UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

MODELOVANIE KREDITNÉHO ZLYHANIA PRI KRÁTKODOBÝCH RETAILOVÝCH ÚVEROCH

DIPLOMOVÁ PRÁCA

2018 Bc. Nikolas MÁRKUS

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

MODELOVANIE KREDITNÉHO ZLYHANIA PRI KRÁTKODOBÝCH RETAILOVÝCH ÚVEROCH

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. RNDr. Ján Boďa, CSc.

Bratislava 2018 **Bc. Nikolas MÁRKUS**

97591504

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Nikolas Márkus
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Modelovanie kreditného zlyhania pri krátkodobých retailových úveroch
Credit default modeling of short-term retail loans

Anotácia: Cieľom práce je opis, analýza a porovnanie niekoľkých prístupov, ktoré sa používajú pri odhade kreditného zlyhania podľa finančného štandardu IFRS 9. Implementované metódy následne aplikovať na reálne údaje slovenských klientov.

Vedúci: doc. RNDr. Ján Boďa, CSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Dátum zadania: 26.01.2017

Dátum schválenia: 27.01.2017 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou sa chcem veľmi pekne poďakovať svojmu vedúcemu diplomovej práce doc. RNDr. Jánovi Boďovi, CSc. za ochotu a pomoc pri vedení diplomovej práce. Ďalej sa chcem poďakovať mojim kolegom: Majovi, Maťovi a Janke za pomoc pri príprave podkladových dát, konzultácii ohľadne finančného štandardu IFRS 9, za odporúčanú literatúru a odborné rady, ktoré mi pomohli pri písaní tejto práce. Ďakujem kolegom aj za vytvorenie vhodných pracovných podmienok na tvorbu mojej diplomovej práce.

Abstrakt

MÁRKUS, Nikolas: Modelovanie kreditného zlyhania pri krátkodobých retailových úveroch [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Ján Boďa, CSc., Bratislava, 2018, 83s.

Táto diplomová práca sa zaoberá modelovaním kreditného rizika pri krátkodobých retailových úveroch poskytnutých slovenským klientom. Prvou časťou práce je predstavenie a porovnanie klasifikačných metód, ako sú zovšeobecnené lineárne modely a klasifikačné stromy spolu s náhodnými lesmi, pomocou ktorých skúmame možnosti modelovania úverového zlyhania. Obsahuje podrobný popis procesu hľadania optimálnych modelov a metódy selekcie signifikantných premenných, ktoré podľa modelov najviac vplývajú na kreditné správanie klientov. Druhou časťou práce je modelovanie kreditného rizika na už schválených úveroch spolu s porovnaním metód na výpočet celoživotných pravdepodobností zlyhania na úveroch v súlade s novým štandardom IFRS 9. Tento nový štandard nadobudol platnosť k 1.1.2018 a mení metodiku výpočtu opravných položiek pre aktíva. Hlavným výsledkom práce je porovnanie rôznych štatistických metód v oboch častiach a nájdenie charakteristických čŕt, ktoré vplývajú na kreditné zlyhanie úverov.

Kľúčové slová: Markovove reťazce, zovšeobecnené lineárne modely, klasifikačné stromy a náhodné lesy, kreditné riziko, pravdepodobnosť zlyhania, IFRS 9.

Abstract

MÁRKUS, Nikolas: Credit default modeling of short-term retail loans [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Ján Bod'a, CSc., Bratislava, 2018, 83p.

This master's thesis deals with modelling of credit risk for short-term retail loans provided to Slovak clients. First part of the thesis is an introduction and comparison of classification methods such as general linear models and classification trees along with random forests, by which we examine options of modelling a credit default. Detailed description of the process for finding optimal models and selection methods looking for significant variables which have most impact on client's credit behavior is included as well. The second part of the thesis is the modelling of credit risk on already approved loans together with a comparison of methods for calculating lifetime probabilities of default on loans in accordance with the new IFRS 9 standard. This new standard was enforced on 1. January 2018 and changes the methodology for the calculation of expected loss on credit assets. The main result of the thesis is the comparison of the different statistical methods in both parts and the identification of characteristic features affecting the credit default of loans.

Keywords: Markov chain, generalized linear models, classification trees and random forrests, credit risk, probability of default, IFRS 9.

Obsah

Úvod

Finančné inštitúcie sú vystavené rôznym typom rizík, avšak najvýznamnejšie riziko vzhľadom k objemu úverov a hypoték je úverové resp. kreditné riziko. Banky čelia riziku, že dlžníci nebudú schopný resp. nebudú ochotný splácať svoje záväzky a tým vytvoria pre banku stratu. Odhadovanie kreditného rizika a zlepšovanie skóringových modelov je pre banky nevyhnutné. Tieto modely sa musia frekventovane kalibrovať, aby dokázali čo najlepšie rozoznať žiadosti o úver od klientov s nižšou kredibilitou a z toho vyplývajúce vyššie kreditné riziko.

Kvantifikovanie kreditného rizika je nevyhnutnou súčasťou každého odboru riadenia kreditných rizík, nakoľko straty vzniknuté práve z kreditného rizika sa v účtovníctve vykazujú pomocou tzv. opravných položiek, ktoré znižujú hodnoty pohľadávok. Ku dňu 1.1.2018 nadobudol platnosť nový účtovný štandard *IFRS 9*, ktorý nahrádza štandard *IAS 39* a dôsledkom prechodu na nový štandard sa mení metodika výpočtu opravných položiek. Hlavnou zmenou je prechod z modelu *vzniknutej straty* na model *očakávanej straty*. V dôsledku tohto prechodu na nový štandard musia banky implementovať nové matematicko-štatistické modely. Keďže tieto modely nie sú presne dané, úlohou bánk je vymyslieť také modely, aby optimalizovali svoje výnosovo-rizikové pomery v závislosti od portfólia klientov na ktorých sú zameraní.

Cieľom práce je opis, analýza a porovnanie niekoľkých prístupov, ktoré sa používajú pri odhade kreditného zlyhania podľa finančného štandardu *IFRS 9*. Implementované metódy následne aplikovať na reálne údaje slovenských klientov.

Prácu sme rozdelili do štyroch kapitol, kde prvé dve kapitoly sú teoretické. V teoretickej časti si postupne uvedieme celý matematický aparát potrebný na ďalšie využitie v práci. Uvedieme základnú teóriu a modely zovšeobecnených lineárnych modelov [1, 5, 11]. Ďalej predstavíme možnosti využitia regularizačných metód pri klasifikácii kreditného zlyhania v prípade viacerých vstupných premenných, ktoré majú vplyv na kreditné správanie klientov. V ďalšej kapitole charakterizujeme rôzne prístupy modelovania kreditného zlyhania pomocou klasifikačných stromov a náhodných lesov [12, 13, 15].

Praktická tretia kapitola je venovaná problematike klasifikácii kreditného rizika pri schvaľovaní a oceňovaní úverov. Ďalej sa zaoberá opisu dát a procesu vytvárania a porovnávania rôznych klasifikačných modelov. Kapitola obsahuje výsledky z porovnania jednotlivých klasifikačných a selekčných metód z pohľadu presnosti klasifikácie. Spomenuté sú aj výhody a nevýhody jednotlivých metód.

Posledná štvrtá kapitola sa venuje modelovaniu kreditného zlyhania na už schválených úveroch. V kapitole sú podrobne popísané rôzne metódy odhadu 12 mesačných a celoživotných pravdepodobností zlyhania podľa nového štandardu *IFRS 9*. Pri odhade využívame najmä Markovove reťazce [8] a čiastočne aj analýzu prežívania [3, 7]. Pomocou týchto metód sledujeme zmeny ratingových skúpin daných klientov a z nich analyzujeme možnosti odhadu pravdepodobnosti zlyhania pre dlhšie časové obdobia.

Na analýzu dát a implementovanie vlastných modelov sme použili voľne dostupný štatistický softvér *R* [20] s rôznymi doplnkovými balíkmi, najmä [18, 21, 23].

1 Zovšeobecnené lineárne modely

V tejto kapitole sa zameriame na modelovanie pomocou zovšeobecnených lineárnych modelov, tiež známych ako *GLM* z angl. *generalized linear models*. Na začiatku si stručne zhrnieme základy teórie *GLM* modelov. Následne porovnáme niekoľko typov modelov a metódy na redukciu počtu vysvetľujúcich premenných. Téoriu sme čerpali najmä z [1, 5].

1.1 Teória GLM modelov

Zovšeobecnené lineárne modely sú trieda modelov vďaka ktorým vieme modelovať aj kategoriálne premenné. Vysvetľujúce premenné, pomocou ktorých modelujeme vysvetľované premenné, môžu byť teda nie len spojité, ale aj kategoriálne premenné. Trieda *GLM* modelov sa využíva v pokročilejších štatistických analýzach, kde sa riešia zovšeobecnené situácie oproti jednoduchým lineárnym modelom:

- Vysvetľovaná premenná má iné rozdelenie ako je normálne, prípadne može byť aj kategoriálna premenná, t. j. nemusí byť spojitá.
- Závislosť medzi vysvetľujúcimi a vysvetľovanými premennými nemusí byť lineárna.

Najznámejším modelom z triedy *GLM* je základný lineárny model:

$$E(Y_i) = \mu_i = x_i^T \beta; \qquad Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \tag{1}$$

kde Y_i sú nezávislé náhodné premenné a transponovaný vektor x_i^T reprezentuje i-ty riadok matice plánu X.

Pri zovšeobecnených modeloch sa modeluje transformovaná stredná hodnota vysvetľovanej premennej Y:

$$g(\mu_i) = x_i^T \beta, \tag{2}$$

kde g je monotónna, diferencovateľná funkcia nazývaná ako *spojovacia funkcia* (angl. *link function*). V jednoduchom lineárnom modeli z rovnice (1) je transformácia vo forme $g(\mu_i) = \mu_i$.

Ak by sme to zhrnuli, GLM modely majú tri zložky:

1. Vysvetľované premenné $Y_1, \ldots, Y_N$, pre ktoré predpokladáme, že majú rovnaké pravdepodobnostné rozdelenie z exponenciálnej triedy rozdelení.

2. Sada parametrov β a vysvetľujúcich premenných

$$X = \begin{pmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{Np} \end{pmatrix},$$

3. Monotónna spojovacia funckia $g(\cdot)$:

$$g(\mu_i) = x_i^T \beta; \qquad \mu_i = E(Y_i).$$

1.2 Logistická regresia

Logistická regresia je najznámejší a najpoužívanejší GLM model na modelovanie kategoriálnych dát. Je zároveň aj najpoužívanejší model pri credit scoringu. Mnoho kategoriálnych závislých premenných je binárnych (t. j. nadobúdajú len dve možné kategórie). V práci sa venujeme modelovaniu binárnej premennej $Y = \{\text{default, non-default}\}$ pomocou vysvetľujúcich premenných X. Vysvetľujúce premenné X pozostávajú z dvoch častí. Prvá časť predstavuje kreditné vlastnosti potenciálnych klientov, ktorí žiadajú o úver. Druhá časť charakterizuje zmluvné podmienky žiadaného úveru, ako sú napríklad : výška úveru, mesačná splátka, dĺžka úveru a podobne. Každé pozorovanie (žiadosť o úver) chceme klasifikovať ako schválený, resp. zamietnutý úver. Týmto klasifikáciam priradíme hodnoty 0 a 1, pričom budeme modelovať pravdepodobnosť nastania javu, že žiadateľ o úver zlyhá a nebude plniť zmluvné podmienky voči banke.

Bernoulliho rozdelenie binárnej náhodnej premennej priraďuje pravdepodobnosť $P(Y = 1) = 1 - P(Y = 0) = p$, kde $p = E(Y) = \mu$.

V logistickej regresii pracujeme s podmienenými pravdepodobnosťami. Nech $p(x) = P(Y = 1 \mid X = x) = 1 - P(Y = 0 \mid X = x)$ pre binárne závislú premennú

Y a vysvetľujúce parametre X. V tejto práci predstavujú vysvetľujúce premenné X charakteristické údaje o úveroch a žiadateľov o úver. Logistická regresia predpokladá nelineárnu závislosť medzi $p(x)$ a x, vyjadrenú pomocou monotónnej logistickej funkcie:

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}. \tag{3}$$

Pomocou logistickej funkcie (3) vieme nájsť tvar funkcie, ktorá bude spĺňať vlastnosti GLM modelu, konkrétne nájdeme spojovaciu funkciu pre logistickú regresiu, ktorá spĺňa (2). Najskôr si však definujeme pojem *šanca* (angl. ODD), ktorý nám pomôže pri interpretácii odhadnutých parametrov logistickej regresie.

$$ODD = \frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p). \tag{4}$$

Zo vzťahu (4) po úprave dostávame spojovaciu funkciu pre logistickú regresiu:

$$logit(p(x)) = \ln\ ODD = \ln \frac{p(x)}{1 - p(x)} = \beta^T X = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{5}$$

1.3 Probit

V prípade logistickej regresie s jednou premennou x a $\beta > 0$, je logistická funkcia $p(x)$ z (3) rastúca funkcia a pripomína tvar distribučnej funkcie spojitej náhodnej premennej. To naznačuje, že pre modely s binárnou odozvou môže mať tvar $p(x) = F(x)$, pre nejakú distribučnú funkciu F. Využitím celej škály týchto distribučných funkcií, môžeme posúvať resp. tvarovať túto funkciu tak, aby fitovala čo najlepšie binárnu odozvu. Často sa používa práve distribučná funkcia normálneho rozdelenia $\mathcal{N}(\mu, \sigma^2)$. Nech $\Phi(\cdot)$ označuje distribučnú funkciu štandardizovaného normálneho rozdelenia $\mathcal{N}(0, 1)$. Použitím Φ, dostávame nasledovný model:

$$p(x) = \Phi(\beta^T X). \tag{6}$$

V modeli (6) sa tvar pravdepodobnostnej funkcie mení v závislosti od voľby parametrov β. V prípade, že Φ je rýdzo rastúca, existuje k nej inverzná funkcia a rovnicu (6) vieme ekvivalentne upraviť do tvaru s *probit* spojovacou funkciou (2):

$$probit(p(x)) = \Phi^{-1}(p(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{7}$$

Nevýhodou *probit* modelu je ťažšia interpretácia odhadnutých koeficientov.

1.4 Regularizačné metódy

Rozšírením GLM modelov sú $LASSO$ (Least Absolute Shrinkage and Selection Operator) a *ridge* regresia (tiež známa ako hrebeňová regresia).

Tieto dve metódy patria medzi regularizačné metódy a v lineárnych regresných modeloch sa využívajú najmä pri redukcii počtu vysvetľujúcich premenných, v ktorých sa vyskytuje *multikolinearita* (korelácia medzi vysvetľujúcimi premennými), alebo pri dátach s malým počtom záznamov. Pri použití týchto metód riešime optimalizačné úlohy s ohraničením na parametre.

1.4.1 Ridge regresia

Prvá regularizačná metóda, ktorú využívame v tejto diplomovej práci je *ridge* regresia. Pri tejto metóde riešime nasledujúcu optimalizačnú úlohu:

$$\hat{\beta}^{ridge} = \arg\min_{\beta}\{\|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}. \tag{8}$$

Parameter $\lambda \geq 0$ je penalizačný parameter, ktorý udáva veľkosť regularizácie. Čím je väčšia hodnota λ, tým väčšia je veľkosť regularizácie. Idea penalizácie súčtu štvorcov parametrov sa využíva aj v neurónových sieťach. Ekvivalentne sa dá úloha (8) prepísať do tvaru:

$$\min_{\beta} \|Y - X\beta\|^2, \qquad s.t. \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{9}$$

z ktorej vieme explicitne určiť veľkosť ohraničenia t na súčet štvorcov parametrov. Ku každému parametru λ z (8) vieme priradiť práve jeden parameter t z (9).

Tento typ regresie sa využíva, ak máme veľa korelovaných vysvetľujúcich parametrov v lineárnom regresnom modeli. V takom prípade sú odhady koeficientov slabšie a vykazujú vyššiu varianciu. Efekt parametra s vyššou pozitívnou hodnotou môže byť vynulovaný s podobne veľkou negatívnou hodnotou premennej, ktorá je vysoko korelovaná s prvou premennou. Tento jav sa dá zmierniť práve použitím ohraničenia na koeficienty z rovnice (9). Riešenie *ridge* regresie nie je invariantné na škálovanie vstupných hodnôt a preto sa doporučuje štandardizácia vstupných dát pred riešením rovnice (8).

Výhody *ridge* regresie sú napríklad v prípade, ak $p \gg n$, t. j. máme málo pozorovaní s veľkým počtom pozorovaných premenných. V prípade multikolinearity majú vysoko korelované vysvetľujúce premenné približne rovnako odhadnuté parametre. Odhad je vychýlený, ale má menšiu varianciu a menšiu strednú kvadratickú chybu MSE. Optimalizačná úloha má explicitné riešenie. Nevýhodou *ridge* regresie je vlastnosť, že reguluje veľkosti parametrov smerom k nule, ale neprodukuje zjednodušené modely, čo sa týka počtu vysvetľujúcich premenných.

1.4.2 LASSO

Metóda $LASSO$ je podobná metóde *ridge*, ale s odlišnou penalizačnou funkciou. Odhady touto metódou dostaneme riešením nasledujúcej úlohy kvadratického programovania:

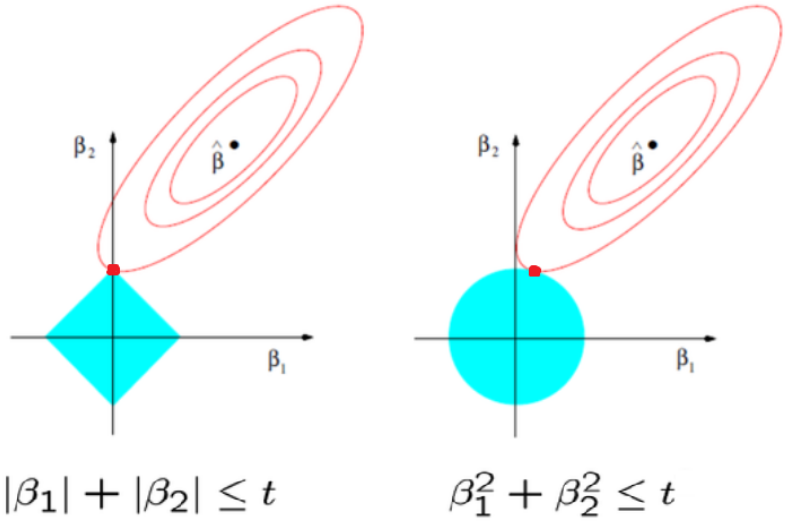
$$\hat{\beta}^{lasso} = \arg\min_{\beta}\{\|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}, \tag{10}$$

kde λ má rovnakú interpretáciu ako v predošlej metóde (8).

Úlohu vieme podobne, ako (9) ekvivalentne prepísať do tvaru:

$$\min_{\beta} \|Y - X\beta\|^2, \qquad s.t. \sum_{j=1}^{p} |\beta_j| \leq t, \tag{11}$$

Všimnime si podobnosť s optimalizačnou úlohou *ridge* regresie z (8), alebo (9). Rozdiel je vo forme penalizačnej funkcie, kde *ridge* regresia používa ℓ_2 normu vektora β, zatiaľčo pri $LASSO$ (11) používame ℓ_1 normu. Táto metóda nemá explicitné riešenie v porovnaní s hrebeňovou regresiou.



Obr. 1: Grafické porovnanie LASSO a ridge regresie v 2D optimalizačnej úlohe (zdroj: [11])

Z vlastnosti $LASSO$ regresie vyplýva, že pri voľbe nízkej hodnoty t docielime, aby niektoré vysvetľujúce premenné mali nulový koeficient, z čoho vyplýva, že $LASSO$ nám zredukuje počet vysvetľujúcich premenných. V opačnom prípade, ak zvolíme t dostatočne veľké, konkrétne $t \geq \|\beta^{ls}\|_1$, tak riešením úlohy (8) je $\beta^{lasso} = \beta^{ls}$, t. j. odhad metódou najmenších štvorcov. V prípade vysoko korelovaných vysvetľujúcich premenných vyberie len jednu premennú a ostatným priradí nulové hodnoty.

Táto metóda je lepšia ako zvyčajné metódy na automatickú selekciu premenných ako *forward*, *backward* alebo *stepwise* metóda. V prípade, že skupina vysvetľujúcich premenných sú vysoko korelované medzi sebou, tak $LASSO$ vyberie len jednu premennú a ostatné zredukuje na nulovú hodnotu.

1.5 Elastic net

Kombináciou predošlých dvoch metód vznikne metóda *elastic net*. V tejto metóde sa rieši nasledujúca optimalizačná úloha:

$$\min_{\beta} \|Y - X\beta\|^2, \qquad s.t. \quad \sum_{j=1}^{p} \left((1-\alpha)\beta_j^2 + \alpha|\beta_j|\right) \leq t,$$

kde parameter $\alpha \in \langle 0, 1\rangle$ je parameter zmiešania, ktorý určuje s akými váhami budú prispievať jednotlivé metódy *LASSO* a *ridge* do výsledného modelu. Pri špeciálnej voľbe parametra $\alpha = 0$ dostávame *ridge* regresiu a v prípade $\alpha = 1$ dostávame *LASSO* regresiu.

V tejto úlohe optimalizujeme už cez dve premenné (α, λ) narozdiel od predchádzajúcich dvoch metód. V tretej kapitole si detailnejšie vysvetlíme proces optimalizácie. Narozdiel od *LASSO* a *ridge* regresie, kde optimalizujeme cez vektor parametrov sa v metóde *elastic net* optimalizuje cez mriežku parametrov.

2 Klasifikačné stromy a náhodné lesy

Medzi ďalšie populárne klasifikačné metódy patria klasifikačné stromy a náhodné lesy. V tejto kapitole si stručne zhrnieme základnú charakteristiku stromových modelov, pravidlá delenia a porovnanie jednoduchých a zložených modelov vytvorených z klasifikačných stromov. Teóriu sme primárne čerpali z [12, 13, 15].

Stromové klasifikačné modely patria medzi modely založené na vytváraní pravidiel, podľa ktorých dáta delíme do cieľových skupín. Výhodou základných klasifikačných stromov je pomerne jednoduchá interpretovateľnosť. V práci sa však venujeme okrem jednoduchých klasifikačných stromov aj náhodným lesom, ktoré vznikajú kombináciami klasifikačných stromov.

Časti stromov sa nazývajú *uzly*, ktoré delíme na dva základné typy: *terminálny uzol* a *neterminálny uzol*. Každý neterminálny uzol je tzv. rodičom dvoch rozličných potomkových uzlov, okrem koreňového uzla, ktorý je na vrchole klasifikačného stromu. V niektorých extrémnych prípadoch sa stáva, že klasifikátor určí koreňový uzol zároveň za terminálnym uzlom. Je to najmä v prípadoch, kedy modelovacia binárna premenná má nevyvážený pomer odoziev. Klasifikácia kreditného zlyhania patrí práve k týmto prípadom a v podkapitole 3.4 si bližšie charakterizujeme tento prípad (počet nezlyhaných úverov výrazne prevyšuje počet zlyhaných úverov).

2.1 Algoritmus rekurzívneho delenia

Algoritmus na budovanie rozhodovacieho stromu začína v koreňovom uzle s celou trénovacou množinou. Pre túto trénovaciu množinu sú vopred známe vstupné atribúty a cielová premenná s hodnotami $\{0, 1\} \equiv \{\text{nezlyhanie, zlyhanie}\}$. Existuje mnoho rôznych verzií klasifikačných stromov, ktoré vznikajú odlišnými metódami delenia jednotlivých častí stromu. Ako prvé sa určí rozdeľovacie kritérium, ktoré najlepšie rozdeľuje dáta (v našom prípade úvery) podľa vstupných popisných premenných do cieľovej binárnej premennej. V práci používame algoritmus, v ktorom z každého neterminálneho uzla vedú dve vetvy, pričom jedna vetva spĺňa a druhá vetva nespĺňa

nejaké rozhodovacie kritérium. Kritérium, ktoré sa na výber testovacej premennej použije na príslušnej úrovni vetvenia [15], závisi od charakteru výstupnej premennej. Základná idea rastu stromu súvisí s čistotou cieľovej premennej v uzloch. Pod zvyšovaním čistoty sa myslí zvyšovanie dominancie jednej hodnoty z cieľovej binárnej premennej v dcérskych uzloch. Pri posudzovaní kvality delenia resp. vetvenia používame *Giniho index*:

$$Gini(\tau) = 1 - \sum_{l} \left(\frac{N(\tau, l)}{N(\tau)} \right)^2,$$

kde τ označuje uzol, l je trieda vysvetľujúcej kategorickej premennej, $N(\tau, l)$ je početnosť kategórie l v uzle τ a $N(\tau)$ je celková početnosť vzorky v uzle τ. V prípade, že *Gini index* je blízky 0, tak ďalšie delenie nie je potrebné. Táto situácia nastáva, ak $N(\tau)$ a $N(\tau, l)$ je blízke 1, t. j. majoritnú časť uzla τ tvorí jedna kategória.

Rozhodovanie o ďalšom delení uzla závisí od miery zmeny *Giniho indexu*. Túto mieru zmeny označujeme *GiniGain* a je definovaná nasledovne:

$$GiniGain(\tau, s) = Gini(\tau) - \frac{N(\tau_q)}{N(\tau)} Gini(\tau_q) - \frac{N(\tau_r)}{N(\tau)} Gini(\tau_r), \tag{12}$$

kde τ_q a τ_r sú dva potomkové uzly vytvorené z uzla τ a s je konkrétne deliace kritérium. V prípade signifikantnej veľkosti $GiniGain(\tau, s)$ sa uzol τ rozdelí na ďalšie dva uzly, pričom sa použije optimálne deliace kritérium s, ktoré dosiahne maximálnu veľkosť $GiniGain(\tau, s)$.

V prípade spojitých resp. ordinálnych premenných je proces hľadania ideálneho deliaceho bodu priamočiary [10], nakoľko sa hodnoty dajú jednoducho zoradiť a môže sa vopred určiť konečný počet deliacich bodov $\theta_i^{(j)}$, $j = 1, \ldots, n_i$ a tým vytvoriť rozhodovacie pravidlá $X_i < \theta_i^{(1)}, \ldots, X_i < \theta_i^{(n_i)}$, ktoré sa budú používať pri rekurzívnom hľadaní optimálnych pravidiel delenia. Podobne aj binárne premenné sú jednoduché na výpočet, nakoľko existuje len jedno možné rozhodovacie pravidlo (v prípade, ak nepripúšťame chýbajúce hodnoty). Vo všeobecnosti pre nominálne premenné s viacerými možnými hodnotami tvoriace konečnú množinu M_i, vytárajú skupinu rozhodovacích pravidiel tvaru $X_i \in M_i^{(j)}$, $j = 1, \ldots, n_i$, kde $n_i = 2^{|M_i|-1} - 1$. Takýto exponenciálny

rast môže mať za následok rýchle navýšenie výpočtovej zložitosti. Pri výbere vhodného deliaceho kritéria musí algoritmus vyhodnotiť všetky možné kritéria pre jednotlivé vstupné premenné a ku každej vyhodnotiť kvalitu rozdelenia (napr. pomocou *Gini-Gain*).

1 Vytvorenie nového uzla τ, ktorý je buď koreňový, alebo potomok existujúceho neterminálneho uzla ;
2 Rozhodnutie, či nový uzol τ má byť terminálny. Uzol je terminálny, ak obsahuje iba jednu triedu cieľovej premennej, alebo dosiahol vopred stanovenú podmienku na prerušenie ďalšieho delenia ;
3 Ak je uzol τ terminálny, priraď tomuto uzlu triedu cieľovej premennej, podľa vopred stanovených pravidiel ;
4 Ak uzol τ je neterminálny, priraď tomuto uzlu rozhodovacie pravidlo (s možnými odpoveďami áno/nie) podľa najlepšieho informačného zisku (12) ;
5 Vráť sa do bodu 1

Algoritmus 1: CART [10]

2.2 Viac stromové modely

V tejto diplomovej práci analyzujeme aj dva zložené klasifikátory: *náhodné lesy* a stromy vzniknuté bootstrapovou agregáciou (angl. *bagged trees*). Všetky ostatné klasifikátory sú tvorené individuálnymi modelmi. Zložené klasifikátory kombinujú výsledné predikcie viacerých základných modelov. V tomto prípade sa kombinujú výsledky základných klasifikačných stromov a spojením týchto stromov vznikajú *viac stromové modely*. Myšlienka kombinovania viacerých modelov zvyšuje predikčnú silu klasifikátora a zároveň znižuje varianciu výslednej predikcie. Klasifikačné stromy sú citlivé na zmeny vstupných dát a malá zmena môže mať za následok komplexnú zmenu štruktúry stromu. Proces tvorby zložených klasifikačných modelov sa skladá z dvoch krokov. V prvej časti sa vytvorí množina individuálnych stromových modelov a v druhej časti je vytvorená finálna predikcia z individuálnych stromov. Vo všeobecnosti sa predikcie zložených klasifikátorov $E(x, M)$ zo vstupného vektora čŕt x a súboru modelov M

vypočítajú:

$$E(x, M) = \frac{1}{N} \sum_{n=1}^{N} \beta_n M_n(x),$$

kde N je počet modelov M_n, $n = 1, 2, \ldots N$, z ktorých sa zložený model vytvára, β_n sú váhy jednotlivých individuálnych predikcií $M_n(x)$.

2.3 Bootstrapová agregácia

Bootstrapová agregácia (ďalej len *bagging*) je špeciálnym algoritmom [13, 15] na tvorenie zložených klasifikačných modelov. Myšlienkou algoritmu je trénovanie modelu na bootstrapovej vzorke dát, čo znamená náhodný výber dát s opakovaním (možnosť opakovania vstupných dát). Rozsah výberu je obyčajne rovnaký, ako rozsah pôvodnej sady dát. Schéma algoritmu je popísana v Alg. 2.

Input: Dátová sada, N = počet stromov

Output: Priemerná výkonnosť modelu z N iterácií

1 **for** *i = 1 to N* **do**

2 Výber náhodnej bootstrapovej vzorky z pôvodnej sady dát ;

3 Trénovanie modelu na bootstrapovej sade dát, pričom sa používa *CART* metóda bez orezávania stromu ;

4 Testovanie modelu na nepoužitých bootstrapových dátach (*out of bag*), prípadne na inej testovej sade ;

5 Výpočet výkonnosti modelu

6 **end**

Algoritmus 2: Bagging

Pri tomto algoritme nie je potrebné manuálne vytvárať tréningovú a testovaciu sadu, nakoľko sa za testovaciu sadu môžu vziať tie žiadosti o úver, ktoré neboli v jednotlivých bootstrapových vzorkách. V práci však budeme porovnávať modely na rovnakých testovacích sadách. Podobne ako pri ostatných zložených klasifikátoroch je tažká interpretácia výsledkov.

2.4 Náhodné lesy

Náhodné lesy [4] sú rovnako založené na trénovaní pomocou bootstrapových vzoriek dát. Okrem toho používajú pri vzniku klasifikačných stromov parameter *mtry* (značenie v súlade s balíkom `randomforest`), ktorý pri každom delení uzla zvolí počet náhodne vybraných premenných, pre ktoré sa hľadá optimálne rozdeľovacie kritérium podľa informačného zisku (12). Druhá vstupná premenná do algrotimu náhodného lesa je počet stromov N, z ktorých je les tvorený. Redukciou množiny opisných premenných sa znižuje korelácia medzi jednotlivými klasifikačnými stromami.

Input: Dátová sada, N = počet stromov, *mtry* = počet premenných
Output: Priemerná výkonnosť modelu z N iterácií

```
1 for i = 1 to N do
2     Výber náhodnej bootstrapovej vzorky z pôvodnej sady dát;
3     Trénovanie modelu na bootstrapovej sade dát pomocou algoritmu CART
        bez orezávania stromu s dodatočným pravidlom ;
4     forall neterminálne uzly τ do
5         Náhodný výber mtry premenných na proces delenia ;
6     end
7     Testovanie modelu na nepoužitých bootstrapových dátach (out of bag),
        prípadne na inej testovej sade ;
8     Výpočet výkonnosti modelu ;
9 end
```

Algoritmus 3: Náhodný les

3 Klasifikácia kreditného zlyhania

Táto kapitola je venovaná krátkemu opisu kreditného rizika, definícii kreditného zlyhania a následné riešenie klasifikačnej úlohy. Obsahuje podrobný popis alogritmu, ktorým sa porovnávajú štatistické metódy z prvých dvoch kapitol. Okrem popisu algoritmov sa v kapitole nachádza aj krátka charakterizácia dátovej sady, nad ktorou budujeme klasifikačné modely.

3.1 Kreditné riziko

Kreditné riziko, alebo aj tzv. úverové riziko, vyjadruje neschopnosť resp. neochotu protistrany plniť svoje záväzky voči veriteľovi. Banky radia toto riziko medzi tie významnejšie riziká a preto je dôležité mu prikladať dostatočnú pozornosť. Kreditné riziko vzniká zo všetkých bankových produktov, pri ktorých dochádza k požičaniu nejakého množstva peňazí klientom. Tieto bankové produkty delíme na homogénne portfólia. Každé portfólio má svoje charakteristické vlastnosti. Dve najväčšie skupiny tvoria retailové a korporátne (firemné) úvery. V rámci týchto dvoch skupín sú jednotlivé úvery zaradené do niekoľkých portfólii. V retailovom segmente to môžu byť krátkodobé úvery s nižšou poskytnutou sumou, alebo aj dlhodobé hypotekárne úvery. Kreditné riziko môže vznikať nepriamo napríklad aj pri debetných kartách, kde je povolené prečerpanie v prípade nedostatku finančných prostriedkov na osobných účtoch.

Existencia kreditného rizika je s nenulovou pravdepodobnosťou na každom úvere. Udalosti vedúce ku kreditnému riziku sú najmä nečakané osobné výdavky, strata zamestnania, ale aj zdravotné problémy, ktoré vedú často krát k neskorému splácaniu finančných záväzkov.

3.1.1 Kreditné zlyhanie

Ako sme už spomenuli v úvode, banky musia účtovne vykazovať stratu vo forme opravných položiek. Finančné štandardy (či už *IAS 39*, alebo *IFRS 9*) rozdeľujú úvery

do skupín podľa veľkosti kreditného rizika. V praxi môžu nastať aj situácie, kde jeden klient má viacej úverov v jednej banke, pričom každý jeden úver sa vyhodnocuje zvlášť. V princípe to nemusí byť nutne úver, ale ľubovolné iné aktívum (produkt banky), ktoré podlieha vyhodnocovaniu kreditného rizika. V takom prípade sa môže stať, že u klienta sa významne navýšilo riziko resp. došlo k zlyhaniu len na jednom z úverov, avšak došlo ku *zlyhaniu na úrovni klienta* a tým pádom sa všetky produkty, ktoré klient vlastní dostanú do skupiny s významným navýšením kreditného rizika. V našej práci sa venujeme iba *zlyhaniu na úrovni úveru*. Hovoríme, že u konkrétneho dlžníka nastalo zlyhanie (default), ak nastala aspoň jedna z nasledujúcich udalostí:

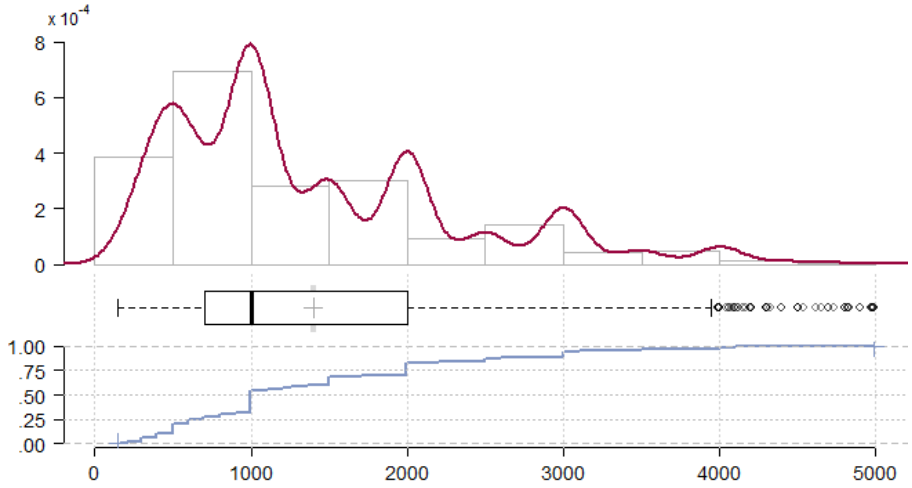
- Dlžník mešká so splátkou viac ako 90 dní.
- Banka považuje za nepravdepodobné, že dlžník splatí svoj záväzok riadne a včas.

Dobu po splatnosti v čase t, budeme v práci označovať DPD_t z anglického (*day past due*). Hodnoty DPD_t sú udávané v dňoch a predstavujú počet dní, ktoré klient mešká s najskoršou nezaplatenou splátkou (v plnej sume) podľa splátkového kalendára.

3.2 Aplikačné dáta

Klasifikačné modely sme aplikovali na dátovú sadu 10059 poskytnutých úverov slovenských klientov. Dáta sú tvorené zo žiadostí o krátkodobé úvery, ktoré obsahujú základné charakteristiky žiadateľov o úver a informácie o poskytnutých úveroch. Na Obr. 2 je zobrazená distribúcia výšky poskytnutých úverov.

Okrem údajov zbieraných pri schvaľovaní úveru sa nachádza aj informácia, či spotrebiteľ počas životnosti úveru meškal so splátkou viac ako 90 dní. Túto udalosť označujeme za zlyhanie spotrebiteľa. Z pôvodnej dátovej sady sme vybrali len tie úvery, pri ktorých bola schválená aspoň dvojročná maturita. Dátová sada obsahuje buď zlyhané úvery, ktoré nemusia byť nutne ukončené, alebo už splatené úvery. Toto kritérium je klúčové, nakoľko zlyhanie klienta prirodzene závisí od dĺžky životnosti úveru. Novoschválené úvery, ktoré ešte nie sú splatené, nemusia vykazovať známky znehodnotenia aktíva,



Obr. 2: Hustota a distribúcia poskytnutej výšky úverov
(zdroj: vlastné spracovanie)

avšak po nejakom čase môže nastať zlyhanie. Ďalšie kritérium bola životnosť úveru. Pri pôvodných krátkodobých úveroch dochádzalo často krát k predčasnému splateniu. Zahrnutím týchto úverov by sme podceňovali váhu zlyhaných úverov. Od úverov sme požadovali, aby čas medzi poskytnutím a ukončením úveru v dôsledku predčasného splatenia bol aspoň 1 rok.

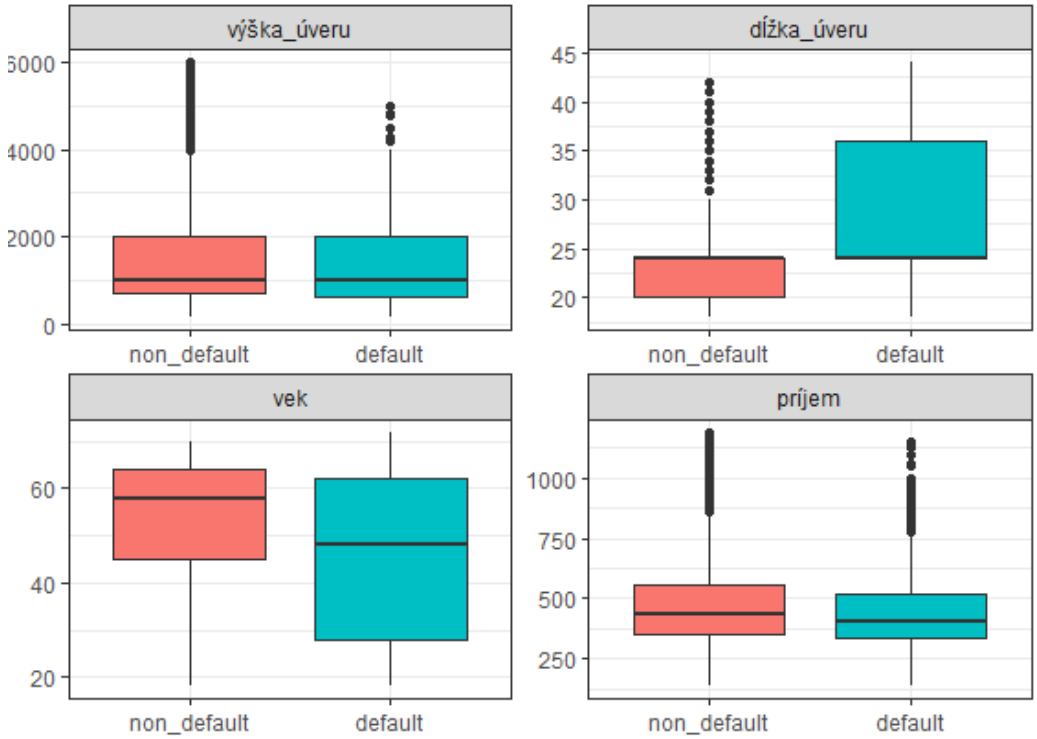
Binárnu premennú kreditného zlyhania modelujeme pomocou 11 premenných, z ktorých 4 sú kvantitatívne a 7 je kategoriálnych. Základná vizualizácia kvantitatívnych dát sa nachádza na Obr. 3:

Modelovacia premenná

Ako sme už spomínali v predošlých častiach práce, modelovacou premennou je úverová charakteristika, či daný úver bol, alebo nebol zlyhaný. V Tab.1 sa nachádzajú pomery zlyhaných a nezlyhaných úverov.

Tabuľka 1: Tabuľka početnosti podľa zlyhania úverov

Kategória	Absolútna početnosť	Relatívna početnosť
Nezlyhaný	8994	89,41 %
Zlyhaný	1065	10,59 %



Obr. 3: Vizualizácia kvantitatívnych premenných pomocou boxplotov (zdroj: vlastné spracovanie)

V klasifikačných úlohách je častokrát problém, ak sú hodnoty modelovacej premennej nevyvážené. V kreditnom skórovaní sú nevyvážené dáta veľmi častým javom a pri klasifikačných modeloch sa používajú rôzne techniky, ktoré tento nepomer berú do úvahy.

3.3 Diskriminačné miery

Za mieru diskriminácie v kreditnom riziku považujeme schopnosť modelu rozoznať zlých klientov od dobrých. V tejto práci na modelovanie používame iba tzv. „modely s učiteľom", kde je známa hodnota cieľovej premennej. Na základe testovacej dátovej sady sa model snaží naučiť predikovať cieľovú premennú na základe vstupných záznamov o jednotlivých klientoch žiadajúcich o úver. Následne sa výkonnosť modelu testuje na testovacej sade dát.

Príklad matice zámen (angl. *confusion matrix*) môžeme vidieť v Tab. 2, v ktorej sa

Tabuľka 2: Matica zámen (angl. *confusion matrix*)

Matica zámen	Predikcia = nezlyhal	Predikcia = zlyhal
Úver nezlyhal	TN	FP
Úver zlyhal	FN	TP

nachádzajú údaje o výsledkoch klasifikácie do cieľových tried. Označenia častí tabuľky sú z angličtiny: TN (true negative), TP (true positive), FN (false negative), FP (false positive), ktoré značia, či model predikoval rovnakú, alebo odlišnú triedu (false/true) cieľovej premennej(negative/positive, resp. 0/1), ako je aktuálna hodnota úveru.

Vysvetlivky k Tab. 2:
TP (true positive) je počet správne zaradených prípadov do triedy zlyhaných úverov.
FP (false positive) je počet nesprávne zaradených prípadov do triedy zlyhaných úverov.
TN (true negative) je počet správne zaradených prípadov do triedy nezlyhaných úverov.
FN (false negative) je počet nesprávne zaradených prípadov do triedy nezlyhaných úverov.

Z početností v Tab. 2 možno vypočítať niekoľko charakteristík, ktoré slúžia ako kritériá na porovnávanie modelov. Prvou charakteristickou veličinou je *presnosť* (angl. *accuracy*), ktorá vyjadruje relatívnu početnosť správne klasifikovaných cieľových premenných a jeho výpočet je nasledovný:

$$\text{Presnosť} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{13}$$

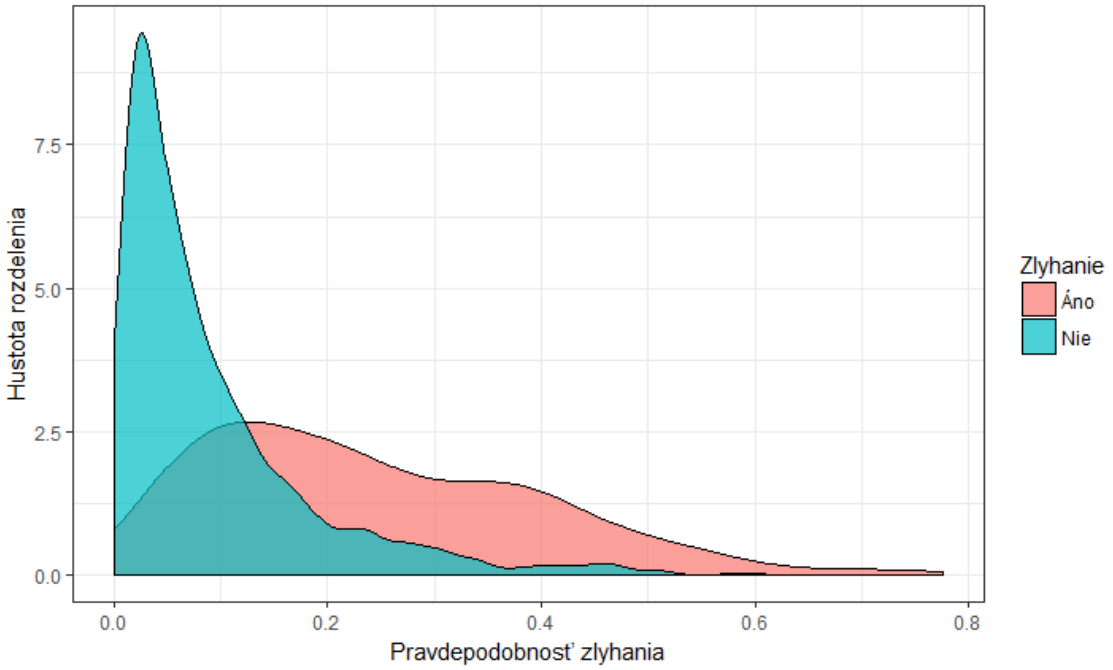
Ak by nás zaujímala iba úspešnosť správneho zaradenia dobrých klientov, vhodným ukazovateľom je *specificita* (angl. *specificity*) modelu

$$\text{Specificita} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14}$$

Pri modelovaní zlyhania banky viac zaujíma odhaľovanie zlých klientov, nakoľko straty z delikventných klientov sú väčšie, ako zisky z riadne splatených úverov. Úspešnosť odhaľovania zlých klientov je možné počítať pomocou *senzitivity*

$$\text{Senzitivita} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

Početnosti v matici zámen (Tab.2) závisia od voľby prahového bodu γ (angl. *cutoff*). Úvery, ktorých odhadnutá pravdepodobnosť zlyhania prekročí vopred stanovený prahový bod, sú zamietnuté. Doteraz spomenuté diskriminačné miery sa môžu vypočítať vždy len pre konkrétne vybranú hraničnú hodnotu. Matica zmien v sebe obsahuje len informáciu, či úver bol, alebo nebol zamietnutý, ale neobsahuje akési „skóre" daného úveru. Nevieme teda určiť, či odhadnutá pravdepodobnosť zlyhania bola tesne, alebo vysoko nad prahovou hodnotou.

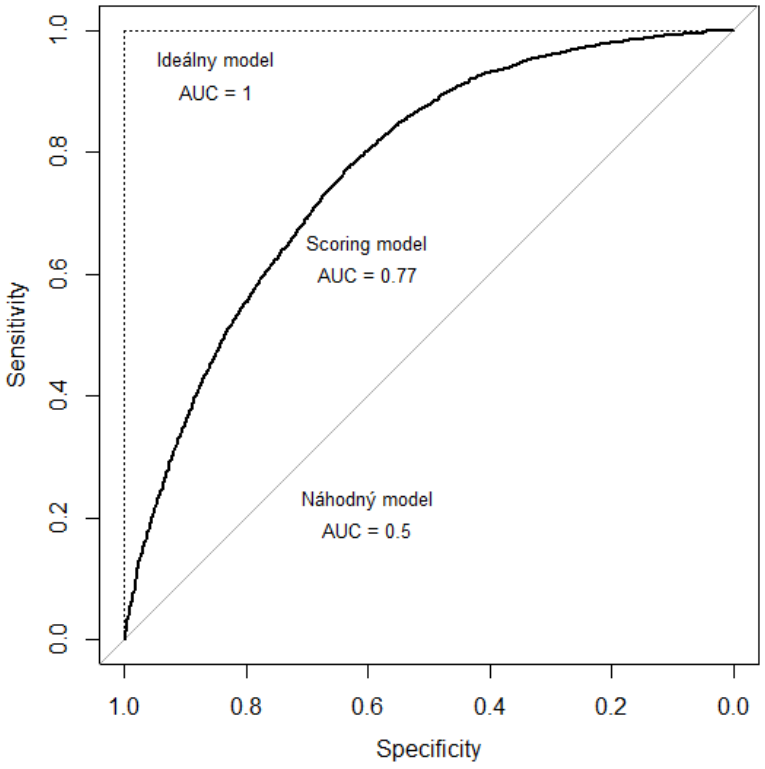


Obr. 4: Hustoty rozdelení odhadov pravdepodonosti zlyhania na testovacej sade dát (zdroj: vlastné spracovanie)

Dobrý klasifikátor je taký, ktorý zlým žiadostiam o úver odhaduje vysoké pravdepodobnosti zlyhania $\widehat{PD} = 1$ a naopak dobrým žiadostiam nízku pravdepodobnosť zlyhania $\widehat{PD} = 0$. Na Obr. 4 je jedna ukážka výsledných hustôt odhadnutých pravdepodobností pre úvery z testovacej sady dát. Aj pre zlyhané úvery existujú vstupné parametre, ktoré sú vyhodnotené ako málo rizikové.

3.3.1 ROC krivka

Na porovnávanie výkonnosti modelov v tejto diplomovej práci používame tzv. *ROC* krivku (*Receiver Operating Characteristic curve*). Táto krivka zohľadňuje výkonnosť modelov pre rôzne hodnoty prahových bodov (ozn. γ). V diagrame *ROC* sú na osi x hodnoty *1-specificita* (v závislosti od hodnoty prahového bodu γ), ktoré vyjadrujú podiel prípadov nesprávne zaradených nezlyhaných úverov. Na osi y sa nachádza *senzitivita* modelu, ktorá vyjadruje podiel správne odhalených zlyhaných úverov. Túto krivku môžeme charakterizovať ako krivku na množine $\langle 0, 1\rangle \times \langle 0, 1\rangle$, kde jednotlivé body sú tvorené súradnicami $[1 - \text{specificita}(\gamma), \text{senzitivita}(\gamma)]$.



Obr. 5: Vizuálna ukážka *ROC* krivky
(zdroj: vlastné spracovanie)

Číselnou charakteristikou *ROC* krivky je obsah plochy pod touto krivkou, ktorú značíme *AUC* (z anglického *area under the curve*). *ROC* krivka ideálneho klasifikátora má plochu $AUC = 1$ a *ROC* krivka prechádza postupne bodmi (0,0) $\rightarrow$ (0,1) $\rightarrow$ (1,1).

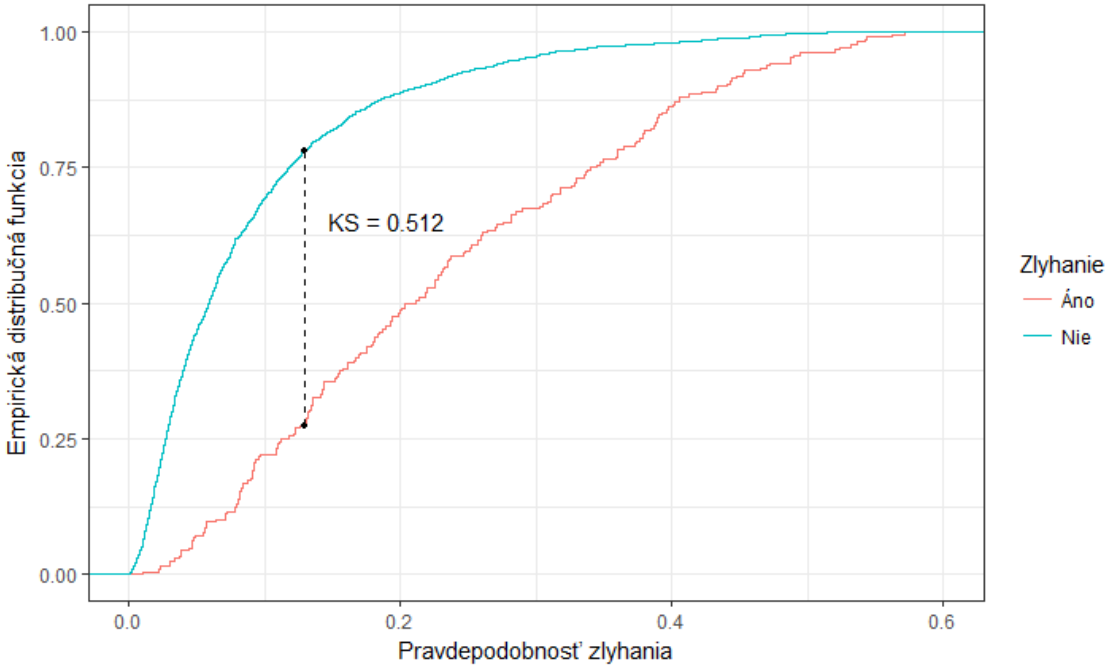
Naopak *ROC* krivka klasifikátora s náhodnou úspešnosťou je tvorená spojnicou bodov (0,0) $\rightarrow$ (1,1), ktorej plocha $AUC = 0,5$. Ukážka *ROC* krivky sa nachádza na Obr. 5. Pri kreditnom skóringu je tiež známy *Gini* koeficient, ktorý sa dá vyjadiť:

$$Gini = 2 \times AUC - 1$$

Výhodou *Gini* koeficientu je obor hodnôt medzi 0 a 1, zatiaľ čo *AUC* nadobúda hodnoty 0,5 až 1.

3.3.2 KS štatistika

Ďalšou diskriminačnou mierou je Kolmogrovova-Smirnovova štatistika (KS), ktorá je definovaná ako maximálna vertikálna vzdialenosť medzi empirickými distribučnými funkciami odhadov pravdepodobností zlyhania pre zlyhané a nezlyhané úvery. Na Obr. 6 sa nachádza vizuálna ukážka KS štatistiky vytvorená pri validácii *probit* modelu.



Obr. 6: Vizualizácia Kolmogrovovej-Smirnovovej štatistiky na skrátenom intervale

3.4 Validácia GLM modelov

Navrhli sme program v softvéri *R* [20], pomocou ktorého sme realizovali a porovnávali klasifikáčné metódy opísané v prvých dvoch kapitolách. Na validáciu modelov používame opakovanú cross-validáciu, ktorú aplikujeme na vopred danú štatistickú metódu a vektor resp. mriežku parametrov cez ktoré optimalizujeme.

Cross-validácia je algoritmus, ktorý rozdelí tréningovú a testovaciu sadu v nejakom pomere (v závislosti od voľby parametra k) do niekoľkých sád dát (ďalej *foldov*). Nakoľko v dátach je pomerne vysoká nevyváženosť, datá budeme deliť do sád náhodne, avšak zachováme relatívny pomer medzi zlyhanými a nezlyhanými úvermi z Tab.1. V Alg. 4 uvádzame pseudokód opakovanej cross-validácie.

Input: Dátová sada, k = počet foldov, n = počet opakovaní
Output: Priemerná AUC z $k \times n$ iterácií
for *i = 1 to n* **do**
    Náhodne rozdelenie dátovej sady na k častí (foldov) so zachovaním relatívnych pomerov z Tab.1 ;
    for *i = 1 to k* **do**
        Testovacia sada $\leftarrow$ i -ty fold;
        Tréningová sada $\leftarrow$ $k-1$ zvyšných foldov;
        Natrénuj model na tréningovej sade;
        Validácia modelu na testovacej sade;
        Výpočet AUC ;
    end
    Výpočet priemernej AUC z k validácii modelov
end
Výpočet priemernej AUC z n opakovaní

Algoritmus 4: n krát opakovaná k-fold cross-validácia

Na meranie výkonnosti modelov používame *ROC* krivku, ktorá porovnáva pomer správne a nesprávne zaradených úverov pri celej škále hraničných hodnôt odhad-

nutých pravdepodobností zlyhania. Mierou výkonnosti je plocha pod touto krivkou (ozn. AUC), ktorú chceme maximalizovať.

Procedúra hľadania optimálnych modelov je vykonávaná pomocou vytvoreného algoritmu na optimalizáciu a balíku `caret`[18]. Pseudokód tohto algoritmu sa nachádza v schéme Alg. 5.

Input: Mriežka parametrov cez ktorú optimalizujeme
Output: Nájdenie optimálnych parametrov podľa AUC
1 Nastavenie vektora resp. mriežky parametrov, cez ktoré optimalizujeme ;
2 Výber štatistického modelu ;
3 **forall** *kombinácie parametrov* **do**
4 | 10 krát opakovaná 5 fold cross-validácia;
5 | Výpočet priemernej AUC pre jednotlivé ROC krivky
6 **end**
7 Nájdenie optimálnej kombinácie parametrov ;
8 Fitovanie modelu na celej sade dát

Algoritmus 5: Hľadanie optimálnych parametrov pre fixne zvolenú metódu

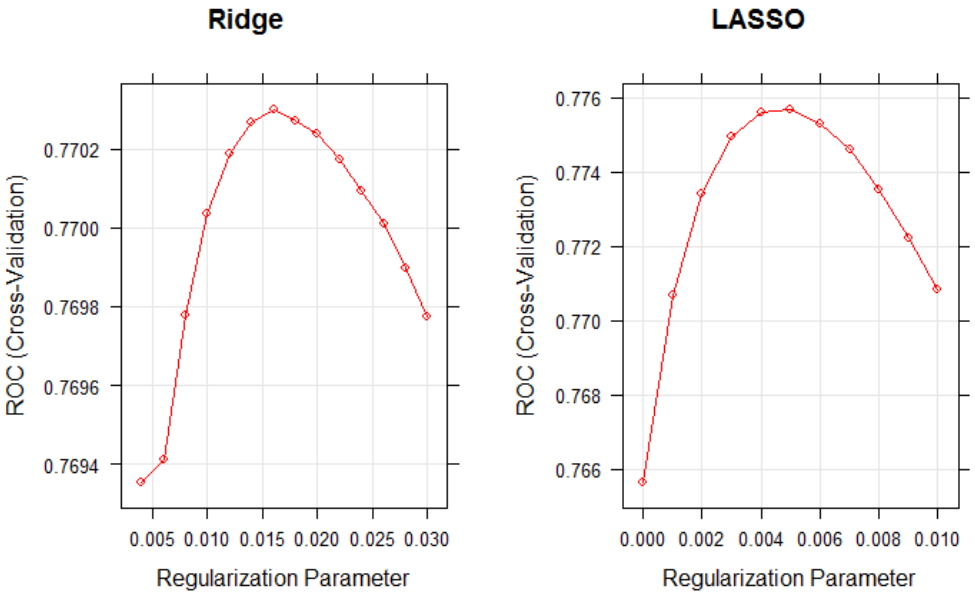
3.4.1 Logit a probit

Pri logistickej a probit regresii používame v balíku `caret` funkciu `glm()` s parametrami `family=binomial(link = logit)` a `family=binomial(link = probit)`. Pre tieto metódy nie sú dodatočné parametre, cez ktoré by sme mohli optimalizovať model. Výsledky uvádzame v podkapitole 3.4.3.

3.4.2 Regularizačné metódy

V rámci regularizačných modelov hľadáme optimálne parametre penalizácie. Tieto parametre hľadáme vzhľadom na maximalizáciu AUC pod ROC krivkou. V prípade *LASSO* a *ridge* regresie sa optimalizuje vzhľadom na jeden parameter λ, kým pri metóde *elastic net* sa optimalizuje aj vzhľadom na parameter zmiešania α. Jeden

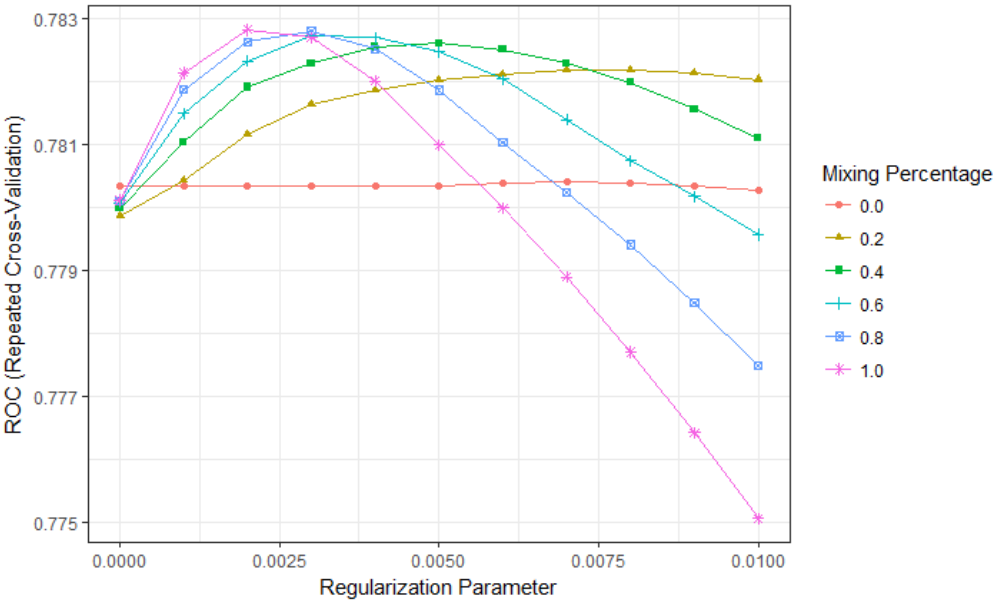
z možných priebehov optimalizácie cez jeden parameter môžeme vidieť na Obr. 7.



Obr. 7: Optimalizácia vektora parametrov λ v metódach *ridge* a *LASSO*

(zdroj: vlastné spracovanie)

V prípade metódy *elastic net* hľadáme optimálne parametre modelu na mriežke tvorenej kombináciami hodnôt parametrov α a λ.



Obr. 8: Optimalizácia parametrov α a λ v *elastic-net* metóde

(zdroj: vlastné spracovanie)

Na Obr. 8 je zobrazený jeden z možných priebehov optimalizácie cez mriežku. Optimálne hodnoty parametrov na zvolenej mriežke sú $\alpha = 1$ a $\lambda = 0,002$.

3.4.3 Porovnanie GLM modelov

Po aplikovaní Alg. 5 na všetky *GLM* modely spomenuté v prvej kapitole, môžeme porovnať sumárne výsledky schopnosti klasifikovať úverové zlyhanie. Vykonali sme 10 krát opakovanú 5 fold cross-validáciu, t. j. výsledky sú z 50-tich validácií.

Tabuľka 3: Priemerné výsledky *AUC* pre *GLM* modely

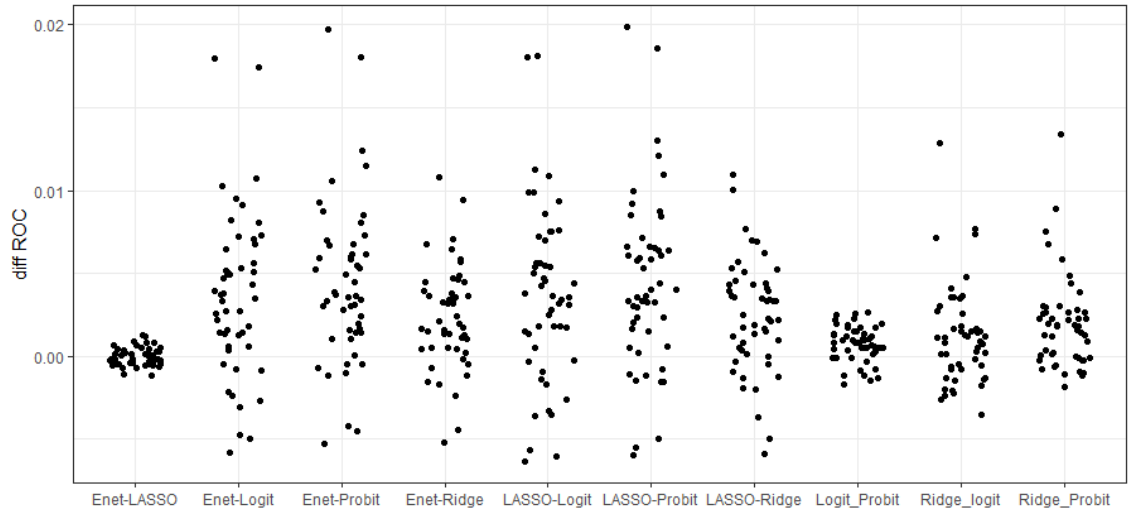
AUC	*Probit*	Logit	Ridge	LASSO	Elastic_net
$\overline{x}$	0,7784	0,7792	0,7804	0,7828	0,7828
σ	0,0169	0,0170	0,0162	0,0164	0,0164

Priemerná výkonnosť aj variancia diskriminačnej miery *AUC* bola veľmi podobná (viď Tab. 3). Podobné výsledky pripisujeme faktu, že cieľovú premennú modelujeme pomocou nízkeho počtu vysvetľujúcich premenných, ktoré nie sú signifikantne korelované. Navyše pri metóde $Elastic-net$, kde sa kombinuje metóda *LASSO* a *ridge*, bola zvolená *LASSO* metóda s plnou váhou.

Keďže sme všetky modely porovnávali na rovnakých testovacích sadách dát, zaujímalo nás aj párové porovnanie výsledkov na jednotlivých sadách. Výsledky párového porovnania predikčnej sily sú na Obr. 9. Zobrazené sú rozdiely hodnôt *AUC* pre všetky dvojice metód na 50-tich validáciach. V článku [9] je popísaný súvis medzi *AUC* a Wilcoxonovou štatistikou. Testy normality (Kolmogrovov-Smirnovov a Shapirov-Wilkovov test) nezamietli normalitu *AUC* hodnôt na hladine významnosti 5%, ale kvôli spomínanému súvisu s Wilcoxonovou štatistikou použijeme na testovanie signifikantnosti rozdielu Wilcoxonov párový znamienkový test. Testujeme hypotézu:

$$\boldsymbol{H_0} : m_d \leq 0 \text{ vs. } \boldsymbol{H_1} : m_d > 0, \tag{16}$$

kde m_d je medián párových rozdielov *AUC* pre dva rôzne modely.



Obr. 9: Porovnanie výsledných rozdielov hodnôt *AUC* jednotlivých *GLM* modelov pri všetkých kombináciach tréningových resp. testovacích sád z cross-validácie

(zdroj: vlastné spracovanie)

V Tab. 4 sa nachádzajú p-hodnoty párových testov (16) spolu so symbolom výsledku testu. Pri testovaní sa počíta $m_d = m_i - m_j \equiv m_{ij}$, kde m_{ij} je medián rozdielov *AUC* pre metódy nachádzajúce sa v i-tom riadku a v j-tom stĺpci Tab. 4. Pre kombináciu *Elastic-net* a *LASSO* sme vykonali aj test rovnosti $m_d = 0$. Výsledkom testu je p-hodnota=0,463 a teda nezamietame hypotézu, že medián rozdielov je nula.

Tabuľka 4: Výsledky párového znamienkového Wilcoxonovho testu

P-value	Probit	Logit	Ridge	LASSO	Elastic net
Probit		$<$	$<$	$<$	$<$
Logit	$3{,}632\times10^{-5}$		$<$	$<$	$<$
Ridge	$2{,}266\times10^{-6}$	0,006377		$<$	$<$
LASSO	$4{,}473\times10^{-7}$	$1{,}339\times10^{-5}$	$8{,}518\times10^{-6}$		$=$
Elastic net	$1{,}302\times10^{-7}$	$4{,}102\times10^{-6}$	$2{,}486\times10^{-6}$	0,7714	

3.5 Validácia stromových modelov

V tejto podkapitole sa zameriame na možnosti modelovania zlyhania so špeciálnymi prístupmi k nevyváženým dátam, nakoľko zlyhaných úverov je proporcionálne málo (viď. Tab. 1).

Na hľadanie pravidiel rozdeľovania a vytvárania klasifikačných stromov používame doplnkový balík `rpart` (classification and regression trees) [21]. Celý algoritmus môžeme zhrnúť v pár krokoch, viď Alg. 1, ktorý môže mať viaceré variácie v závislosti od vopred stanovených pravidiel delenia, alebo charakteristík na celkový klasifikačný strom.

V prípade, ak by sme použili iba jednoduchú funkciu `rpart()` (bez ďalších dodatočných vstupných parametrov) z doplnkového balíka `rpart`, výstupom algoritmu je klasifikačný strom pozostávajúci iba z koreňového uzla, ktorý určí, aby všetky klasifikované hodnoty boli *zlyhané*. V takomto prípade má klasifikátor *presnosť* (13) približne 90%, nakoľko chybne zaradené sú len tie úvery, ktoré zlyhali. Snahou vhodného klasifikátora je však odhaliť zlé žiadosti o úver.

Veľkosť klasifikačného stromu vieme meniť pomocou parametra *cp* (*complexity parameter*), ktorý je štandardne nastavený na hodnotu 0.01. Ide o akúsi hraničnú hodnotu informačného „zisku“, ktorá povolí ďalšie delenie stromu iba v prípade, ak zníženie relatívnej cross-validačnej chyby je vyššia ako daný parameter *cp*. Na nájdenie optimálneho parametra existujú pomocné funkcie `printcp()`, `plotcp()`, pomocou ktorých vieme určiť optimálne hodnoty parametra *cp*. V nasledujúcich podkapitolách (3.5.1 - 3.5.3) porovnáme jednotlivé metódy na vylepšenie klasifikačnych stromov pri nevyvážených dátach s parametrom $cp = 0.001$. Pri takto nízkej hodnote *cp* máme k dispozícii pomerne rozvinutý strom, pri ktorom však dochádza k javu zvanému *overfitting* („preučenie”). Pri overfittingu sa model učí príliš detailne a snaží sa správne klasifikovať aj šum alebo náhodné fluktuácie. Takýto „preučený”model nie je vyhovujúci pre zaraďovanie údajov v testovacej množine, nakoľko model popisuje až príliš podrobne dáta z trénovacej množiny. Tento rozvinutý strom sa „osekáva”(prunning), čo znamená výber menšieho podstromu, ktorý má najlepšiu schopnosť predikcie na testovacej sade dát.

3.5.1 Undersampling

Pri tejto metóde sme z pôvodnej tréningovej sady postupne zredukovali počet nezlyhaných úverov. Záujmom bolo sledovať, ako sa vylepšuje predikčná sila klasifikátora pri postupnom vyrovnávaní pomeru nezlyhaných a zlyhaných úverov v trénovacej sade dát. Pomer zlyhaných a nezlyhaných úverov označme k. Postupovali sme podľa podkapitoly 3.5, t. j. nechali sme strom narásť do väčšej hĺbky a vybrali optimálny podstrom z cross-validácie. Postupne sme zaznamenali priemerné hodnoty AUC a priemerné poradie na všetkých 50 testovacích sadách, podobne ako pri validácii GLM modelov v podkapitole (3.4), pre všetky pomery $k \in \{1, 2, \ldots, 7, 8\}$.

Tabuľka 5: Priemerné výsledky validácie pre metódu undersampling

Pomer k	1	2	3	4	5	6	7	8
AUC	0,721	0,713	0,697	0,675	0,641	0,632	0,622	0,603
Poradie	1,86	2,20	3,32	4,60	5,33	5,98	6,49	6,76

Z výsledkov v Tab. 5 je pozorovateľná citlivosť voľby pomeru tried cieľovej premennej. Z priemerného poradia možno konštatovať, že najlepšie predikčné schopnosti majú klasifikátory, ktoré sa učia pri pomeroch 1:1, alebo 2:1 (pomer nezlyhaných ku zlyhaným úverom). Nevýhodou malého pomeru je nízka efektívnosť využitia dostupných dát. Do ďalšieho porovnávania napriek tomu použijeme pomer $k = 1$.

3.5.2 Prioritné pravdepodobnosti

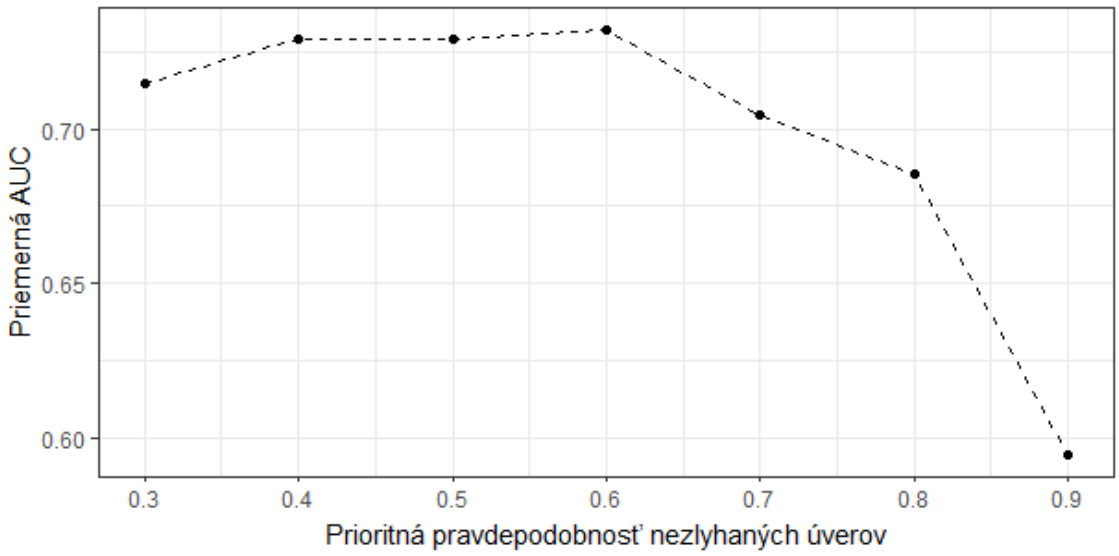
Ďalšou technikou na vylepšenie predikčnej schopnosti nevyvážených dát je nastavenie prioritných pravdepodobností. Základné nastavenie prioritných pravdepodobností v balíku `rpart` [21] je

$$(\pi_1, \pi_2) = \left(\frac{\#\,\text{nezlyhaných úverov}}{\#\,\text{všetkých úverov}}, \frac{\#\,\text{zlyhaných úverov}}{\#\,\text{všetkých úverov}} \right),$$

kde jednotlivé početnosti sú z tréningovej množiny. Približný odhad týchto základných pravdepodobností sa nachádza v Tab. 1. Pri zvyšovaní pravdepodobnosti výskytu zlyhaných úverov sa zvyšuje dôležitosť správnej klasifikácie tejto triedy.

Tabuľka 6: Priemerné výsledky validácie pre metódu prioritných pravdepodobností

(π_1, π_2)	(0.3, 0.7)	(0.4, 0.6)	(0.5, 0.5)	(0.6, 0.4)	(0.7, 0.3)	(0.8, 0.2)	(0.9, 0.1)
AUC	0,715	0,729	0,729	0,732	0,705	0,685	0,594
Poradie	3,60	2,72	2,54	2,36	4,50	5,32	6,96



Obr. 10: Priemerná výkonnosť klasifikátorov v závislosti voľby prioritnej pravdepodobnosti (zdroj: vlastné spracovanie)

Do celkového porovnávania metód použijeme prioritné pravdepodobnosti $(\pi_1, \pi_2) = (0.6, 0.4)$.

3.5.3 Matica strát

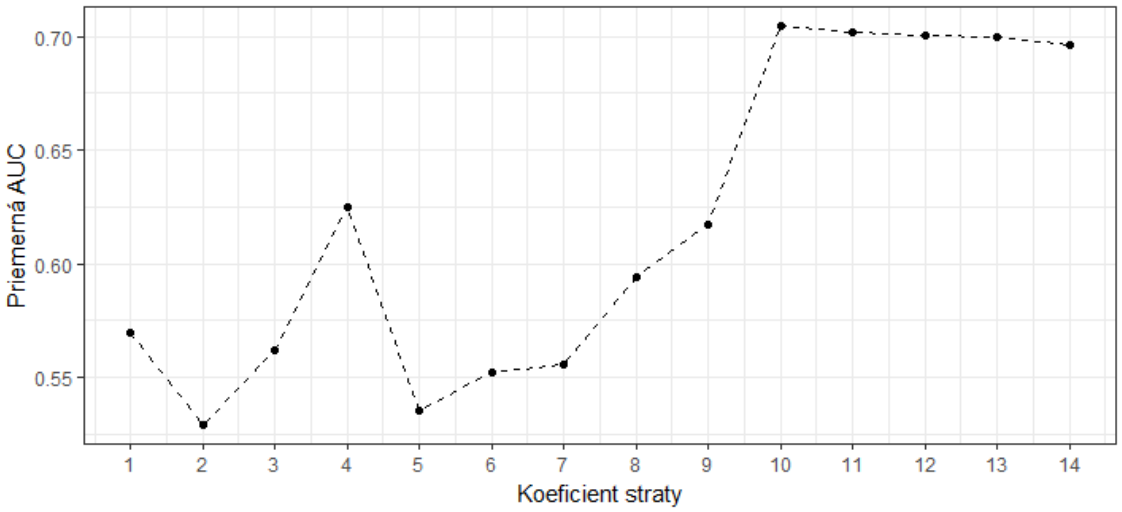
V tejto kapitole sme už v matici zámen (pozri Tab. 2) charakterizovali všetky možné scenáre, ktoré môžu nastať pri klasifikácii. Pomocou *matice strát* (angl. loss matrix), nastavujeme pomerové váhy strát pre jednotlivé scenáre. Nech matica strát L je nasle-

dujúceho tvaru:

$$L = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ s & 0 \end{pmatrix}. \tag{17}$$

Pre banky je výhodnejšie zamietnúť dobrého klienta, ako keby nezamietli delikventného klienta, ktorý nesplatí svoje záväzky. Porovnávame teda čistý ušlý zisk na úrokoch vs. straty v dôsledku nesplatenia úveru.

Pre správne predikované triedy sú hodnoty straty nulové (hlavná diagonála matice L). Pre scenár FP pracujeme so stratou veľkosti $L(1,2) = 1$. Ostáva nám scenár FN, kedy klasifikátor nezamietne úver, ktorý zlyhá. V rámci analýzy pozorujeme predikčnú schopnosť v závislosti od koeficientu straty s, ktorý predstavuje koľkonásobne väčšia je strata v dôsledku zlej predikcie zlyhaného úveru voči zlej predikcie nezlyhaného úveru.



Obr. 11: Priemerná výkonnosť klasifikátorov v závislosti voľby koeficientu straty s (zdroj: vlastné spracovanie)

Z Obr. 11 môžeme pozorovať ustálenie predikčnej schopnosti pre $s \geq 10$. Priebeh pre hodnoty $s < 10$ je pomerne nestabilný. Oproti metódam undersampling a prioritné pravdepodobnosti bola najviac citlivá na správnu voľbu vstupného parametra. Vo viacerých prípadoch ani nízka hodnota *cp* nestačila na rozvinutie koreňového uzla. Do celkového porovnávania metód použijeme maticu strát z rovnice (17) s parametrom $s = 10$.

3.5.4 Bagging

Prvým použitým viac stromovým modelom je bootstrapová agregácia charakterizovaná v podkapitole (2.3). Pomocou baggingu vieme znížiť varianciu výsledných predikcií a zvýšiť predikčnú silu modelu. Rovnako, ako pri validácii GLM modelov (3.4) používame 10 krát opakovanú 5 fold cross-validáciu (Alg.4) kombinovanú s baggingom (Alg. 2). Na *bagging* používame doplnkový balík `ipred` [19].

Tabuľka 7: Porovnanie predikčnej sily metódy bagging v závislosti od počtu stromov

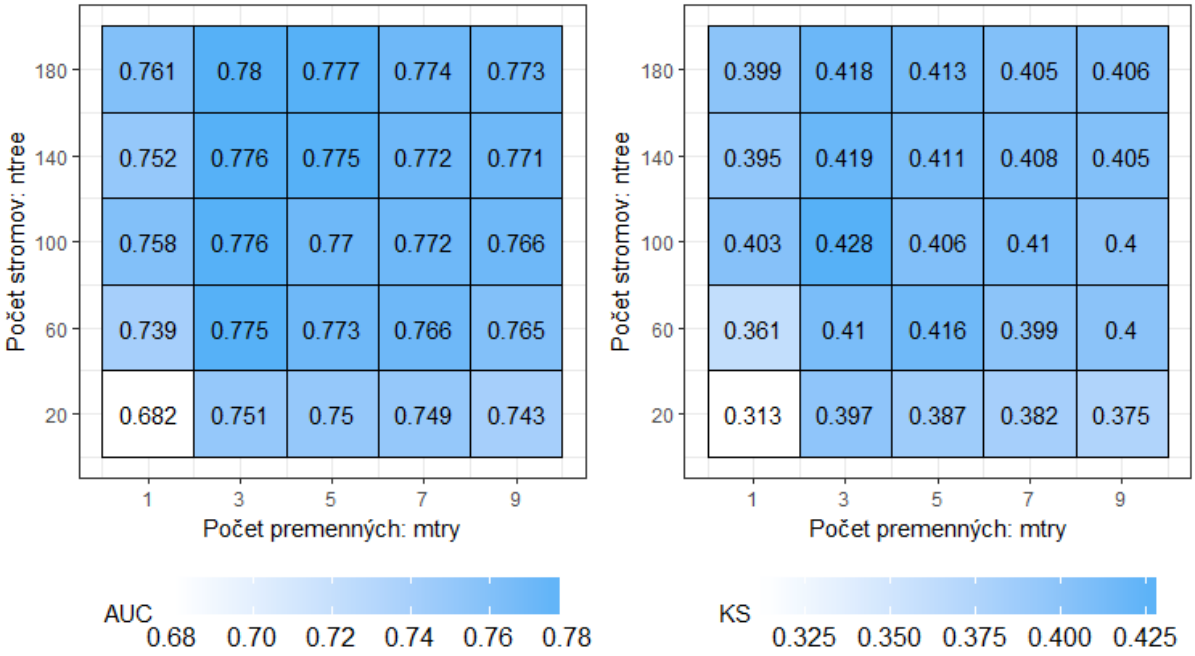
Počet stromov	5	10	15	25	35	45	55	75
AUC	0,690	0,712	0,734	0,760	0,753	0,760	0,760	0,762
KS	0,317	0,334	0,356	0,401	0,383	0,387	0,389	0,396

Z výsledkov závislosti predikčnej sily od počtu stromov v algoritme bagging (Tab. 7), sme za optimálnu vstupnú hodnotu do ďalších porovnávaní zvolili hodnotu $N = 25$ (počet stromov). Táto hodnota je zároveň aj predvolenou možnosťou vo funkcii `bagging` z balíka `ipred`.

3.5.5 Náhodné lesy

Poslednou použitou klasifikačnou metódou v tejto kapitole sú náhodné lesy. Táto metóda je rovnako založená na bootstrapových výberoch z trénovacej množiny dát. Podrobnejšie sme túto metódu charaketrizovali v podkapitole 2.4 a pseudokód tohto algoritmu je v Alg. 3. Na hľadanie vhodných vstupných parametrov používame balíček `randomForests`. V súlade s použitým balíkom značíme počet použitých stromov $ntree$ a počet vybraných premenných pri výbere optimálneho deliaceho kritéria $mtry$. Základné nastavenie týchto dvoch parametrov v balíku `randomForrest`je $ntree = 500$ a $mtry = \lfloor\sqrt{ncol(X)}\rfloor$, kde $ncol(X)$ je počet vysvetľujúcich premenných. V našom prípade je základná hodnota parametra $mtry$ rovná $\lfloor\sqrt{11}\rfloor = 3$. Aj pri tejto metóde analyzujeme citlivosť vstupných premenných na výkonnosť modelu. V rámci analýzy sme hľadali optimálne vstupné parametre na mriežke tvorenej dvoma spomínanými premennými.

Základnú hodnotu $ntree = 500$ sme znížili najmä kvôli časovej úspore z pohľadu efektivity, nakoľko modely s vyšším počtom stromov nemali výrazne lepšiu predikčnú silu.

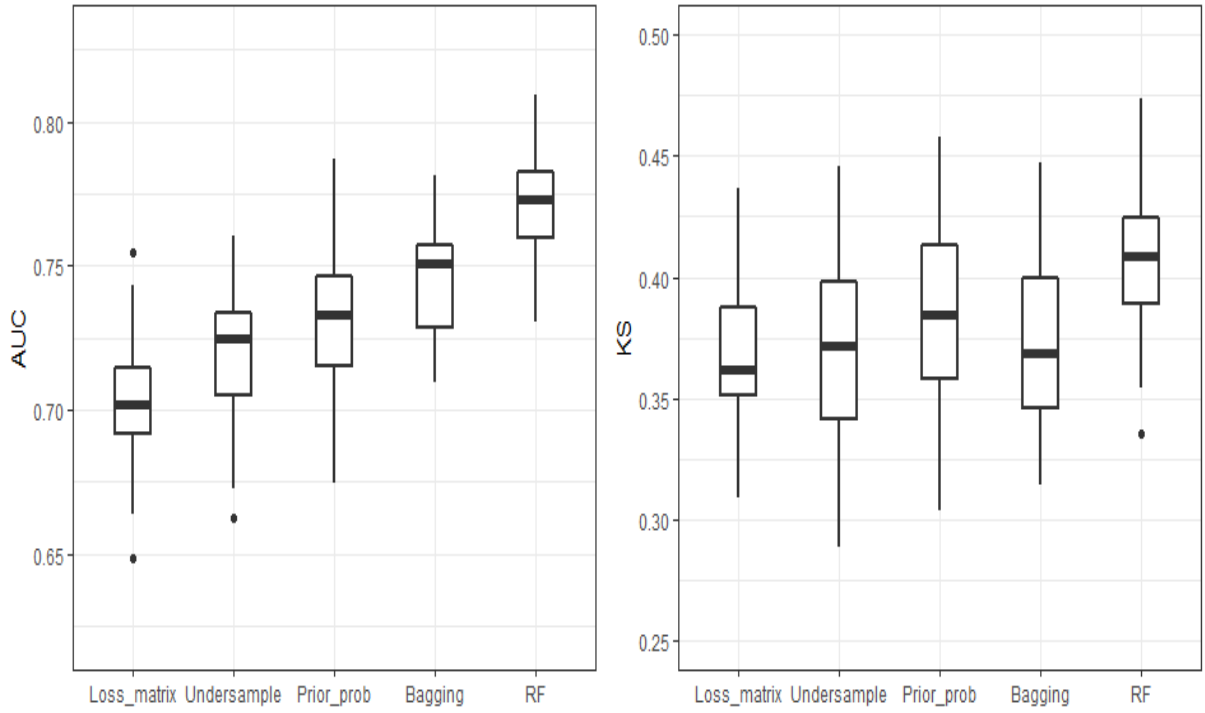


Obr. 12: Priemerná výkonnosť klasifikátorov v závislosti voľby parametrov *ntree* a *mtry* (zdroj: vlastné spracovanie)

Na Obr. 12 sa nachádzajú priemerné výsledky získané na 2 krát opakovanej 5 fold cross-validácii. Do ďalšieho porovnávania použijeme vstupné hodnoty $[mtry, ntree] = [3, 100]$.

3.5.6 Porovnanie stromových modelov

V podkapitole 3.5 sme postupne pre jednotlivé metódy analyzovali predikčné sily v závislosti od vstupných parametrov daných metód. Na základe analýzy sme vybrali optimálne hodnoty vstupných parametrov a následne sme tieto modely s optimálnymi parametrami porovnali.



Obr. 13: Vizuálne porovnanie výsledných diskriminačných hodnôt AUC (vľavo) a KS (vpravo) pre 10 krát opakovanú 5 fold cross-validáciu

(zdroj: vlastné spracovanie)

Na Obr. 13 sa nachádzajú boxploty diskriminačných štatistík, vytvorených z validácii modelov na 50-tich testovacích sadách dát. Náhodné lesy (ozn. RF) boli výkonnostne najlepšie. V oboch diskriminačných mierach dosahovali najvyššiu priemernú hodnotu a zároveň najnižšiu varianciu. Najväčšiu varianciu resp. smerodajnú odchýlku mala metóda prioritných pravdepodobností (viď Tab. 8).

Tabuľka 8: Smerodajné odchýlky diskriminačných mier pre stromové modely

σ	Undersample	Prior prob.	Loss matrix	Bagging	RF
AUC	0,02383	0,02651	0,02055	0,01878	**0,01667**
KS	0,03691	0,03741	0,03311	0,03405	**0,02810**

4 Pravdepodobnosť zlyhania podľa IFRS 9

V tejto kapitole sa zaoberáme odhadom pravdepodobnosti zlyhania za účelom výpočtu opravných položiek podľa novej metodiky *IFRS 9*. Opravná položka je účtovný nástroj, ktorý sa používa vo forme účtovného vykazovania vzniknutých resp. očakávaných strát, obvykle s frekvenciou jedného mesiaca. Používá sa na premietnutie zníženej hodnoty pohľadávky na stranu pasív v dôsledku nesplnenia zmluvných podmienok. Princíp a metodika výpočtu je daná účtovnými štandardmi. Ku dňu 1.1.2018 nadobudol platnosť nový účtovný štandard *IFRS 9*, ktorý mení metodiku výpočtu.

Meranie zníženej hodnoty finančných aktív podľa *IFRS 9* je založené narozdiel od štandardu IAS 39 na modeloch očakávaných kreditných strát. Podľa IAS 39 štandardu sa aktíva rozdeľujú na zlyhané a nezlyhané. Podľa nového štandardu *IFRS 9* (International Financial Repporting Standards) platného od 1.1.2018, pribudne tretia skupina aktív, ktoré sú znehodnotené, t. j. nie sú ešte zlyhané, ale vykazujú signifikantný nárast rizika zlyhania. Hladinu významnosti nárastu kreditného rizika si banky určujú podľa vlastne vytvorených metodík a nie sú jednotne dané pre všetky inštitúcie. Typickým príkladom významného navýšenia rizika je napríklad meškanie so splátkou. Bližší opis signifikantného narástu opisujeme v podkapitole (4.2).

4.1 Kvantifikovanie kreditného rizika

Kvantifikovanie kreditného rizika sa nachádza v každej procesnej časti úveru. Od schvaľovania a prvotného oceňovania úveru, cez priebežný monitoring a reporting úverového portfólia, až po prípadné vymáhacie procesy. Do kvantifikovania vstupujú rôzne zložky kreditného rizika, ktoré si priblížime v nasledujúcich častiach (4.1.1 až 4.1.3):

4.1.1 Pravdepodobnosť zlyhania

Pravdepodobnosť zlyhania vyjadruje pravdepodonosť, že úver za nejaké časové obdobie zlyhá. V tejto práci ho značíme PD (angl. *probability of default*) a pravdepodobnosť

zlyhania v intervale (t_1, t_2) značíme:

$$PD_{t_1,t_2} = P(\exists t \in (t_1, t_2) : DPD_t > 90). \tag{18}$$

V prípade jednomesačnej pravdepodobnosti zlyhania $PD_{t,t+1}$ budeme pravdepobobnosť značiť skrátene PD_t. Pre celoživotnú pravdepodobnosť zlyhania zavedieme nasledovné značenie:

$$PD_{EM} = PD_{t_0,T} = P(\exists t \in \langle t_0, T\rangle : \ DPD_t > 90),$$

kde T predstavuje očakávanú splatnosť (maturitu) úveru.

Odhad PD je pre banky dôležitý, nakoľko sa odhaduje pri viacerých procesných etapách riadenia úverových rizík. Prvý krát sa pravdepodobnosť zlyhania odhaduje pri schvaľovaní úverov. Klient, ktorý má záujem o úver vyplní dotazník, ktorý je následne vyhodnotený klasifikačným modelom. Výstupom klasifikačného modelu je pravdepodobnosť zlyhania, ktorý sa porovnáva s vopred určenou hraničnou hodnotou. V prípade, že odhadnutá PD je vyššia ako hraničná hodnota, úver je zamietnutý. V opačnom prípade sa úver ďalej oceňuje a vypočítava sa úrok pre daný úver. V súčasnosti je už typický tzv. *risk based pricing*, kde sa banka na každého klienta pozerá ako na individuálnu osobu a podľa rizikovosti klienta upravuje výslednú úrokovú sadzbu. Čím je klient menej rizikový, tým menšiu úrokovú sadzbu dostane. Úroková prirážka za rizikovosť je známa ako *riziková marža* a slúži na krytie očakávaný strát v dôsledku zlyhaných úverov.

V tejto diplomovej práci sa venujeme krátkodobým retailovým úverom a za základnú časovú jednotku považujeme 1 mesiac.

4.1.2 Expozícia v čase zlyhania

Expozícia pri zlyhaní *EAD* (*Exposure at Default*) je expozícia k aktuálnému dátumu zlyhania, ktorá je tvorená súčtom zostatkovej istiny, úrokov vyplývajúcich zo splátkového kalendára, nabehlých nezaplatených úrokov a poplatkov, od ktorého sú odpočítané nerozlišené počiatočné poplatky (pri vzniku zmluvy) a pripočítané nerozlišené provízie, ktoré banka vypláca predajcom, alebo obchodným maklérom. V princípe sa jedná

o všetky finančné dlžoby klienta upravené o prípadné ďalšie výdavky banky spojené s daným úverom, ktoré sa však klienta netýkajú.

V čase výpočtu straty zo zníženej hodnoty je známa účtovná hodnota úveru alebo pohladávky, avšak v okamihu prípadného zlyhania nie je známa táto účtovná hodnota. Expozícia pri zlyhaní je ďalším parametrom pre výpočet strát, ktorý je treba odhadnúť. Výška EAD sa počíta na základe:

- Splátkového kalendára.
- Odhadu čiastky nevyčerpaného limitu (ak existuje), ktorá bude v čase zlyhania načerpaná.
- Odhadu očakávaného predčasného splatenia.

4.1.3 Stratovosť v prípade zlyhania

Strata v prípade zlyhania je označovaná LGD zo skratky *loss given default*. Vyjadruje percentuálny podiel nevymožiteľnej čiastky v prípade zlyhania úveru. Hodnoty tejto veličiny sú z intervalu $\langle 0, 1\rangle$ a sú rôzne pre rozličnú škálu bankových produktov.

V prípade hypotekárnych úverov je hodnota LGD obvykle nízka. Pri hypotéke banka spravidla požaduje od dlžníka, aby ako kolaterál založila svoj dom či byt. To znamená, že pri nedodržaní podmienok splácania hypotéky má banka právo stať sa vlastníkom tejto nehnuteľnosti. Nehnuteľnosť môže predať a získať tak späť zvyšnú dlžnú expozíciu v čase zlyhania.

V prípade krátkodobých nekrytých retailových úverov je hodnota LGD vyššia. Po zlyhaní klienta sa úver dostáva do vymáhacieho procesu, kde sa veriteľovi obvykle podarí získať istú percentuálnu časť z dlžnej expozície. V literatúre, ale aj v praxi sa často krát objavuje v súvislosti s LGD aj *miera návratnosti*, ktorá je doplnkovou veličinou k LGD. Označuje sa RR (z angl. *recovery rate*). Hodnota RR je vypočítaná nasledovne:

$$RR = \frac{1}{EAD} \sum_{t=1}^{T} \frac{CF_t}{(1+r)^t},$$

kde CF_t sú dodatočné platby po zlyhaní v mesiaci t ponížené o náklady vzniknuté pri procese vymáhania. Peňažné toky sú diskontované efektívnou úrokovou mierou r. Do dodatočných platieb môžeme zahrnúť aj prípadný odpredaj pohľadávky v mesiaci T, čím sa proces vymáhania končí. Stratovosť sa následne vypočíta zo vzťahu

$$LGD = 1 - RR.$$

4.2 Segmentácia a výpočet ECL

Kolektívne posúdenie finančných nástrojov je podľa *IFRS 9* nutné previesť na báze homogénnych skupín aktív vychádzajúcich zo segmentácie portfólia podľa podobných úverových rizík a produktových charakteristík. V tejto podkapitole popisujeme postupne výpočet ocákávaných strát ECL (*expected credit loss*) v troch rôznych skupinách. Tieto skupiny značíme v tejto práci podľa zaužívaného značenia z článku [6] do tzv. *stage*. Aktíva sú na začiatku každej periódy (zvyčajne jeden mesiac) zaradené do troch *stage-ov* zoradených podľa veľkosti vykazujúceho kreditného rizika.

4.2.1 Stage 1

Do *Stage 1* sú zaradené aktíva po prvotnom vykázaní a následne tie, u ktorých nedošlo od okamihu prvotného vykázánia k výraznému zvýšeniu úverového rizika. Očakávaná úverová strata sa v prípade týchto aktív vypočíta pre obdobie 12 mesiacov, ozn. ECL_{12} od dňa výpočtu (zvyčajne začiatok kalendárneho mesiaca):

$$ECL_{12} = \sum_{t=1}^{min(EM,12)} \frac{PD_t \times LGD_t \times EAD_t}{(1+r)^t},$$

kde

- EM je očakávaná splatnosť (expected maturity)
- PD_t je marginálna pravdepodobnosť zlyhania v mesiaci t

– EAD_t je výška expozície pri zlyhaní v čase t

– LGD_t je strata (percentuálna) pri zlyhaní v čase t

– r je mesačná efektívna úroková miera

Očakávaná splatnosť je v tejto skupine aktív väčšinou blížiaca sa kontraktuálna splatnosť úveru podľa vopred dohodnutého splátkového kalendára. V praxi sa však toto obdobie znižuje o koeficient predčasného splatenia podľa historických údajov o predčasnom splatení. V tejto práci sa nebudeme zaoberať skracovaním očakávanej splatnosti.

V prípade aktív, ktoré sa nachádzajú v *Stage 1*, nedošlo k signifikantnému nárastu kreditného rizika. V rámci pôvodného finančného štandardu *IAS 39* sa pre tieto aktíva netvorili opravné položky, nakoľko tento štandard bol založený na modeli nastalých strát (v zmysle zníženia hodnoty pohľadávky v súvislosti s navýšením kreditného rizika). V prípade nového štandardu *IFRS 9* sa opravné položky musia počítať aj pre kreditne najlepších klientov.

Do kategórie *Stage 1* patria aktíva, ktoré meškajú so splátkou najviac 30 dní a zároveň prechod medzi aktuálne vykazovaným behaviorálnym ratingom a prvotným aplikačným ratingom, získaným pri schvaľovaní úveru, nie je signifikantne odlišný.

4.2.2 Stage 2

Do *Stage 2* sú zaradené aktíva, u ktorých došlo k výraznému navýšeniu úverového rizika od okamihu prvotného vykázánia. Očakávaná úverová strata sa v prípade týchto aktív, ktoré sú segmentované do *Stage 2*, vypočíta pre celoživotné obdobie, ozn. ECL_{EM} od dňa výpočtu (zvyčajne začiatok kalendárneho mesiaca):

$$ECL_{EM} = \sum_{t=1}^{EM} \frac{PD_t \times LGD_t \times EAD_t}{(1+r)^t},$$

kde

– EM je očakávaná splatnosť (expected maturity)

– PD_t je marginálna pravdepodobnosť zlyhania v mesiaci t

– EAD_t je výška expozície pri zlyhaní v čase t

– LGD_t je strata (percentuálna) pri zlyhaní v čase t

– r je mesačná efektívna úroková miera

V pôvodnom štandarde *IAS 39* sa pre znehodnotené pohľadávky počítala iba 12 mesačná očakávaná strata, avšak po zmene štandardu na *IFRS 9* je potrebné počítať celoživotné očakávané straty. Táto zmena metodiky je z pohľadu výpočtu opravných položiek klúčovou zmenou. Banky musia k 1.1.2018 implementovať vlastné štatistické metódy na odhad doživotných pravdepodobností zlyhania. Prístup k tejto problematike je voľný, t. j. v štandarde nie je presne zadané, akým spôsobom sa má táto pravdepodobnosť modelovať.

Pri segmentácii aktív do jednotlivých skupín berieme do úvahy dva indikátory výrazného zvýšenia kreditného rizika:

- Dni po splatnosti
- Prechody medzi ratingovými triedami

Dni po splatnosti

Za vznik výrazného zvýšenia kreditného rizika považujeme okamih, kedy je splátka v omeškaní 31 až 90 dní. Počet dní po splatnosti sa kumuluje dovtedy, kým nie je splatená plná čiastka splátky.

Prechody medzi ratingovými triedami

Dalším indikátorom výrazného zvýšenia úverového rizika je výrazné zhoršenie aktuálnej behaviorálnej ratingovej triedy oproti aplikačnej ratingovej triede priradenej aktívu v deň jeho prvotného rozoznania. Behaviorálny rating sa určuje pomocou interných skórovacích modelov, ktoré zohľadňujú správanie klienta pri priebežnom splácaní úveru.

Významnosť zhoršenia ratingovej triedy si všetky banky určujú podľa vlastnej metodiky. V princípe sa jedná o tabuľky, ktoré obsahujú všetky možné dvojice prechodov z ratingu i do ratingu j a informáciu, či daný prechod predstavuje signifikantný nárast rizika nesplatenia.

V prípade, že je splnená aspoň jedna z vyššie uvedených kritérii, je aktívum zaradené do *Stage 2*.

4.2.3 Stage 3

Do kategórie *Stage 3* patria aktíva, ktoré meškajú so splátkou viac ako 90 dní. V *Stage 3* sa nachádzajú tzv. zlyhané úvery. Tieto úvery majú automaticky najhorší možný rating.

V prípade znehodnotených aktív, zaradených do *Stage 3* je výpočet opravných položiek nasledovný:

$$ECL = Exp_t \times LGD_t,$$

kde Exp_t je aktuálna výška expozície v čase t a LGD_t je strata (percentuálna) pri zlyhaní v čase t. V praxi sa LGD_t mení v závislosti od dĺžky trvania zlyhania, t. j. čím dlhšie je úver zlyhaný, tým vyššia je očakávaná percentuálna strata z hodnoty pohľadávky.

4.3 Markovove reťazce

V tejto časti práce si charakterizujeme základnú teóriu Markovových reťazcov [8], pomocou ktorých modelujeme pravdepodobnosť zlyhania úverov.

Markovov reťazec je stachastický proces $\{X_t\}_{t\geq 0}$, ktorý vytvára postupnosť náhodných premenných $X_0, X_1, \ldots$, s realizáciami $x_0, x_1, \ldots$, ktoré spĺňajú *Markovovu vlastnosť*. Konečná množina S obsahuje všetky možné *stavy* (realizácie) náhodnej premennej X.

Definícia 4.1. Markovova vlastnosť. Hovoríme, že reťazec $\{X_t\}_{t\in T}$ má Markovovu vlastnosť, ak pre každé $n = 0, 1, 2, \ldots$, pre všetky časy $t \geq 0, t \in T$ a pre všetky stavy $x_0, x_1, \ldots, x_n \in S$ platí:

$$Pr(X_{t_{n+1}} = x \mid X_0 = x_0,\ X_1 = x_1, \ldots, X_n = x_n) = Pr(X_{n+1} = x \mid X_n = x_n).$$

Matica prechodu

Pravdepodobnosť zlyhania odhadujeme na základe prechodovej (migračnej) matice. Migračná matica obsahuje pravdepodobnosti prechodov medzi dvomi stavmi v stanovenom časovom období (napr. jeden mesiac). Pravdepodobnosť prechodu z jedného stavu do druhého stavu počítame metódou kohort. Pravdepodobnosť prechodu zo stavu i do stavu j počas základného časového intervalu Δt je daná vzťahom:

$$p_{ij}^{\Delta t} = \frac{N_{ij}}{N_i}, \tag{19}$$

kde N_i predstavuje počet úverov v stave i na začiatku intervalu Δt a N_{ij} predstavuje počet úverov presunutých zo stavu i do stavu j na konci intervalu Δt.

Migračná matica M s n kreditnými stavmi je reprezentovaná štvorcovou $n \times n$ maticou, kde zložky matice $M(i, j)$ sú tvorené pravdepodobnosťami prechodov zo stavu i do stavu j, čo môžeme zapísať $M(i, j) = p_{ij}$.

Z vlastností Markovových reťazcov vyplýva, že v prípade, ak matica ${}_{\Delta t}M_1$ popisuje pravdepodobnosti prechodov z jedného stavu do druhého za jedno základné časové obdobie Δt, tak pravdepodobnosť prechodu po dobu $k \times \Delta t$ je popísaná v matici M_k:

$${}_{\Delta t}M_k = {}_{\Delta t}\, M_1^k. \tag{20}$$

Pre jednoduchší zápis budeme pre základné obdobie $\Delta t = 1$ mesiac vynechávať index migračného obdobia, t. j. ${}_1M_t \equiv M_t$. Nech matica M_k (k-ta mocnina matice M_1) má nasledujúcu štruktúru:

$$M_k = \begin{pmatrix} CP_{11}^{(k)} & \cdots & CP_{1,n-1}^{(k)} & CP_{1n}^{(k)} \\ \vdots & \ddots & \vdots & \vdots \\ CP_{n-1,1}^{(k)} & \cdots & CP_{n-1,n-1}^{(k)} & CP_{n-1,n}^{(k)} \\ 0\% & \cdots & 0\% & 100\% \end{pmatrix},$$

kde $CP_{ij}^{(k)}$ predstavuje kumulatívnu pravdepodobnosť prechodu zo stavu i do stavu j v časovom intervale od t_0 do $t_0 + k \times \Delta t$. Kumulatívne pravdepodobnosti zlyhania stavov i v časovom intervale od t_0 do $t_0 + k \times \Delta t$ sa nachádzajú v poslednom stĺpci matice M_k, pre ktoré platí značenie z rovnice (18):

$$CP_{in}^{(k)} = PD_{t_0,t_0+k}. \tag{22}$$

Pre pravdepodobnosť zlyhania v presne k-tom intervale Δt od času výpočtu t_0, tzv. marginálna pravdepodobnosť v čase k:

$$PD_{k-1,k} = CP_{in}^{(k)} - CP_{in}^{(k-1)}. \tag{23}$$

V tejto diplomovej práci budeme za stavy i a j považovať aplikačné a behaviorálne ratingy. Aplikačný rating sa vypočítava z aplikačných údajov pri žiadosti o úver. Z odhadnutej pravdepodonosti zlyhania na aplikačných údajoch sa úvery zaradia do prvotných rizikových ratingov. Následne sa pre každý mesiac počíta behaviorálny rating, ktorý sleduje správanie klienta pri platení svojich záväzkov. Zohladňuje napr. počet dní po splatnosti alebo aj transakčné údaje klienta na osobnom účte. V práci používame 13 ratingových skupín z množiny R:

$$R = \{\text{A1, A2, A3, B1, B2, B3, C1, C2, C3, D, E, F, Default}\}, \tag{24}$$

z ktorých jeden je rating *Default*, ktorý úver dosiahne v prípade splnenia podmienok na zlyhanie (viď. podkapitola 3.1.1). Ratingy v množine R (24) sú zoradené od najlepšieho (najmenšie kreditné riziko) po najhorší (najväčšie kreditné riziko).

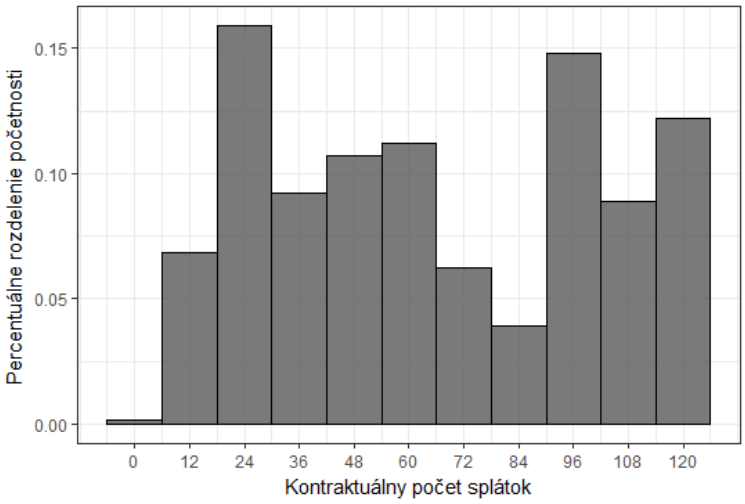
Posledný riadok matice M_k predstavuje kumulatívne pravdepodobnosti prechodov zo zlyhania do ratingov z množiny R. Stav Default nazveme *absorpčným*, čo znamená, že v prípade dosiahnutia tohto stavu si úver už nemôže vylepšiť tento rating. V praxi ak nejaký úver zlyhá, ešte stále môže v kontraktuálnej lehote splatiť úver. V prípade zlyhania a následného opätovného splácania podľa dohodnutého splátkového kalendára, sa úver dostáva do karantény, ktorá trvá niekoľko mesiacov. V tomto čase má úver stále

rating=Default, a v prípade bezproblémových platieb mu je po karanténe opätovne počítaný behaviorálny rating.

Pri odhade migračných matíc tak dostávame nenulové pravdepodobnosti vylepšenia zlyhaného ratingu. Týchto prípadov je však zanedbateľne malý počet a v práci sme po odhade migračných matíc umelo upravili zlyhaný rating na absorpčný stav tak ako je uvedené v matici M_k (21).

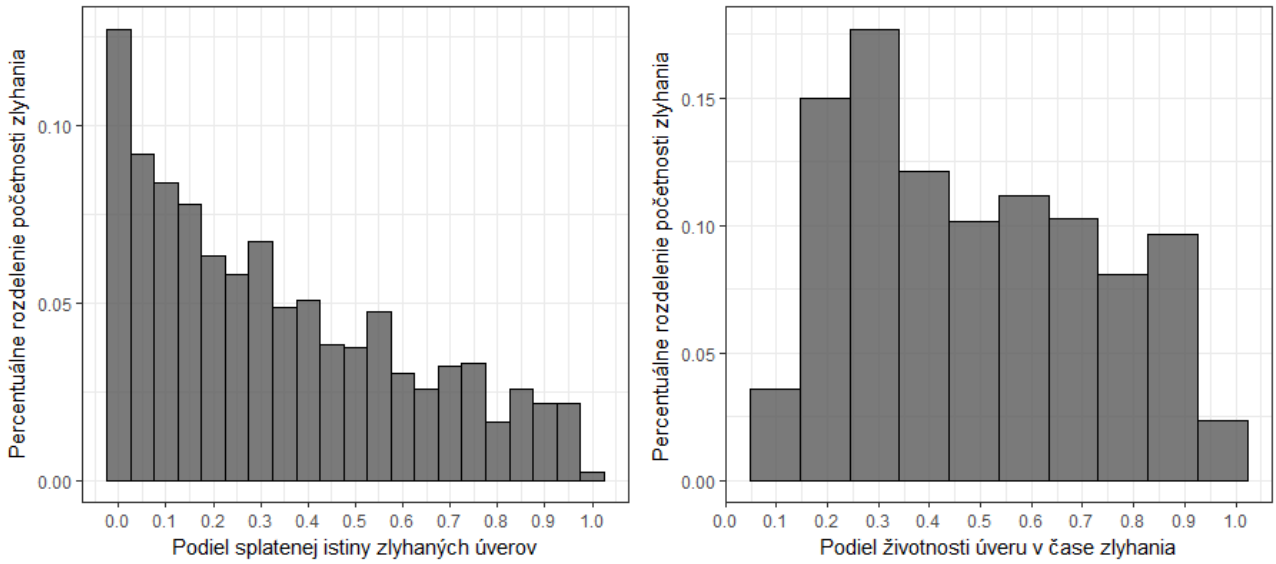
4.4 Vývoj úverov v čase

Na analýzu a porovnanie štatistických metód odhadu pravdepodobnosti zlyhania v rozpätí 12 mesiacov použijeme opäť interné dáta jednej slovenskej banky. Na modelovanie použijeme časové rady s mesačnými údajmi o stavoch jednotlivých úverov. Zaujímať nás budú najmä zmeny behaviorálnych ratingov a správanie zlyhaných úverov. K dispozícii máme 2 081 616 mesačných záznamov o 142 179 spotrebných úveroch v období rokov 2014 až po začiatok roka 2018.



Obr. 14: Percentuálne rozdelenie kontraktuálneho počtu splátok (zdroj: vlastné spracovanie)

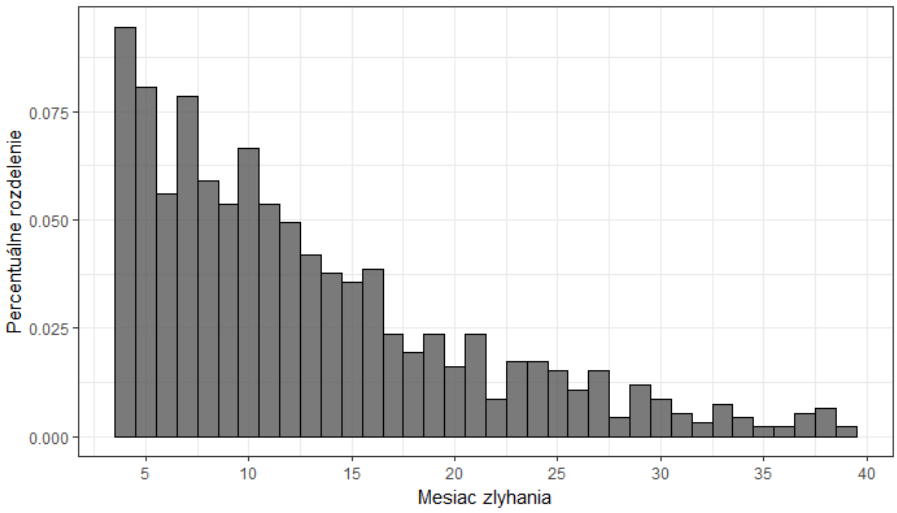
Podiel kontraktuálnej doby splatnosti v mesiacoch môžeme vidieť na Obr. 14. Keďže máme pomerne krátke časové okno dát, väčšina úverov je ešte pred splatnosťou.



Obr. 15: Percentuálne rozdelenie splatenej istiny (vľavo) a životnosti (vpravo) úveru v čase zlyhania

(zdroj: vlastné spracovanie)

Pri odhade miery zlyhania v závislosti od časových parametrov, používame len úvery, ktoré mali kontraktuálnu splatnosť najneskôr na konci časového okna, ktoré máme k dispozícii. V opačnom prípade by sme mali vizuálne výstupy skreslené, nakoľko by sa nadhodnocovala početnosť zlyhaných úverov.



Obr. 16: Percentuálne rozdelenie mesiacu zlyhania od začiatku životnosti úveru

(zdroj: vlastné spracovanie)

Na Obr. 16 je zobraznená absolútna životnosť úveru v čase zlyhania. Počet mesiacov od vzniku kontraktu značíme *MOB* (angl. Month on Book). Môžeme pozorovať, že najväčšia frekvencia zlyhaných úverov je v najkratšej možnej dobe (t. j. *MOB*=4).

4.5 Interval spoľahlivosti

Na odhad intervalu spoľahlivosti (skrát. IS) pre 12 mesačnú pravdepodobnosť zlyhania použijeme Waldov interval spoľahlivosti. Waldov IS [8] predpokladá, že sa pozorované premenné riadia binomickým rozdelením. Nech náhodná premenná X predstavuje udalosť, či úver zlyhá (1), alebo nezlyhá (0). Ak n je počet sledovaných úverov a PD je pravdepodobnosť zlyhania pre nejaký časový krok, tak očakávaný počet zlyhaných úverov pre časový krok rovná $n \times PD$ s disperziou $n \times PD \times (1 - PD)$. Majme súbor pozorovaní X_i, $i = 1, 2, \ldots, n$ (úverov). Zaujíma nás výberový priemer z pozorovaní, t. j. odhadnutá pravdepodobnosť $\widehat{PD}$:

$$\widehat{PD} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Pre dostatočne veľké n sa podľa centrálnej limitnej vety odhad $\widehat{PD}$ riadi približne normálnym rozdelením $N(\mu, \frac{\sigma^2}{n})$:

$$\widehat{PD} \sim N\left(\widehat{PD}, \frac{\widehat{PD}(1 - \widehat{PD})}{n}\right)$$

Veľkosť $(1 - \alpha)\%$ Waldovho intervalu spoľahlivosti je potom:

$$IS_W = \widehat{PD} \pm \kappa \sqrt{\frac{\widehat{PD}(1 - \widehat{PD})}{n}}, \tag{25}$$

kde κ je $(1 - \frac{\alpha}{2})$ kvantil štandardizovaného normálneho rozdelenia.

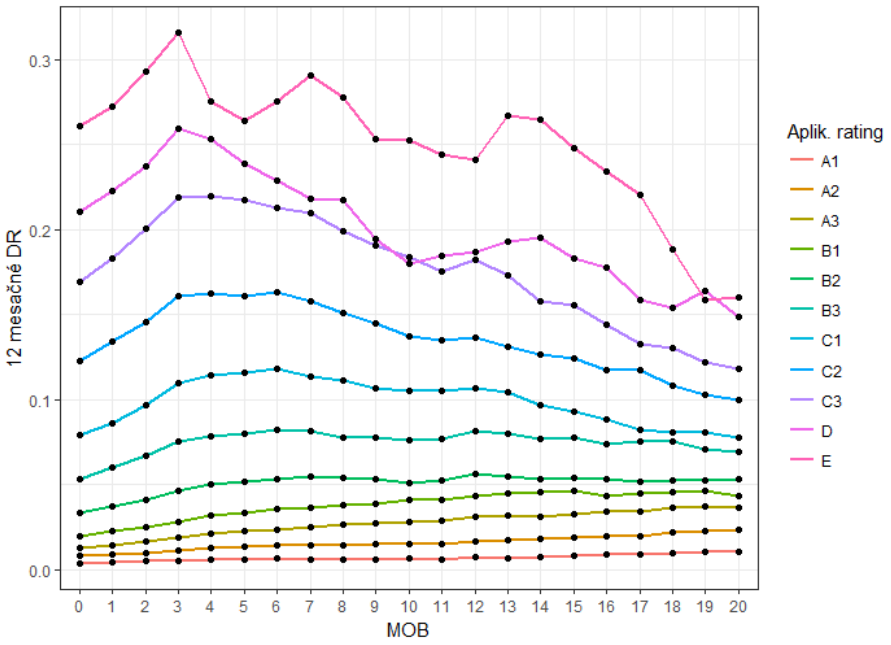
4.6 Ročná miera zlyhania

Podľa štandardu *IFRS 9* sú všetky banky povinné počítať minimálne 12-mesačné pravdepodobnosti zlyhania. V tejto podkapitole porovnáme rôzne prístupy k odhadu pravdepodobnosti zlyhania. Na odhad mier zlyhania sme použili všetky úvery, ktoré

mali k dispozícii aspoň 12 mesačnú životnosť. Nech R_i predstavuje rating úveru pre $MOB = i$, potom odhad 12 mesačných mier zlyhania (označ. DR z default rate) v závislosti od aplikačného ratingu R_0 a MOB-u počítame:

$$DR^{12}(\text{MOB}, R_0) = Pr[\exists t \in (\text{MOB}, \text{MOB+12}\rangle : R_t = \text{Default} \mid R_{MOB} \neq \text{Default}] \tag{26}$$

Mieru zlyhania z rovnice (26) pre konkrétny MOB a aplikačný rating R_0 vypočítame ako podiel zlyhaných úverov v čase (MOB, MOB+12⟩ z celkového počtu nezlyhaných úverov (v konkrétnom MOB-e), ktoré majú aplikačný rating R_0. Na Obr. 17 je zobrazený časový priebeh mier zlyhania pre jednotlivé aplikačné úvery.



Obr. 17: Časový priebeh 12 mesačných mier zlyhania pre jednotlivé aplikačné ratingy v závislosti od počtu mesiacov životnosti úveru.

(zdroj: vlastné spracovanie)

Môžeme pozorovať signifikantný nárast do tretieho mesiaca, ktorý predstavuje posledný možný mesiac, v ktorom ešte nemôže nastať zlyhanie podľa pravidla $DPD > 90$.

4.6.1 Časový krok migračnej matice

V prvom rade analyzujeme vplyv voľby Δt na úspešnosť fitovania 12 mesačnej miery zlyhania. Postupne sme zvolili za časový krok $\Delta t = \{1, 2, 3, 4, 6, 12\}$ mesiacov. Pre jednotlivé voľby Δt sme odhadli matice prechodu podľa postupu z rovnice (19). Následne sme vypočítali kumulatívne pravdepodobnosti 12 mesačných zlyhaní z matíc

$$M_{12},\ {}_2M_6,\ {}_3M_4,\ {}_4M_3,\ {}_6M_2,\ {}_{12}M_1.$$

Kumulatívne pravdepodobnosti zlyhania sa nachádzajú v poslednom stĺpci týchto matíc.

Pri výbere vhodnej krivky je snahou vybrať najpodobnejšiu krivku ku krivke *DR estimated*, ktorej hodnoty sú vypočítane ako $DR^{12}(MOB = 0, R_0)$ z rovnice (26). V tejto podkapitole použijeme značenie $\widehat{PD}$ pre odhadnuté 12 mesačné miery zlyhania. Úlohu hľadania najpodobnejšej krivky zlyhaní sme riešili dvomi spôsobmi.

Prvý spôsob obsahoval hľadanie kriviek pomocou regresie. Keďže odhadnuté miery zlyhania používajú pri odhade početnosť aplikačných ratingov, rozhodli sme sa pre regresiu s váhami pre jednotlivé ratingy. V Tab. 9 sa nachádza percentuálne rozdelenie aplikačných ratingov v použitej vzorke dát.

Tabuľka 9: Percentuálne rozdelenie početnosti aplikačných ratingov

Rating	A1	A2	A3	B1	B2	B3
w	14,362%	10,391%	14,912%	13,926%	12,842%	11,744%

Rating	C1	C2	C3	D	E	F
w	9,481%	6,770%	4,371%	0,822%	0,375%	0,004%

Podobnosť kriviek sme kvantifikovali pomocou váženého súčtu štvorcov rezíduí:

$$wRSS = \sum_{r \in R} w_r \left(\widehat{PD}_r - {}_{\Delta t}\, \widehat{DR}_r \right)^2, \tag{27}$$

kde ${}_{\Delta t}\widehat{DR}_i$ sú odhadnuté 12 mesačné pravdepodobnosti zlyhania pomocou časovo diskrétnych Markovových reťazcov (DTMC).

Tabuľka 10: Porovnanie odhadnutých 12 mesačných mier zlyhania pomocou diskrétnych Markovových reťazcov (DTMC) metódou súčtu vážených štvorcov rezíduí

Δt	1	2	3	**4**	6	12
wRSS $\times 10^6$	67,358	97,565	51,079	**41,851**	62,529	248,13
Poradie	4	5	2	**1**	3	6

Výsledné vážené súčty štvorcov odchýlok pomocou DTMC (*discrete time Markov chain*) sa nachádzajú v Tab. 10. Pre jednoduchšie porovnanie sme výsledné $wRSS$ dodatočne prenásobili konštantou 10^6.

Druhý prístup, ktorý sme aplikovali na hľadanie najlepších migračných matíc je pomocou intervalov spoľahlivosti. V prvom rade sme vypočítali Waldové intervaly spoľahlivosti IS_{Wr} charakterizované v podkapitole 4.5 pre jednotlivé ratingové skupiny r. Následne sme odhadli diskrétne Markovove migračné matice s krokom Δt, ktoré sme vynásobili $12/\Delta t$ krát (resp. vypočítali mocninu stupňa $12/\Delta t$) , čím sme vypočítali 12 mesačné kumulatívne matice prechodu. V poslednom stĺpci týchto matíc sa nachádzajú 12 mesačné miery zlyhania ${}_{\Delta t}\widehat{DR}_r$ pre ratingy $r \in R$. Definujme δ_r:

$$\delta_r = \begin{cases} 1 & {}_{\Delta t}\widehat{DR}_r \in IS_{Wr}, \\ 0 & {}_{\Delta t}\widehat{DR}_r \notin IS_{Wr}. \end{cases}$$

Naším cieľom je nájsť také migračné matice, pre ktoré maximalizujeme $\sum_{r \in R} \delta_r$. Jednoducho povedané, suma $\sum_{r \in R} \delta_j$ predstavuje počet ratingov, ktoré sa nachádzajú v 95% IS_W pre jednotlivé migračné matice. V Tab. 11 sa nachádzajú odhadnuté miery pomocou DTMC metódy spolu s intervalmi spoľahlivosti. Hodnoty v Tab. 11 sú zaokrúhlené na dve desatinné miesta, avšak hodnoty δ_r počítame s nezaokrúhlenými hodnotami IS_{Wr} resp. ${}_{\Delta t}\widehat{DR}_r$. Pre aplikačný rating F nemáme dostatočný počet úverov na odhad IS_W.

V Tab. 12 sa nachádza ukážka odhadnutej migračnej matice, ktorá najlepšie odhadovala 12 mesačné miery zlyhania $\widehat{PD}$ podľa oboch prístupov hľadania migračnej matice.

Tabuľka 11: Waldove intervaly spoľahlivosti (95%) pre 12 mesačné pravdepodobnosti zlyhania spolu s odhadnutými pravdepodobnosťami zlyhania pomocou časovo diskrétnych Markovových reťazcov (DTMC) s krokom Δt

DTMC				Δt					
Rating	$\widehat{PD}$	IS_W		1	2	3	4	6	12
A1	0,38%	0,28%	0,48%	0,22%	0,20%	0,37%	0,46%	0,48%	0,48%
A2	0,80%	0,63%	0,97%	0,38%	0,39%	0,66%	0,81%	0,85%	0,86%
A3	1,29%	1,10%	1,47%	0,65%	0,63%	0,93%	1,11%	1,17%	1,28%
B1	1,97%	1,74%	2,21%	1,36%	1,34%	1,74%	1,97%	2,06%	2,20%
B2	3,34%	3,01%	3,66%	2,46%	2,48%	3,00%	3,25%	3,24%	3,65%
B3	5,35%	4,94%	5,77%	4,04%	3,85%	4,26%	4,47%	4,62%	5,24%
C1	7,94%	7,39%	8,48%	7,00%	6,89%	7,17%	7,41%	7,77%	8,64%
C2	12,31%	11,53%	13,09%	11,84%	12,03%	12,30%	12,85%	13,49%	15,56%
C3	16,91%	15,81%	18,00%	16,54%	17,55%	17,87%	18,35%	19,58%	22,85%
D	21,07%	18,66%	23,47%	22,88%	24,24%	23,68%	23,36%	23,23%	23,33%
E	26,12%	22,05%	30,18%	31,36%	34,40%	32,88%	32,16%	31,42%	31,59%
F	—	—	—	42,49%	47,74%	46,71%	46,12%	45,46%	45,00%
$\sum_{r \in R} \delta_r$				3	2	5	8	6	6

Tabuľka 12: 12 mesačná kumulovaná matica prechodu ${}_4M_3$ pre metódu DTMC s $\Delta t = 4$

Rating	A1	A2	A3	B1	B2	B3	C1	C2	C3	D	E	F	Default
A1	37,4%	11,3%	28,7%	10,0%	2,6%	4,4%	2,7%	0,7%	0,4%	0,4%	0,3%	0,5%	0,5%
A2	20,2%	8,1%	36,2%	14,7%	4,1%	6,5%	4,9%	1,3%	0,7%	0,8%	0,6%	1,1%	0,8%
A3	12,7%	6,3%	38,7%	16,9%	4,8%	7,5%	6,1%	1,7%	0,8%	1,1%	0,8%	1,5%	1,1%
B1	5,6%	3,3%	28,6%	22,5%	6,9%	10,8%	9,7%	2,9%	1,4%	1,9%	1,4%	2,9%	2,0%
B2	3,1%	2,0%	20,4%	20,9%	7,8%	13,2%	12,9%	4,2%	2,0%	3,0%	2,3%	5,1%	3,2%
B3	2,2%	1,5%	16,7%	19,2%	7,9%	14,1%	14,2%	4,8%	2,3%	3,5%	2,8%	6,4%	4,5%
C1	1,2%	0,9%	11,7%	16,0%	7,2%	13,8%	15,6%	5,5%	2,7%	4,5%	3,7%	9,6%	7,4%
C2	0,6%	0,5%	7,3%	11,8%	5,9%	12,2%	15,3%	6,0%	3,0%	5,6%	4,7%	14,2%	12,9%
C3	0,4%	0,3%	4,9%	9,0%	4,7%	10,4%	14,3%	6,0%	3,1%	6,1%	5,3%	17,2%	18,3%
D	0,2%	0,2%	3,1%	6,3%	3,6%	8,4%	12,8%	5,9%	3,1%	6,6%	5,9%	20,6%	23,4%
E	0,1%	0,1%	1,7%	3,8%	2,3%	5,8%	10,0%	5,1%	2,8%	6,5%	6,1%	23,5%	32,2%
F	0,0%	0,0%	0,5%	1,4%	0,9%	2,6%	5,7%	3,6%	2,1%	5,5%	5,7%	25,8%	46,1%
Default	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	100,0%

Problémom voľby časových krokov $\Delta t \neq 1$ je výpočet 1 mesačných pravdepodobností zlyhania. Pre výpočet opravných položiek (či už úvery v Stage 1, Stage 2, alebo v Stage 3) je potrebný výpočet medzimesačných pravdepodobností zlyhania (pozri podkapitolu 4.2). Pri odhade týchto hodnôt sme použili Markovove reťazce so spojitým časom - CTMC (*continuous-time Markov chain*), pomocou ktorých vieme vypočítať pravdepodobnosti prechodov aj medzi časovými krokmi Δt.

Spojité reťazce [2] sú charakterizované takou maticou Q, že $\exp(tQ)$ definuje maticu prechodu pre interval [0,t], kde $\exp(\cdot)$ značí maticovú exponenciálu. Markovove reťazce so spojitým časom sú nadstavbou reťazcov s diskrétnym časom, nakoľko vieme odhadovať pravdepodobnosti prechodu aj na užších časových intervaloch. Pre maticu prechodu s diskrétnym časom ${}_{\Delta t}M$ nájdeme generujúcu maticu Q:

$$M(t) = e^{tQ} = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}. \tag{28}$$

Matica $M(t)$ definovaná v (28) je stochastická matica pre všetky $t \geq 0$ práve vtedy, ak $Q = (q_{ij})$ spĺňa nasledujúce podmienky [2]:

(i) $0 \leq -q_{ii} \leq \infty$ pre všetky $i = 1, \ldots, 8$;

(ii) $q_{ij} \geq 0$ pre všetky $i \neq j$;

(iii) $\sum_{j=1}^{n} q_{ij} = 0$ pre všetky $i = 1, \ldots, 8$.

V teórii Markovových reťazcov sa takéto matice nazývajú aj ako generátory. Odhad takejto matice vykonávame pomocou doplnkového balíka `ctmcd`, ktorý ponúka viac metód odhadu. Pri odhade matice Q [14] sa využíva inverzný vzťah z rovnice (28):

$$Q = \log(M) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (M - I)^k. \tag{29}$$

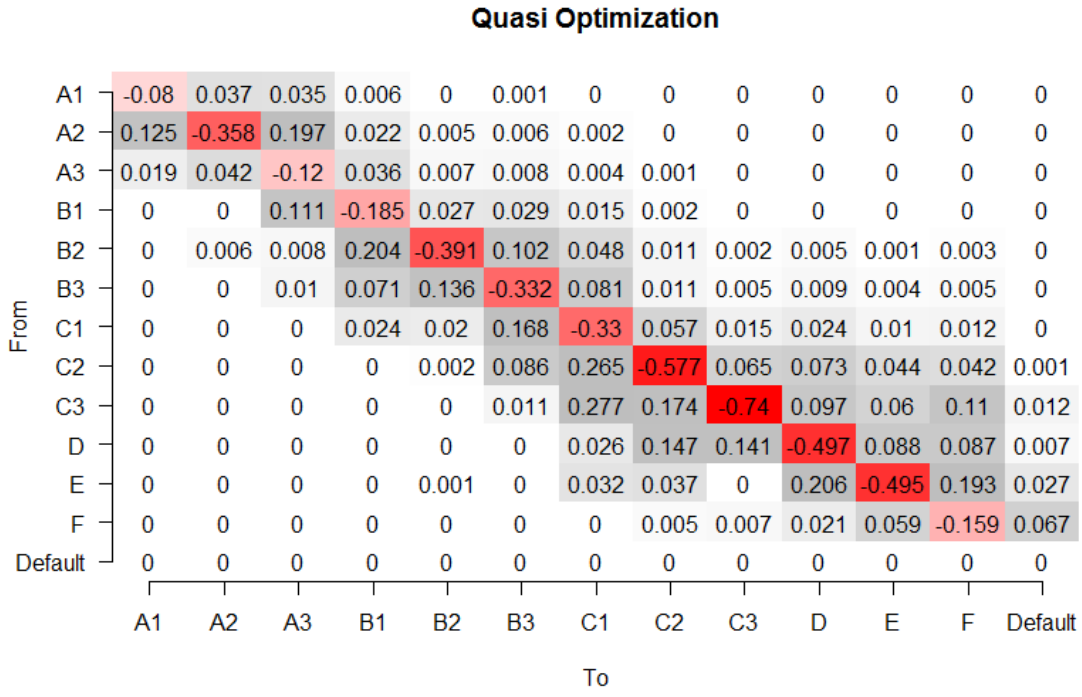
Porovnávame tri metódy (*diagonal adjustment*, *weighted adjustment* a *quasi optimization*), ktorých proces odhadu je popísaný v dokumentácii [14] k doplnkovému balíku `ctmcd`.

Pri odhade generátora Q z troch spomenutých metód sme najprv vybrali tú, ktorej exponenciála bola najpodobnejšia diskrétnej matice prechodu ${}_{\Delta t}M$ v zmysle minimálnej hodnoty Euklidovskej normy rozdielu matíc:

$$||M - exp(Q)||_2 = \left(\sum_i^n \sum_j^n (M_{ij} - \exp(Q)_{ij})^2 \right)^{1/2} \tag{30}$$

Na výpočet exponenciály $\exp(Q)$ sme použili funkciu `expm(·)` z doplnkového balíka `Matrix` [17], ktorá vypočíta maticovú exponenciálu.

Na Obr. 18 je ukážka Q matice odhadnutej z diskrétnej matice prechodu ${}_6M_1$. Matica Q je navyše predelená $\Delta t = 6$, čím dostávame logaritmus jednomesačnej matice prechodu.

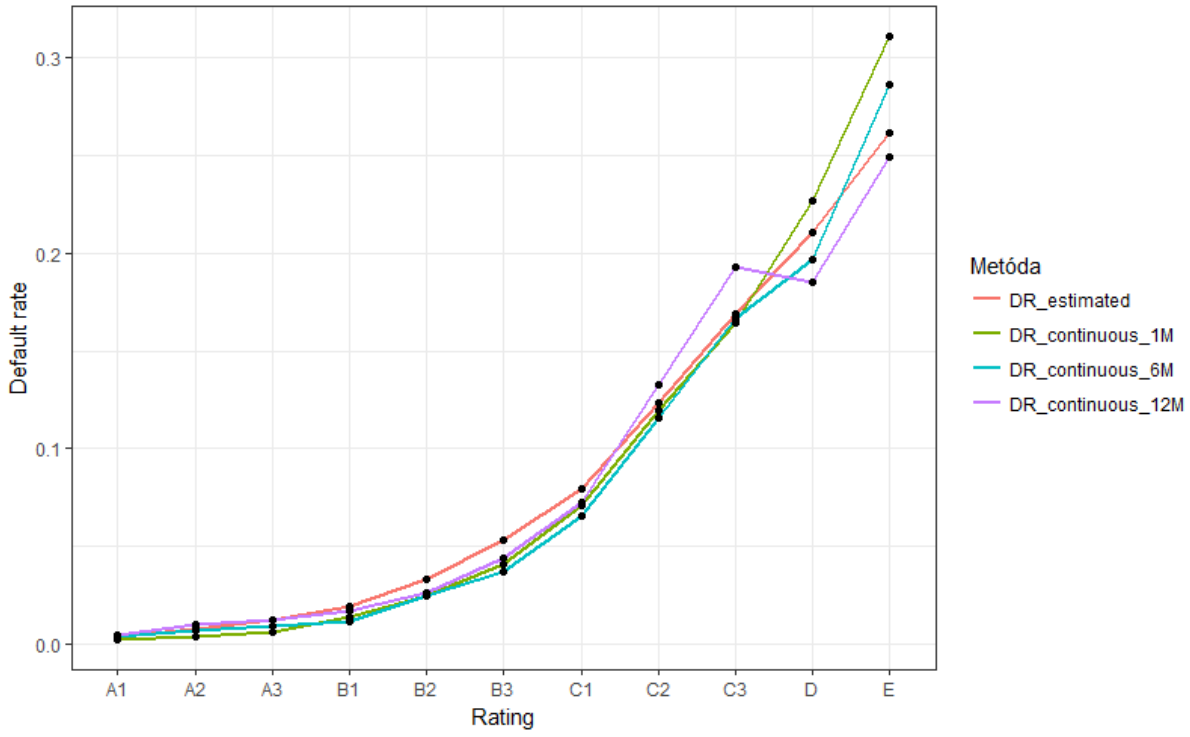


Obr. 18: Preškálovaný odhad generujúcej matice Q z diskrétnej matice ${}_6M_1$ na dobu jedného mesiaca

(zdroj: vlastné spracovanie)

Pre všetky diskrétne matice ${}_{\Delta t}M$, $\Delta t = \{1, 2, 3, 4, 6, 12\}$ sme odhadli generujúce matice Q (podľa procesu vysvetleného v tejto podkapitole), z ktorých sme vypočítali

12 mesačné miery zlyhania. Na Obr. 19 je vizualizácia troch odhadnutých 12-mesačných mier zlyhania.



Obr. 19: Ukážka odhadnutých 12 mesačných mier zlyhania pomocou spojitých Markovových reťazcov pre voľby $\Delta t = \{1, 6, 12\}$ spolu s odhadnutými mierami z reálnych dát

(zdroj: vlastné spracovanie)

Podobne ako pri diskrétnych maticiach (DTMC), sme podobnosť kriviek kvantifikovali pomocou váženého súčtu štvorcov rezíduí:

$$wRSS = \sum_{i \in R} w_i \left(DR^{12}(0, i) - \widehat{DR}_i \right)^2, \tag{31}$$

kde $\widehat{DR}_i$ sú odnadnuté pravdepodobnosti pomocou spojitých Markovových reťazcov (CTMC).

Z Obr. 19 môžeme pozorovať neprirodzený priebeh pravdepodonosti zlyhania pre spojitú migračnú maticu s $\Delta t = 12$. Odhadnutá krivka pre takýto krok je nerastúcou funkciou ratingu a preto sme sa ju rozhodli vylúčiť z porovnávania. V Tab.13 môžeme vidieť zhoršenie výsledkov pri prechode z DTMC na CTMC.

Tabuľka 13: Porovnanie hodnôt $wRSS \times 10^6$ pre metodiky DTMC a CTMC s odhadom Q matice pomocou minimalizácie euklidovskej ℓ_2 normy z rovnice (30)

Δt	1	2	3	4	6	12
DTMC	67,358	97,565	51,079	41,851	62,529	248,13
CTMC - ℓ_2	60,096	92,262	101,447	114,218	74,68	59,054

Prístup z rovnice (30) sa neosvedčil. Najlepšie DTMC migračné matice mali po prechode na spojitú verziu väčšiu chybu $wRSS$ (viď Tab. 13). Pre časový krok $\Delta t = 12$ bola spojitá verzia odhadu mier pravdepodobností zlyhania nerastúcou funkciou ratingu (viď. Obr 19). Z toho dôvodu sme za pravidlo výberu metódy odhadu Q matice vybrali tú metódu, ktorá mala najvyššiu hodnotu $\sum_{r \in R} \delta_r$ pre spojitú verziu odhadu 12 mesačných pravdepodobností zlyhania. Vyberali sme z troch metód (*diagonal adjustment, weighted adjustment* a *quasi-optimization*), ktorých dokumentáciu možno nájsť v [14]. Výsledky pri tomto prístupe sa nachádzajú v Tab. 14, 15.

Tabuľka 14: Porovnanie hodnôt $wRSS \times 10^6$ pre metodiky DTMC a CTMC s odhadom Q matice pomocou prístupu maximalizácie $\sum_{r \in R} \delta_r$

Δt	1	2	3	4	6	12
DTMC	67,358	97,565	51,079	41,851	62,529	248,13
CTMC - IS_w	60,096	71,33111	29,75711	30,20344	26,72683	84,78034
Poradie	4	5	2	3	1	6

Výsledky pomocou spojitých odhadov zlyhaní sa vylepšili v porovnaní s diskrétnymi odhadmi.

Tabuľka 15: Waldove intervaly spoľahlivosti (95%) pre 12 mesačné pravdepodobnosti zlyhania spolu s odhadnutými pravdepodobnosťami zlyhania pomocou časovo spojitých Markovových reťazcov (CTMC) s krokom Δt

CTMC				Δt					
Rating	$\widehat{PD}$	IS_W		1	2	3	4	6	12
A1	0,38%	0,28%	0,48%	0,23%	0,25%	0,44%	0,55%	0,49%	0,54%
A2	0,80%	0,63%	0,97%	0,40%	0,46%	0,77%	1,09%	0,88%	1,13%
A3	1,29%	1,10%	1,47%	0,67%	0,73%	1,05%	1,22%	1,14%	1,29%
B1	1,97%	1,74%	2,21%	1,40%	1,50%	1,98%	2,06%	1,87%	2,02%
B2	3,34%	3,01%	3,66%	2,47%	2,77%	3,63%	3,57%	3,02%	3,17%
B3	5,35%	4,94%	5,77%	4,10%	3,98%	4,72%	4,46%	4,29%	4,92%
C1	7,94%	7,39%	8,48%	7,11%	6,96%	7,10%	6,96%	7,17%	7,92%
C2	12,31%	11,53%	13,09%	11,95%	11,86%	11,69%	11,82%	12,33%	14,21%
C3	16,91%	15,81%	18,00%	16,42%	17,08%	16,90%	16,73%	17,68%	20,45%
D	21,07%	18,66%	23,47%	22,68%	23,35%	22,38%	21,50%	20,71%	20,09%
E	26,12%	22,05%	30,18%	31,08%	33,55%	31,81%	30,58%	28,92%	27,35%
F	—	—	—	42,35%	47,05%	46,04%	45,25%	44,01%	43,07%
$\sum_{r \in R} \delta_r$				3	3	7	6	8	6

4.6.2 Migračné matice závislé od MOB

V podkapitole 4.6.1 sme sa venovali maticiam, ktorým sme menili iba veľkosť migračného kroku Δt. Nazvime takéto matice statické. Statické matice neberú do úvahy podiel celkovej životnosti úveru v čase výpočtu opravných položiek. Odhadnuté migračné matice zohľadňujú akési priemerné správanie prechodov medzi ratingami.

Ďalším možným prístupom je rozdelenie dátovej sady na niekoľko časových podmnožín závislých od počtu mesiacov na účtovnej knihe (MOB). Pri výpočte opravných položiek by sme tým pádom zahrnuli časovú pozíciu úverov v splátkovom kalendári. Na Obr. 16 je zobrazený percentuálny podiel zlyhania úverov v závislosti od mesiaca životnosti a v Tab. 16 sa nachádzajú aj príslušné decily zlyhania. Nazvime takýto prístup ako nehomogénne diskrétne Markovove reťazce, skrátene NHDTMC (*non-*

homogeneous discrete time Markov chain).

Tabuľka 16: Kvantily zlyhania úverov v závislosti od počtu mesiacov na účtovnej knihe

Kvantil	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
MOB	3	5	6	7	9	11	13	15	19	25	39

Rozhodli sme sa rozdeliť časové okno na dve základné časti. Záznamy, ktoré majú $MOB \in \langle 0, 12\rangle$ a záznamy, kde $MOB > 12$. Prvý spomenutý interval budeme ďalej deliť postupne na $k = \{1, 2, 3, 12\}$ častí, na ktorých budeme odhadovať jednomesačné matice prechodu. Druhý interval nebudeme ďalej deliť a odhadneme preň jednomesačnú maticu prechodu. Takýmto procesom odhadneme postupne maticu prechodu ${}_1M_1$ pre MOB_{0-3}, MOB_{4-7}, MOB_{8-12}, MOB_{0-6}, MOB_{7-12}, MOB_{0-12} ,MOB_{13+} a takisto unikátne matice prechodov pre všetky mesiace životnosti: MOB_1, $MOB_2, \cdots MOB_{12}$. Takto vytvorené migračné matice M budeme v tejto podkapitole označovať pravým dolným indexom, napr. M_{7-12} značí jednomesačnú maticu prechodu odhadnutú na záznamoch, kde počty mesiacov životnosti $MOB \in \langle 7, 12\rangle$. Potom pre $DR^{12}(MOB = 0)$ platia nasledujúce vzťahy:

- $k = 1:$ $\quad DR^{12} = M^{12}_{0-12}$
- $k = 2:$ $\quad DR^{12} = M^{6}_{0-6} \times M^{6}_{7-12}$
- $k = 3:$ $\quad DR^{12} = M^{3}_{0-3} \times M^{3}_{4-7} \times M^{3}_{8-12}$
- $k = 12:$ $\quad DR^{12} = \prod_{i=1}^{12} M_i$

Môžeme si všimnúť, že pre $\forall k$ sme pri odhade DR^{12} násobili 12 jednomesačných migračných matíc. Postupnosť násobenia matíc je dôležitá. Podobne by sme mohli vypočítať aj $DR^{12}(MOB = t), t \neq 0$, kde by sa v závislosti od voľby k násobili matice po $MOB = 12$ a následne by sa násobilo pomocou jednomesačnej matice M_{13+} $t-$krát.

Tabuľka 17: Waldove intervaly spoľahlivosti pre 12 mesačné pravdepodobnosti zlyhania spolu s odhadnutými pravdepodobnosťami zlyhania pomocou nehomogénnych časovo dynamických diskrétnych Markovových reťazcov s k deleniami prvých 12 mesiacov životnosti úveru

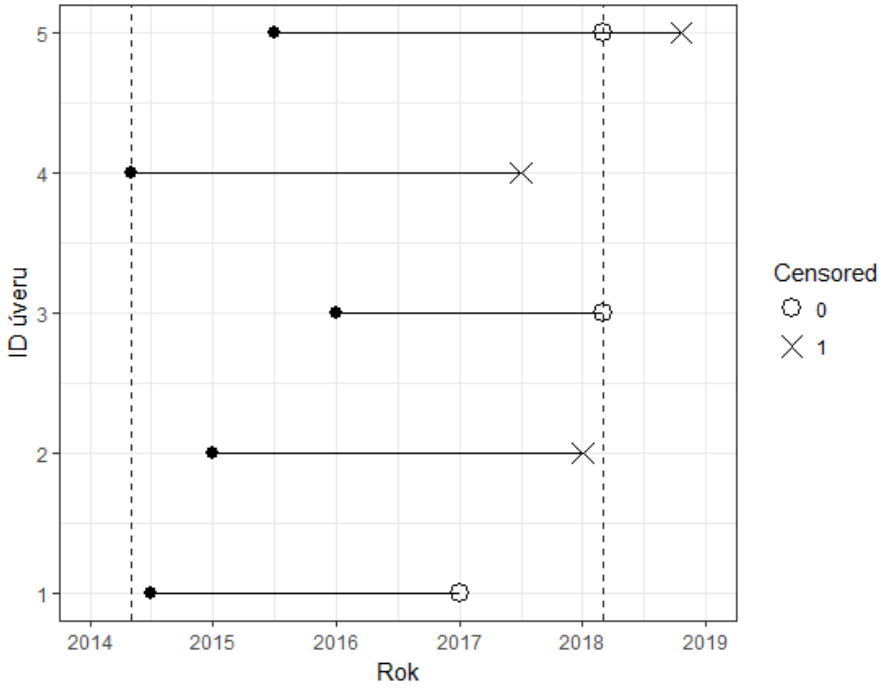
NHDTMC				k			
Rating	$\widehat{PD}$	IS_W		1	2	3	12
A1	0,38%	0,28%	0,48%	0,25%	0,40%	0,44%	0,37%
A2	0,80%	0,63%	0,97%	0,36%	0,65%	0,68%	0,73%
A3	1,29%	1,10%	1,47%	0,66%	1,10%	1,27%	1,35%
B1	1,97%	1,74%	2,21%	1,24%	2,00%	2,37%	2,73%
B2	3,34%	3,01%	3,66%	1,89%	3,24%	3,69%	3,70%
B3	5,35%	4,94%	5,77%	3,29%	4,95%	5,81%	5,88%
C1	7,94%	7,39%	8,48%	7,28%	7,47%	9,29%	10,08%
C2	12,31%	11,53%	13,09%	13,49%	13,10%	14,62%	15,68%
C3	16,91%	15,81%	18,00%	16,83%	20,99%	23,56%	23,51%
D	21,07%	18,66%	23,47%	21,28%	32,55%	37,89%	35,48%
E	26,12%	22,05%	30,18%	27,19%	47,64%	46,78%	42,81%
F	—	—	—	42,88%	56,27%	42,43%	46,76%
$\sum_{r \in R} \delta_r$				3	6	3	3

4.7 Analýza prežívania

Analýza prežívania sa zaoberá modelovaním času do nejakej udalosti (v našom prípade zlyhanie úveru), pričom sa predpokladá, že tento čas závisí od nejakých vlastností (regresorov) daného úveru. Analýza prežívania sa začala používať na odhad času do úmrtia (preto je v názve prežívanie).

Výhodou metód analýzy prežívania je práca s tzv. cenzurovanými pozorovaniami. Jedná sa o prípady, kedy nemôžeme sledovať jedinca počas celej pozorovacej doby. V našom prípade hovoríme o cenzurovaní z pravej strany, kedy vieme o konkrétnom

úvere povedať, či do konca sledovacej doby zlyhal, alebo nie. Naším cenzurovacím časom bude koniec dátovej sady (koniec časového okna), ktorú máme k dispozícii na analýzu, prípadne čas predčasného alebo riadneho splatenia úveru.



Obr. 20: Cenzurované pozorovania, pričom • značí začiatok pozorovania, ◯ predstavuje cenzurovanie a znak × zlyhanie klienta

(zdroj: vlastné spracovanie)

Dáta s ktorými pracujeme v tejto kapitole sú generované na mesačnej báze k ultimu mesiaca. Môže nastať situácia, kedy úver zlyhá napr. v strede mesiaca, avšak do času výpočtu opravných položiek (posledný deň v aktuálnom mesiaci) banka obdrží dodatočnú splátku. Túto skutočnosť v práci zanedbáme. Na Obr. 20 sa nachádzajú rôzne možné scenáre pozorovaných úverov. Pri úveroch s ID $= \{1, 3, 5\}$ hovoríme o cenzurovaných údajoch. V prípade prvého úveru môžeme bez ďalších dodatočných informácií pozorovať buď predčasné alebo riadne splatenie úveru. Pri treťom a piatom úvere máme ukončené sledovacie obdobie, kvôli čomu sa všetky pozorovania cenzurujú. Pri piatom úvere máme uvedenú aj dodatočnú životnosť úveru, počas ktorého úver zlyhal, avšak pri danom časovom okne zaznamenávame tento úver ako cenzurované pozorovanie. Pri úveroch s ID $= \{2, 4\}$ nastala sledovaná udalosť (kreditné zlyhanie) ešte pred

ukončením sledovaného časového okna.

Nech náhodná premenná T s distribučnou funkciou $F(t) = Pr(T \leq t)$ reprezentuje čas do zlyhania. Doplnkom distribučnej funkcie $F(t)$ je funckia prežívania (*survival function*) $S(t) = Pr(T > t) = 1 - F(t)$. Ďalšou základnou opisnou funkciou je tzv. riziková funkcia (*hazard function*):

$$h(t) = \lim_{\Delta t \to 0^+} \frac{Pr[(t \leq T < t + \Delta t) \mid T \geq t]}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t), \tag{32}$$

kde $f(t)$ je hustota distribučnej funkcie.

4.7.1 Kaplanov-Meierov odhad

Najznámejším neparametrickým odhadom funkcie prežitia $S(t)$ je *Kaplanov-Meierov* odhad. Predpokladajme [3], že máme k rôznych časov, v ktorých došlo k udalosti (zlyhaniu úveru). Tieto časy zoradíme tak, aby platilo $t_1 < t_2 < \cdots < t_k$. Kaplanov-Meierov odhad pre $t_1 < t < t_k$ je definovaný:

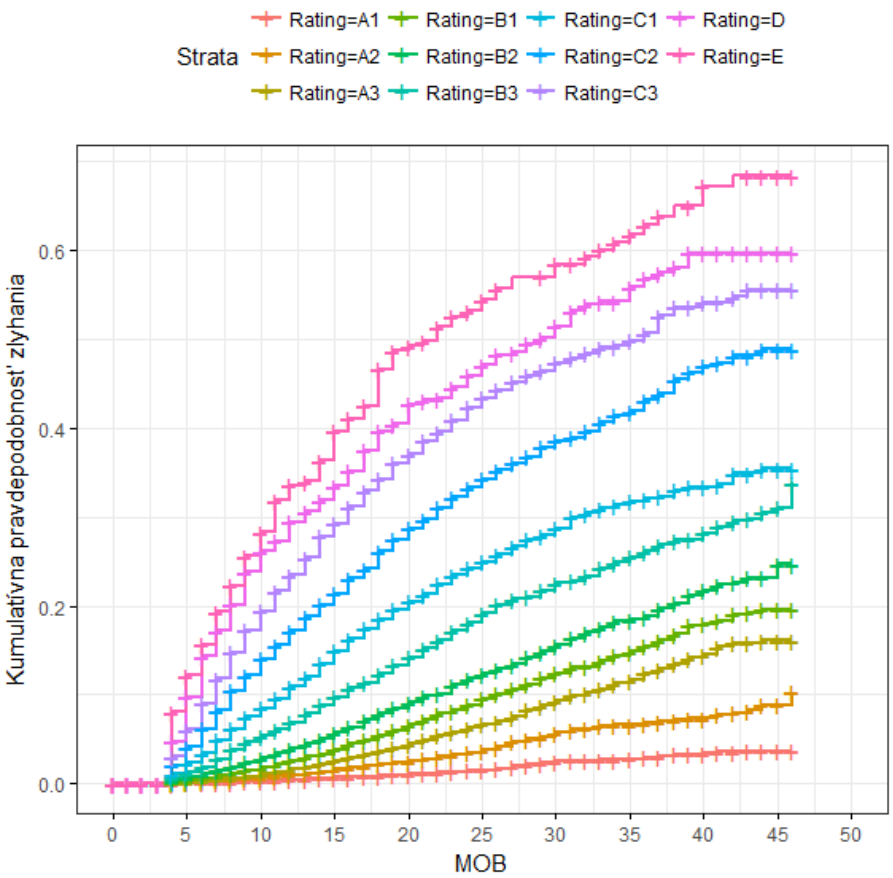
$$S_{KM}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{33}$$

s časom t_i, kedy nastalo aspoň jedno zlyhanie, d_i je počet zlyhaní v čase t_i a n_i je počet zvyšných pozorovaných úverov (nemali do času t_i zlyhanie a neboli cenzúrované).

4.8 Celoživotná miera zlyhania

Podľa štandardu *IFRS 9* sa pre úvery zaradené do *Stage 2* a *Stage 3* (pozri podkapitolu 4.2) počítajú celoživotné očakávané straty. Pri výpočte očakávaných strát (resp. opravných položiek) sa odhadujú medzimesačné pravdepodobnosti zlyhania až po očakávanú splatnosť úveru.

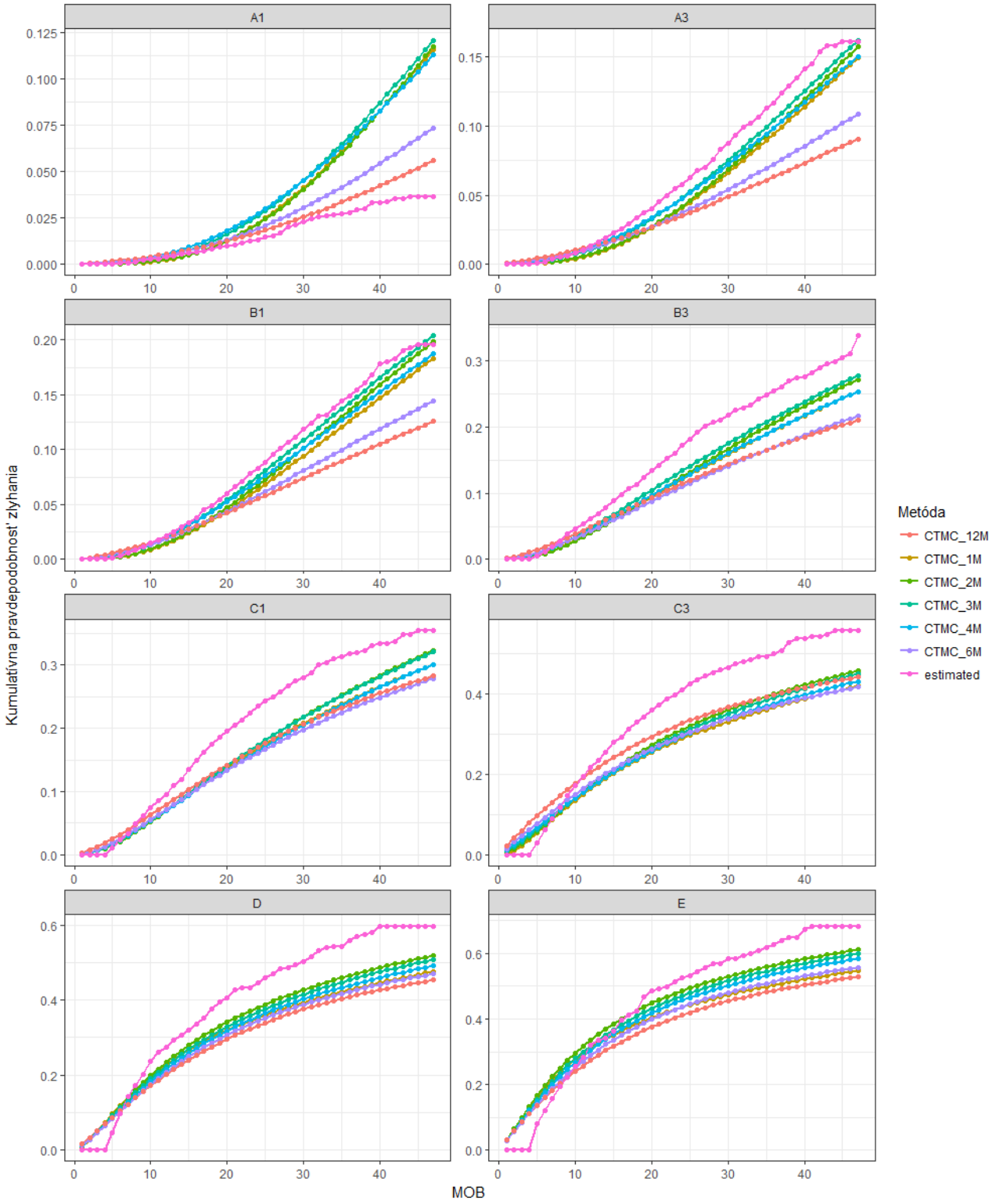
Na odhad celoživotných pravdepodobností zlyhania pre dáta, ktoré máme k dispozícii, použijeme práve Kaplanov-Meierov odhad pomocou funkcie `survfit` z doplnkového balíka `survival`. Kumulatívnu pravdepodobnosť zlyhania vypočítame ako $F(t)$ = 1 - $S_{KM}(t)$. Odhady kumulatívnych pravdepodobností zlyhania pre jednotlivé



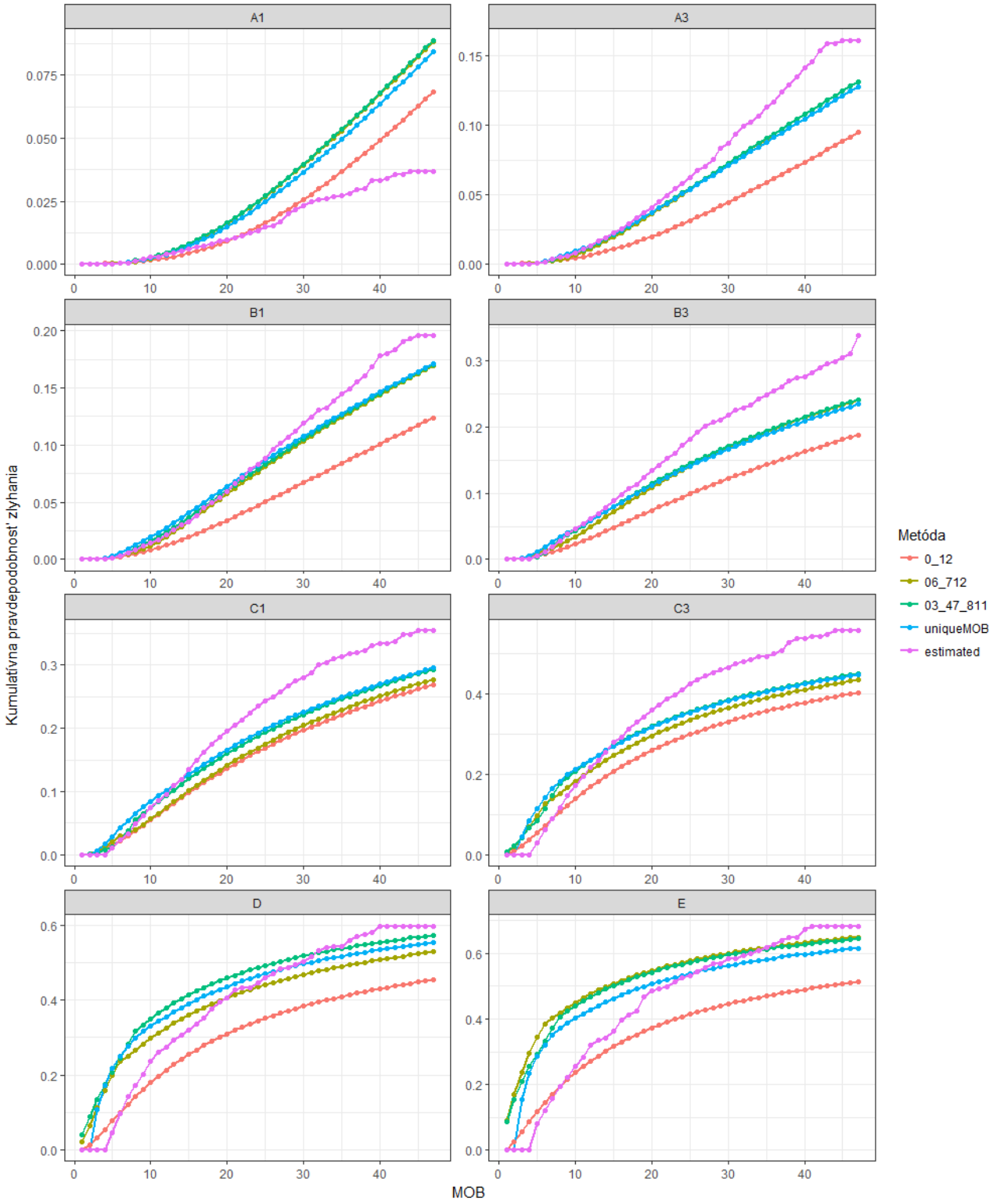
Obr. 21: Odhad kumulatívnej pravdepobobnosti zlyhania $F(t) = 1 - S_{KM}(t)$ pre rôzne ratingy pomocou Kaplanovho-Meierovho odhadu funkcie prežitia

(zdroj: vlastné spracovanie)

ratingy sa nachádzajú na Obr. 21. Keďže naša dátová sada obsahuje pozorovania najvýš pre $MOB = 47$, pokúsime sa porovnať odhady pomocou statických (CTMC) a dynamických (NHDTMC) migračných matíc na tomto skrátenom intervale. Postup odhadu celoživotných pravdepodobností zlyhania pomocou Markovových reťazcov je podobný, ako pri odhade 12 mesačných mier zlyhania v podkapitole 4.6. Rozdiel je len v počte mocnín jednotlivých migračných matíc. Pre kumulatívne odhady mier zlyhaní v MOB-e k platia vzťahy z rovnice (22) s $t_0 = 0$ pre CTMC, a podobne proces odhadu pomocou migračných matíc závislých od životnosti úveru (NHDTMC) popísaný v časti 4.6.2. Kumulatívne odhady podľa metód CTMC a NHDTMC pre niektoré ratingy sú zobrazené na Obr. 22, 23.



Obr. 22: Porovnanie metód statických migračných matíc (CTMC) a odhadnutej kumulatívnej pravdepodobnosti zlyhania $F(t) = 1 - S_{KM}(t)$ pre niektoré ratingové skupiny (zdroj: vlastné spracovanie)



Obr. 23: Porovnanie metód dynamických migračných matíc (NHDTMC) a odhadnutej kumulatívnej pravdepodobnosti zlyhania $F(t) = 1 - S_{KM}(t)$ pre niektoré ratingové skupiny (zdroj: vlastné spracovanie)

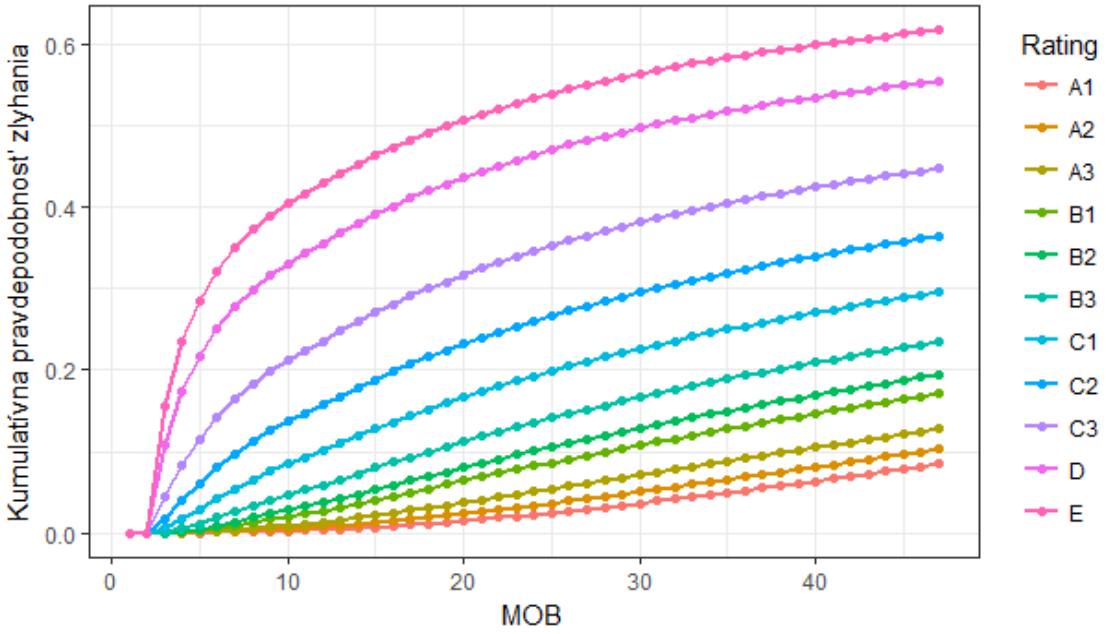
Z Obr. 22, 23 môžeme pozorovať rôznorodé výsledky. Vo väčšine prípadov odhadnuté krivky pomocou migračných matíc dobre fitujú hodnoty na začiatkoch životnosti úverov. Pre málo rizikový rating A1 sú odhadnuté hodnoty nadhodnotené, ratingy A2 až B2 sú pomerne dobre odhadnuté pri oboch metódach CTMC aj NHDTMC, pre stredne rizikové ratingy B3 až C3 sú naopak odhady podhodnotené a najviac rizikové ratingy sú opäť pomerne dobre odhadnuté.

Na porovnanie odhadnutých kriviek sme použili metodiku pomocou intervalov spoľahlivosti (podobne ako pri odhade 12 mesačných pravdepodobností). Výsledky sa nachádzajú v Tab. 18.

Tabuľka 18: Porovnanie odhadov celoživotných pravdepodobností pre metodiky CTMC a NHMTMC pomocou prístupu maximalizácie $\sum_i \sum_{r \in R} \delta_{i,r}$ pre dostupné MOB_i,$i \in \{1, 2, \ldots, 47\}$

Metóda	CTMC - Δt						NHDTMC - k			
Rating	1	2	3	4	6	12	1	2	3	12
A1	9	10	7	0	2	6	18	10	10	14
A2	17	15	17	3	38	23	4	31	37	43
A3	4	5	9	9	5	3	1	4	17	15
B1	3	11	31	12	5	5	2	15	18	13
B2	3	24	41	38	5	6	1	21	19	13
B3	1	1	2	2	1	1	1	1	8	7
C1	1	1	1	2	2	1	1	1	6	6
C2	2	1	2	1	2	2	2	1	3	6
C3	1	1	2	2	2	3	1	4	4	6
D	3	3	3	2	2	2	2	15	23	28
E	10	33	27	15	12	9	10	25	27	31
$\sum_{i}^{47} \sum_{r \in R} \delta_{i,r}$	54	105	142	86	76	61	43	128	172	182

Z výsledkov v Tab. 18 môžeme konštatovať, že najlepšie výsledky mala metóda NHDTMC s rôznymi migračnými maticami pre každý mesiac v prvom roku životnosti úveru ($k = 12$).



Obr. 24: Vizuálna ukážka odhadnutých kumulatívnych pravdepodobností zlyhania pre metódu NHDTMC s $k = 12$

(zdroj: vlastné spracovanie)

Fakt, že odhady pomocou migračných matíc fitujú lepšie najmä pre počiatočné mesiace životnosti sa dá vysvetliť nasledovne. Odhady $F(t) = 1 - S_{KM}(t)$ sú vypočítané v závislosti od aplikačných ratingov. Takéto odhady predstavujú postupné „prežívanie" úverov v závislosti od prvotných ratingových ohodnotení klientov. Markovove migračné matice však zohľadňujú prechody medzi behaviorálnymi ratingami, tzn. zohľadňujú aj časovo premenlivú distribúciu ratingov. Odhady prežívania aplikačných ratingov sú však dôležité pre banku a ešte viac dôležité pre klientov. Na základe týchto mier zlyhaní sú klientom vypočítané rizikové prirážky k úroku. Úroková prirážka je známa pod termínom *risk marža* a slúži na krytie očakávaných strát. Voľba metodiky odhadu pravdepodobnosti zlyhania môže byť závislá aj na strategickom pláne banky. Banky môžu byť viac konzervatívne, kvôli čomu volia metódy s vyšším odhadom PD a sú pripravené aj na stresové scenáre. Podľa štandardu *IFRS 9* [6] sa navyše marginálne PD hodnoty upravujú o tzv. *forward looking element*, ktorý zohľadňuje budúce ekonomické podmienky vplývajúce na výpočet očakávaných strát, čím sa navyšuje hodnota marginálnych PD o násobok nejakej konštanty $(1 + c)$.

4.8.1 Coxov model

Posledným použitým modelom v tejto diplomovej práci je Coxov model. Tento model patrí do triedy modelov analýzy prežívania. V tejto práci by sme chceli len navrhnúť jeden možný model na odhad kreditného zlyhania pre úvery so životnosťou viac ako jeden rok. Chceli by sme tým poukázať na možnosti vytvorenia viackriteriálnych modelov založených na behaviorálnych ratingoch, alebo na premenných, ktoré popisujú charakteristiky úveru v čase životnosti. Coxov model [7] má nasledujúci tvar:

$$h_i(t) = h_0(t) \times \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_p x_{ip}), \tag{34}$$

kde $h(t)$ je riziková funkcia v čase t, $h_0(t)$ predstavuje referenčnú rizikovú funkciu, ktorá obvykle obsahuje informáciu o rizikovosti pre nulové hodnoty regresorov x_1 až x_p (prípadne prvé hodnoty pri kategoriálnych premenných). V literatúre [7] môžeme nájsť aj iný spôsob zápisu modelu (34):

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \tag{35}$$

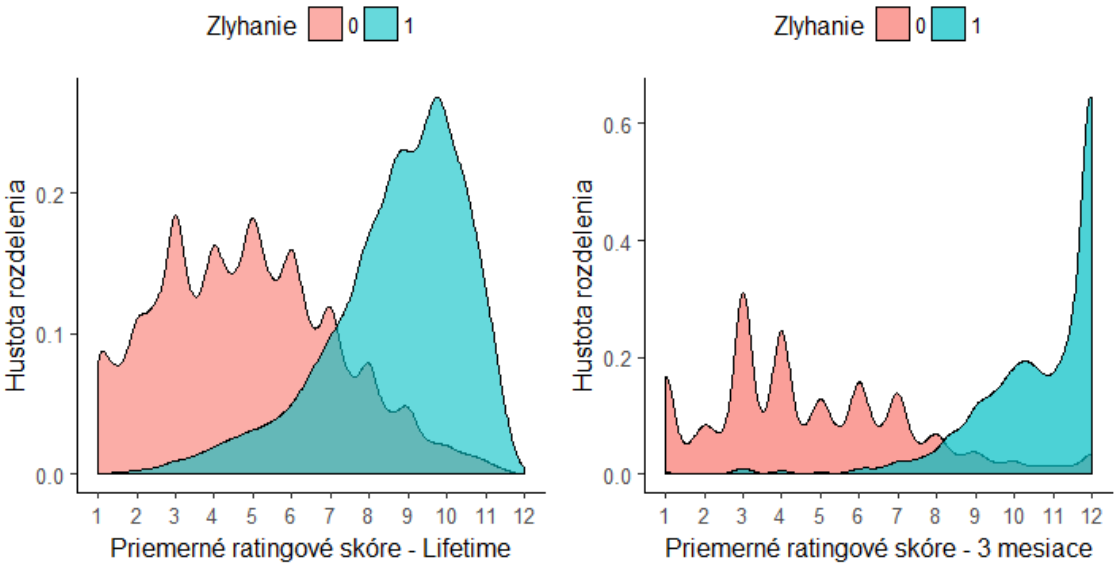
V doterajších modeloch v tejto kapitole sme na odhad pravdepodobnosti zlyhania používali migračné matice odhadnuté výlučne pomocou behaviorálneho ratingu. Snahou vylepšiť modely sme pridali predpoklad, že úvery zlyhávajú najmä v prvých 12-tich mesiacoch životnosti (NHDTMC). V týchto modeloch boli zanedbané niektoré ukazovatele, ktoré by mohli odzrkadliť zvýšené riziko zlyhania.

Majme dvá úvery, ktoré majú aplikačné ratingy A1 a D, ktorých aktuálna životnosť je 20 mesiacov. Predstavme si situáciu, že dané úvery si udržiavali svoje behaviorálne ratingy na týchto hladinách, avšak za posledných pár mesiacov sa úveru s aplikačným ratingom A1 pomerne rýchlo navýšil behaviorálny rating na úroveň D. Za daných situácii sa pre dva úvery vypočíta rovnaká pravdepodobnosť zlyhania (podľa prístupu CTMC a NHDTMC). Prirodzene by sme povedali (podľa nášho názoru), že úver, ktorého behaviorálny rating sa zvýši v krátkej dobe je rizikovejší, ako úver ktorý si drží stabilne rovnakú ratingovú úroveň už dlhšiu dobu. Ďalej sme do modelu pridali aj informáciu o kontraktuálnom počte splátok a o aktuálnom podiely nesplatenej istiny (závislosť možno vidieť na Obr. 15).

Na kvantifikovanie priemernej rizikovosti sme použili skóringovú škálu, ktorá ratingom $r \in R$ priradí číselné hodnoty nasledovne:

$$\{A1, A2, A3, B1, B2, B3, C1, \ldots, E, F\} \rightarrow \{1, 2, 3, 4, 5, 6, 7, \ldots, 11, 12\}.$$

Pomocou tejto škály sme následne vypočítali skóre podľa behaviorálnych ratingov jednotlivých úverov. Následne sme vypočítali priemernú skóringovú hodnotu pre celú životnosť úveru až po dobu cenzúrovania. Za cenzúrovanie sa považujú dva prípady. Prvý prípad je zlyhanie úveru. Druhý prípad je buď splatenie úveru (riadne, predčasné, konsolidácia), alebo ukončenie sledovacieho obdobia (posledný dostupný mesiac v dátovej sade).



Obr. 25: Vizualizácia rozdelenia priemerných skóringových hodnôt pre celoživotné resp. posledných 3 mesiacov pred cenzurovaním. Za čas cenzúrovania sa považuje čas zlyhania, alebo posledný časový rez dostupný o úvere

(zdroj: vlastné spracovanie)

Na Obr. 25 sa nachádza vizualizácia hustôt rozdelení priemerných skóre pre kategóriu zlyhaných a cenzurovaných úverov. Zaveďme značenie premenných:

- `r_apl`: Aplikačný rating
- `istina`: Podiel zostávajúcej istiny

- `splatky`: Počet kontraktuálnych splátok
- `score_Life`: Priemerné skóre počas životnosti pred cenzurovaním
- `score_3M`: Priemerné skóre posledných troch mesiacov pred cenzurovaním
- `pomer_3M_Life`: Podiel `score_3M` a `score_Life`

Pomocou týchto premenných sme navrhli jeden z možných modelov, ktorý mal nasledujúcu štruktúru:

$$h(t) = h_0(t) \times \exp(\beta_1 \cdot r\_apl \ + \ \beta_2 \cdot istina \ + \ \beta_3 \cdot splatky \ + \ \beta_4 \cdot score\_Life \ + \ \beta_5 \cdot score\_3M \ + \ \beta_6 \cdot pomer\_3M\_Life).$$

Na odhad parametrov Coxovho modelu používame funkciu `coxph` z doplnkového balíka `survival`. Premenné navrhnutého Coxovho modelu sú okrem aplikačného ratingu všetky číselné. Pre kategoriálnu premennú `r_apl` sa vytvoria umelé tzv. *dummy* premenné.

Tabuľka 19: Odhady koeficientov spolu so sumarizačnou tabuľkou pre Coxov model

Premenné	$\widehat{\beta}$	$\exp(\widehat{\beta})$	Std. Err.	z	P-value	Signif.
r_apl : A2	-0,11	0,89	0,10	-1,11	0,2665	
r_apl: A3	0,22	1,25	0,09	2,51	0,0119	*
r_apl: B1	0,37	1,44	0,09	4,23	$2{,}30\times10^{-5}$	***
r_apl: B2	0,60	1,83	0,09	6,98	$3{,}00\times10^{-12}$	***
r_apl: B3	0,85	2,34	0,09	9,9	$<2\times10^{-16}$	***
r_apl: C1	1,11	3,02	0,09	12,86	$<2\times10^{-16}$	***
r_apl: C2	1,32	3,73	0,09	15,19	$<2\times10^{-16}$	***
r_apl: C3	1,55	4,69	0,09	17,59	$<2\times10^{-16}$	***
r_apl: D	1,74	5,68	0,10	17,56	$<2\times10^{-16}$	***
r_apl: E	1,85	6,38	0,11	16,48	$<2\times10^{-16}$	***
r_apl: F	2,18	8,82	0,71	3,06	0,0022	**
istina	0,55	1,73	0,01	65,26	$<2\times10^{-16}$	***
splatky	-0,003	0,997	0,0003	-9,51	$<2\times10^{-16}$	***
score_Life	-0,13	0,88	0,02	-6,68	$2{,}30\times10^{-11}$	***
score_3M	0,53	1,70	0,02	32,68	$<2\times10^{-16}$	***
pomer_3M_Life	0,84	2,31	0,09	9,00	$<2\times10^{-16}$	***

Z výsledkov modelu (Tab. 19) sa ukázalo, že nie je signifikantný prechod z aplikačného ratingu A1 na aplikačný rating A2. Rozhodli sme sa ratingy A1, A2 a A3 zlúčiť do ratingu A pre konečný model, ktorého výsledky sa nachádzajú v Tab. 20.

Tabuľka 20: Odhady koeficientov spolu so sumarizačnou tabuľkou pre Coxov model s upravenými hodnotami aplikačných ratingov

Premenné	$\widehat{\beta}$	$\exp(\widehat{\beta})$	Std. Err.	z	P-value	Signif.
r_apl: B1	0,26	1,30	0,05	5,23	$1{,}7\times10^{-7}$	***
r_apl: B2	0,50	1,64	0,05	10,22	$<2\times10^{-16}$	***
r_apl: B3	0,74	2,10	0,05	15,95	$<2\times10^{-16}$	***
r_apl: C1	1,00	2,71	0,05	21,49	$<2\times10^{-16}$	***
r_apl: C2	1,20	3,34	0,05	25,49	$<2\times10^{-16}$	***
r_apl: C3	1,43	4,20	0,05	29,01	$<2\times10^{-16}$	***
r_apl: D	1,62	5,08	0,07	21,12	$<2\times10^{-16}$	***
r_apl: E	1,74	5,70	0,09	20,26	$<2\times10^{-16}$	***
r_apl: F	2,06	7,88	0,71	2,91	0,0036	**
istina	0,54	1,71	0,01	69,73	$<2\times10^{-16}$	***
splatky	-0,003	0,997	0,0003	-9,42	$<2\times10^{-16}$	***
score_Life	-0,13	0,87	0,02	-7,00	$2{,}6\times10^{-12}$	***
score_3M	0,54	1,71	0,02	33,74	$<2\times10^{-16}$	***
pomer_3M_Life	0,79	2,20	0,09	1,84	$<2\times10^{-16}$	***

Výsledky v Tab. 19 ukazujú signifikantnosť nami navrhnutými doplnkovými premennými. Pomocou metód analýzy prežívania by sa dalo vytvoriť široké spektrum modelov s rôznymi vstupnými premennými. Dáta analyzované v tejto práci poskytujú krátke časové okno, na ktorom je pomerne málo ukončených úverov. Pre dlhšie časové okno by bola možnosť vytvoriť oveľa sofistikovanejšie predikčné modely, ktoré by v sebe zahŕňali naozajstné miery zlyhaní, nakoľko by obsahovali len ukončené úvery. Validácia takýchto modelov by mohla prebiehať tak, že by sme podľa opisných premenných rozdelili všetky úvery na niekoľko podskupín, ktoré by v sebe obsahovali podobné úvery na základe vstupných premenných. Pre tieto podskupiny by sa následne odhadla skutočná miera zlyhania, ktorá by sa porovnala s modelovým odhadom (cross-validácia: Alg. 4) a výkonnosť modelu by sa kvantifikovala napr. pomocou *ROC* krivky, prípadne súčtom štvorcov odchýlok medzi skutočnými a modelovými odhadmi mier zlyhaní.

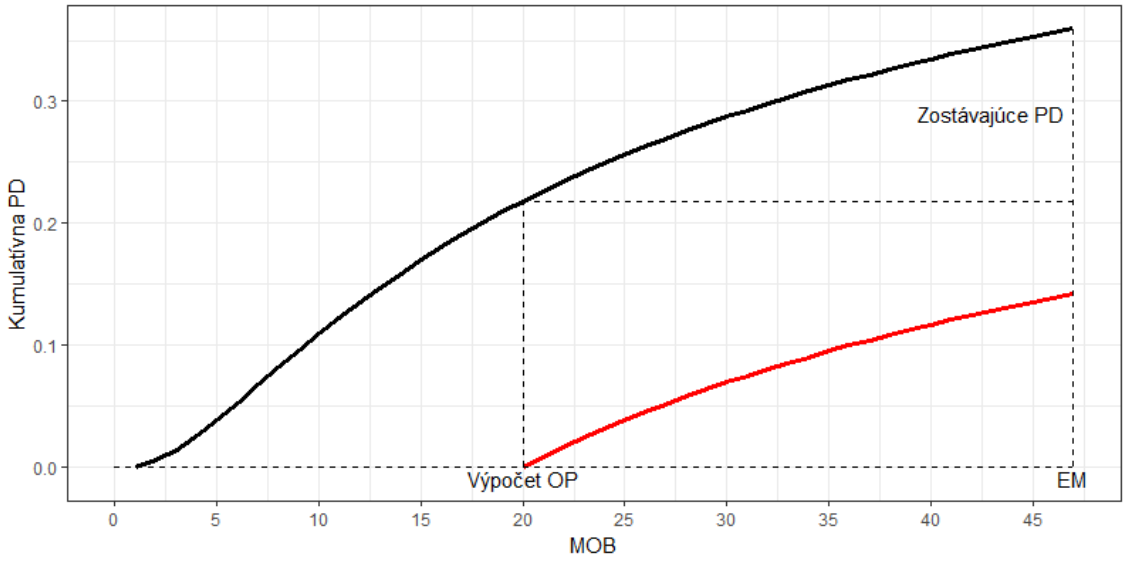
Výpočet mier zlyhaní pre vstupné hodnoty regresorov X do času Δt sa počíta pomocou vzťahu $F(\Delta t|X) = 1-S(\Delta t|X)$. Konkrétnejší výpočet uvádzame v podkapitole 4.9.

4.9 Výpočet medzimesačných pravdepodobností zlyhania

Na záver tejto kapitoly by sme radi ozrejmili spôsob výpočtu medzimesačných a celoživotných pravdepodobností pre úvery v ľubovolných fázach životnosti (MOB-och).

CTMC a NHDTMC

Pre metódy CTMC a NHDTMC sú k dispozícii dva možné prístupy. Prvým z nich je odhad počiatočných celoživotných pravdepodobností zlyhania pre všetky ratingové skupiny. Následne by sa v závislosti od behaviorálneho ratingu a aktuálneho mesiaca životnosti vypočítali zvislé zmeny v danej celožiovotnej PD krivke. Vizualizáciu tejto myšlienky uvádzame na Obr. 26.



Obr. 26: Vizuálna ukážka výpočtu celoživotných, resp. medzimesačných pravdepodobností zlyhania pre rôzne fázy životnosti úveru pomocou počiatočnej celoživotnej *PD* krivky

(zdroj: vlastné spracovanie)

Na Obr. 26 sa nachádza názorná ukážka výpočtu opravnej položky (skr. OP) pre úver, ktorý má životnosť 20 mesiacov s očakávanou splatnosťou o 27 mesiacov. Pre tento

úver sa vypočíta nový behaviorálny rating, pre ktorý zoberieme príslušnú PD krivku. Novou zostávajúcou PD krivkou je krivka vyznačená červenou farbou. Na výpočet očakávaných strát sa postupne zoberú jednomesačné zvislé zmeny PD kriviek, ktoré predstavujú marginálne pravdepodobnosti zlyhania práve v mesiaci $t+1$, $PD_{t,t+1}$.

Druhý prístup je viac „Markovovsky". Tento prístup neberie do úvahy tvar počiatočnej PD krivky. Pre úver v ľubovoľnej fáze životnosti sa začnú počítať odznova pravdepodobnosti prechodov do stavu zlyhania pomocou násobenia migračných matíc. Jediný rozdiel medzi metódami CTMC a NHDTMC je ten, že pri CTMC sa vždy násobí jednou maticou, zatiaľčo pri NHDTMC sa môže násobiť rôznymi maticami pre rôzne životnosti úverov (závislé od voľby počtu delenia k). Tvar PD krivky pre CTMC metódu by bol stále rovnaký a posúval by sa len v čase. Pre medzimesačné pravdepodobnosti použijeme vzťah z rovnice (23), t. j. vypočítame kumulatívne pravdepodobnosti pre začiatok a koniec mesiaca pre ktorý chceme vypočítať marginálnu pravdepodobnosť a tieto hodnoty od seba odpočítame.

Coxov model

Pre Coxov model je proces výpočtu založený na odhade kumulatívnej miery rizikovej funkcie $H_0(t)$ a od vektora vstupných hodnôt X. Z rovnice (32) sa dajú ďalej odvodiť vzťahy:

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t h(u)du\right) \\ &= \exp\left(-\int_0^t h_0(u)\mathrm{e}^{\beta X}du\right) \\ &= \exp\left(-H_0(t)\right)^{\exp(\beta X)}, \end{aligned} \tag{36}$$

kde $H_0(t) = \int_0^t h_0(u)\,du$ je základná kumulatívna riziková funkcia v čase t, ktorá sa dá odhadnúť pomocou funkcie `basehaz` v balíku `survival`. Na odhad pravdepodobnosti v čase t do času $t+\Delta t$ použijeme vzťah:

$$PD_{t,t+\Delta t} = 1 - S(\Delta t),$$

kde $S(\Delta t)$ vypočítame zo vzťahu (36).

Záver

V rámci tejto diplomovej práce sme publikovali výsledky získané počas nášho výskumného procesu. Zaoberali sme sa modelovaním kreditného zlyhania na portfóliu retailových úverov. K dispozícii sme mali dáta o krátkodobých spotrebných úverov slovenských klientov.

V prvých dvoch kapitolách sme prezentovali teoretické základy zovšeobecnených lineárnych modelov v kombinácii s regularizačnými metódami. Následne sme predstavili možnosti klasifikácie pomocou klasifikačných stromov a náhodných lesov v prípade nevyváženej modelovacej premennej, ktorá značne komplikuje klasifikačnú úlohu.

Tento štatistický aparát nám poslúžil v tretej kapitole, ktorá bola venovaná krátkej opisnej štatistike úverových dát, predstaveniu implementovaných algoritmov, pomocou ktorých sme vykonávali klasifikáciu a následne sme porovnávali jednotlivé modely podľa úspešnosti klasifikácie. Zamerali sme sa na úspešnosť klasifikátorov podľa AUC a KS štatistiky. Navrhli sme program v softvéri R, pomocou ktorého sme vytvárali, vyhodnocovali a porovnávali jednotlivé modely. Ukázalo sa, že regularizačné metódy môžu byť štatisticky signifikantne lepšie (viď. Tab. 4), ako veľmi využívaná logistická regresia, ktorá je vo väčšine kredit skóringových modeloch implementovaná. Tieto metódy môžu zrýchliť proces tvorby scoringového modelu v prípade modelovania s väčším počtom premenných. Bežnou praxou je manuálna analýza korelácie medzi regresormi a následná redukcia počtu opisných premenných. Azda najväčšou výhodou regularizačných metód je pomerne jednoduchá adaptácia týchto modelov do praxe, nakoľko interpretácia modelov je rovnaká, ako pri zaužívanej logistickej regresii.

To sa o stromových modeloch povedať nedá a častokrát sú analytici obmedzovaní modelmi s jednoduchou interpretáciou. V práci sme sa však venovali aj týmto modelom a pri porovnaní jedno-stromových modelov sme hľadali možnosti vylepšenia predikčných algoritmov nastavením viacerých možných vstupných parametrov, ktoré pomoholi klasifikovať pomerne vysoko nevyváženú cieľovú premennú. Pri všetkých stromových metódach sme optimalizovali vstupné parametre cez vektor alebo mriežku hodnôt, ale do úvahy sme brali aj viaceré rozdelenia do trénovacích a testovacích

dátových sád. Porovnanie výsledkov optimálnych modelov pre jednotlivé metódy sú vizualizované na Obr. 13. Predikčnou silou boli najlepšie náhodné lesy vo všetkých smeroch (najvyššia priemerná hodnota a najnižšia variancia hodnôt AUC a KS spomedzi všetkých stromových modelov).

V ďalšej časti sme sa venovali modelovaniu kreditného rizika na už schválených úveroch. V rámci nového štandardu *IFRS 9* sme sa venovali odhadu 12 mesačných a celoživotných pravdepodobností zlyhania pomocou Markovových reťazcov. Za pomoci behaviorálnych ratingových tried sme sa snažili odhadnúť migračné matice medzi jednotlivými ratingami a tým odhadnúť pravdepodobnosti zlyhania. Venovali sme sa najskôr odhadu diskrétnych migračných matíc (DTMC) s rôznymi krokmi Δt. Vhodnosť modelov sme posudzovali pomocou dvoch prístupov. Jeden prístup bol založený na váženej regresii, kde váhy predstavovali percentuálnu početnosť zastúpenia jednotlivých ratingových tried. Druhý prístup sledoval počet ratingových tried, pre ktoré boli správne určené miery zlyhania v rámci 95%-ných intervalov spoľahlivosti.

Pre odhad mesačných pravdepodobností zlyhania sme pre diskrétne matice odhadli spojité verzie migračných matíc pomocou spojitých Markovových reťazcov (CTMC), vďaka ktorým sme dostali hladké PD krivky na akúkoľvek dobu. Pre odhad spojitých verzií migračných matíc sme použili vlastnú metódu, ktorá značne vylepšila odhadnuté PD hodnoty (pozri Tab. 13, 14).

Z dátovej analýzy vyplynulo, že väčšina úverov zlyháva už v prvom roku životnosti. Vytvorili sme modely (NHDTMC), ktoré špecificky sledujú práve správanie klientov počas prvého roku životnosti. Pri polročnom rozdelení (pozri Tab. 17) model pomerne dobre opisoval správanie prvotných aplikačných ratingov A1 až C2 po dobu 12 mesiacov.

Na odhad celoživotných mier zlyhaní nám poslúžil neparametrický Kaplanov-Meierov odhad. Pre celoživotné odhady s prístupom migračných matíc sa najviac osvedčili nehomogénne Markovove reťazce s unikátnymi migračnými maticami pre každý mesiac životnosti úveru (do jedného roka). Na záver práce sme navrhli aj možnosť modelovania kreditného zlyhania pomocou metódy analýzy prežívania. Konkrétne sme použili Co-

xov model, vďaka ktorému sme vedeli vytvoriť viackriteriálny model, ktorý obsahoval nami vytvorené premenné zohľadňujúce aj iné faktory rizika, ktoré nie sú obsiahnuté v behaviorálnom skóringu. Tieto premenné sa ukázali byť signifikantné (viď Tab. 20) a naznačovali dobré predikčné schopnosti. Problematická bola dátová sada, na ktorej by sme nevedeli vykonať rozumné testovanie výkonnosti.

Práca je prínosná najmä z praktickej stránky, nakoľko obsahuje analýzu rôznych metód na dátach o slovenských klientoch a výsledky analýz sú použiteľné pre prípadnú ipmlementáciu do metodík výpočtu opravných položiek podľa štandardu *IFRS 9*. Nami vytvorené programové kódy sú ľahko upraviteľné pre ďalšie modelovanie. Ďalej je práca určená tým, ktorí by sa radi dozvedeli niečo o základnej metodike výpočtu očakávaných strát podľa finančného štandardu *IFRS 9*. Na záver by sme radi skonštatovali, že naša diplomová práca splnila stanovené ciele.

Zoznam použitej literatúry

[1] AGRESTI, A.: *Categorical Data Analysis, Second Edition*, John Wiley Sons, Inc., 2002.

[2] BLUHM, C., OVERBECK, L., WAGNER, C.: *An Introduction to Credit Risk Modeling*, Chapman Hall, 2003.

[3] BRUNEL, V.: *Lifetime PD Analytics for Credit Portfolios: A Survey*, 2016., Dostupné na adrese (21.10.2017):
http://dx.doi.org/10.2139/ssrn.2857183

[4] CICHOSZ, P.: *Data Mining Algorithms: Explained Using R*, Wiley, 2015.

[5] DOBSON, A. J.: *An Introduction to Generalized Linear Models, Second Edition*, Chapman Hall/CRC, 2001.

[6] ERNST YOUNG: *Impairment of financial instruments under IFRS 9*, 2014, Dostupné na adrese (1.10.2017):
http://www.ey.com/us/en/SearchResults?query=IFRS+9

[7] FOX, J., WEISBERG, J.: *Cox Proportional-Hazards Regression for Survival Data in R,* , 2011., Dostupné na adrese (21.10.2017):
https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf

[8] GUNNVALD, R.: *Estimating Probability of Default Using Rating Migrations in Discrete and Continuous Time*, 2014, Dostupné na adrese (1.4.2018):
https://www.math.kth.se/matstat/seminarier/reports/M-exjobb14/140908.pdf

[9] HANLEY, A. J., MCNEIL, J. B.: *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology vol. 143, 1982, Dostupné na adrese (1.4.2018):
https://pubs.rsna.org/doi/10.1148/radiology.143.1.7063747

[10] HARMAN, R.: *Multivariate Statistical Analysis, Selected lecture notes*, FMFI UK, Bratislava, 2018, Dostupné na adrese (1.4.2018):
http://www.iam.fmph.uniba.sk/ospm/Harman/VSAp.pdf

[11] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J.: *The Elements of Statistical Learning: data mining, inference and prediction, Second Edition*, Springer, 2009, Dostupné na adrese (21.10.2017):
https://web.stanford.edu/ hastie/Papers/ESLII.pdf

[12] JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R.: *An Introduction to Statistical Learning with Applications in R*, Springer, 2013, Dostupné na adrese (13.03.2018):
http://www-bcf.usc.edu/ gareth/ISL/index.html

[13] KUHN, M., JOHSON, K.: *Applied Predictive Modeling*, Springer, 2013.

[14] PFEUFFER, M., JOHSON, K.: *ctmcd: An R Package for Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time Data*, The R Jounral Vol. 9/2, 2017, Dostupné na internete (14.04.2018):
https://journal.r-project.org/archive/2017/RJ-2017-038/RJ-2017-038.pdf

[15] TEREK, M., HORNÍKOVÁ, A., LABUDOVÁ, V.: *Hĺbková analýza údajov*, Iura edition, 2010.

[16] THOMAS, L.C., EDELMAN, D.B., CROOK, J.N.: *Credit Scoring and Its Applications*, SIAM, 2002.

Doplnkové balíky v programe R

[17] BATES D.,MAECHLER M.: *Matrix: Sparse and Dense Matrix Classes and Method*, Dostupné na adrese:
https://cran.r-project.org/web/packages/Matrix/Matrix.pdf

[18] KUHN, M.: *Caret: Classification and Regression Training*, 2017, Dostupné na adrese:
https://CRAN.R-project.org/package=caret

[19] PETERS, A.: *Ipred: Improved Predictors*, 2017, Dostupné na adrese: `https://CRAN.R-project.org/package=ipred`

[20] R Core Team (2017): *R: A Language and Environment for Statistical Computing*,R Foundation for Statistical Computing, Rakúsko, Dostupné na adrese: `https://www.R-project.org/`

[21] THERNEAU T.,ATKINSON B.,RIPLEY B.: *Rpart: Recursive Partitioning and Regression Trees*, Dostupné na adrese: `https://CRAN.R-project.org/package=rpart`

[22] THERNEAU T.: *Survival Analysis*, Dostupné na adrese: `https://cran.r-project.org/web/packages/survival/survival.pdf`

[23] WICKHAM, H.: *Tidyverse*, 2017, Dostupné na adrese: `https://CRAN.R-project.org/package=tidyverse`