt v registru za klienta v prodlení – tj. úhrada dlužné částky po splatnosti v plné výši nebo dlužná částka po splatnosti klesla pod hranici, která je aktivně vymáhána. 
Příznak v SOLUS D	Tento příznak označuje „zesplatněno“ a bude použit v případě zesplatnění pohledávky. 

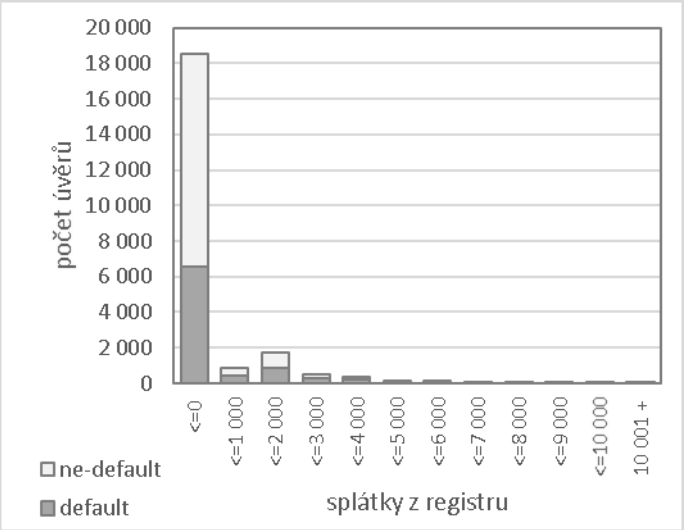
Příznak v SOLUS P	Tento příznak označuje „prodej“ a bude přiřazen v případě, že věřitel již nemá možnost aktualizovat pohledávku z důvodu jejího prodeje. 
Příznak v SOLUS U	Tento příznak označuje „ukončeno“ a bude použit v případě odpisu pohledávky. 

Příznak v SOLUS Z	Tento příznak označuje „zaplaceno“ a bude použit v případě, že věřitel již nepovažuje zařazený subjekt v registru za klienta v prodlení – tj. úhrada dlužné částky po splatnosti v plné výši nebo dlužná částka po splatnosti klesla pod hranici, která je aktivně vymáhána. 
Spoludlužník na smlouvě	
Typ klienta	

Věk klienta	
Typ bydlení	 1 Byt v OV 2 Družstevní byt 3 Nájem – státní byt 4 Nájem od soukr. os. 5 Ostatní 6 U rodičů 7 Vlastní
Volné zdroje	

Měsíční výdaje	
Příjmy celkem	
Splátky úvěrů/leasing	

Splátky úvěrů	
Vyplacená částka	
Vzdělání	 1 Neúplné SŠ 2 Střední vzdělání 3 VOŠ 4 VŠ 5 Vyučení 6 Vyučení s maturitou 7 ZŠ 8 Neuvedeno

Výše splátek z registru SOLUS	
EAD