Vysoká škola ekonomická v Praze

Fakulta financí a účetnictví

Katedra bankovnictví a pojišťovnictví

Studijní obor: Finanční inženýrství

Diplomová práce

Tvorba skóringového modelu

Autor: Bc. Vladislav Hošek

Vedoucí práce: Ing. Milan Fičura, Ph.D.

Čestné prohlášení:

Prohlašuji, že jsem diplomovou práci na téma „Tvorba skóringového modelu“ vypracoval samostatně a všechnu použitou literaturu a podkladové materiály jsem řádně označil a uvedl na konci práce v seznamu.

V Praze dne ……. .....

Podpis studenta

Poděkování:

Chtěl bych poděkovat mému vedoucímu diplomové práce Ing. Milanovi Fičurovi, Ph.D, za důležité rady a pomoc při vedení práce.

Vladislav Hošek

Anotace

Cílem této práce je popsat vytvořit kreditní ratingový model. Tento model je založen na logistické regresi hlavně z toho důvodu, že je nejpoužívanějším a nejdůležitějším nástrojem ve vývoji kreditních ratingových systémů. V druhé kapitole se zabývám teorií s kterou se setkám při tvorbě modelu. Praktická část se pak zabývá datovou analýzou dat použitých v diplomové práci a následným vytvořením, popisem a srovnáním vytvořených logistických modelů. Jde o konkrétní data z bank v USA. První model je vytvořen na užším vzorku dat – je modelován pouze na jednom státě. Druhý model pak využívá celou populaci USA. Celá práce je pak v přílohách doprovozena kódem z programu R.

Klíčová slova:

Basel II, pravděpodobnost defaultu, logistická regrese, šance, datová analýza, probability modelling.

Anotation

The aim of this master's thesis is to create a credit scoring model, which is based on a logistic regression. Th theoretical part of thesis starts theory sorrounding the logistic regression, then moves onto the theory sorrounding model creation. The practical part of the thesis starts with data analysis of given data and creation of several logistic models. First model is created on a smaller sample size – just one US state, the second model deals with the entire population. There is some R code with commentary in the appendix of the thesis.

Key words:

Basel II, probability of default, logistic regression, data analysis, probability modeling

Obsah

Úvod 6
Logistická Regrese 8
Lineární regrese 8
Logistická regrese – jedna proměnná 11
Logistická regrese – více proměnných 15
Tvorba modelu – teoretická část 18
Výběr proměnných do modelu 18
Tvorba logistického modelu 20
Hodnocení kvality modelu 22
Tvorba modelu – praktická část 28
Univariate Analysis 33
Model 1 48
Model 2 54
Závěr 60
Příloha 1 – popisná část univariate analysis 62
Příloha 2 – vytvořené funkce + kód použitý v kapitole Univariate Analysis 78
Příloha 3 – Státy USA a defaultní míry z úvěrů SBA z roku 2006 81
Seznam obrázků a grafů 82
Knižní zdroje 83
Internetové zdroje 83

Úvod

Banky a bankovní systém je centrálním komponentem jakékoliv moderní ekonomiky. V době globalizace se dá i říct, že banky jsou extrémně důležité jak pro chod domácí ekonomiky, tak pro chod světové ekonomiky. Banky jsou závislé na depozitech obyvatelstva – běžných lidí, korporací nebo vlád. Tyto vklady jsou většinou krátkodobé. Příjmy bank jsou pak převážně vázány na poskytování úvěrů či investování, obě dvě tyto operace jsou považovány za dlouhodobé. Je jasné, že vzniká jistá časová nevyrovnanost, která může jednoduše způsobit pád jedné banky a jak ukázaly dvě poslední krize, tak pád jedné banky může dále způsobit pád dalších bank a dojde k zpomalení nejenom domácí ekonomiky ale i světové.

Aby nedocházelo ke krizím, je potřeba banky regulovat nejenom na národní úrovni ale i na mezinárodní úrovni. K tomuto poprvé přispěl v roce 1988 Basel I – soubor pravidel, rad a praktik, které měly minimalizovat rizika spojená s bankovnictvím.

Na Basel I se vztahuje řada kritiky, nejhlavnější body jsou následující:

1) Basel I nebyl citlivý na různá rizika – nebylo diferencováno mezi vysokým, průměrným či nízkým rizikovým chováním bank
2) Neodměňoval dobré chování ani nepenalizoval špatné chování
3) Nezohledňoval technologické pokroky v bankovnictví a financích

Z výše uvedených důvodů přišel v roce 2004 Basel II, který je založen na třech pilířích:

1) Minimální kapitálové požadavky
2) Bankovní dohled
3) Tržní disciplína

První pilíř ukládá bankám povinnost mít sofistikovaný riziko-měřící a řídící systém. Banky si mohou zvolit, zda chtějí používat standardizovaný přístup nebo vlastní přístup[1]. Prvně zmíněný přístup je použitelný pro všechny banky – využívá externího ratingu pro výpočty očekávané ztráty. Druhotně zmíněný přístup je pak navržen tak, aby daná banka mohla plně využít svá nasbíraná data a mohla si sama vypočítat očekávanou ztrátu. Basel II ukládá, že očekávaná ztráta se vypočítá jako:

$$EL = PD \; x \; EAD \; x \; LGD \tag{1}$$

EL značí zmíněnou očekávanou ztrátu, PD značí pravděpodobnost defaultu, EAD značí hodnotu expozice v čase defaultu a LGD značí ztrátu, která vznikne při defaultu.

Definice defaultu je následovná (Basel Committee on Banking Supervision, 2004):

1) Banka předpokládá, že dlužník nebude schopen splatit svoje závazky bance v plné výší.
2) Dlužník se zpozdil s platbou o více než 90 dní.

[1] Volně přeloženo z Internal Rating Based

Basel II uznává jako další zároveň UL- neočekávanou ztrátu. Výpočet tohoto ukazatele je většinou založen na *Value at Risk* modelech. Neočekávaná ztráta je jakousi volatilitou, která nastává kolem očekávané ztráty.

Úspěšné modelování komplexních dat je trochu věda, trochu statistika, trochu zkušenost a trochu běžný rozum.

Kredit skóring je metoda k hodnocení rizika. Cílem této metody je přiřadit každému klientovi nějaké číslo které je vhodným indikátorem rizika, které jeho charakteristiky představují. Před tím, než se v bance začaly používat tyto statistické metody všechno leželo na pracovníkovi banky, který dělal rozhodnutí sám za sebe – řídil se svými pocit, mohl být ovlivněn náladou, jiný pracovník by se rozhodl jinak – zkrátka subjektivní rozhodování. (Anderson, 2007)

Kredit skóring se používá již více než 50 let. Svoje první využití našel v přiřazování různých rizikových indikátorů v oblasti kreditních karet. Jako první model použitý v komerčním bankovnictví byl model z roku 1941, který byl založen na – pohlaví, práci, počet let strávených v práci, počet let strávených na adrese současného bydliště. (Anderson, 2007)

S příchodem počítačů a novými technologickými pokroky bylo jen otázkou času, než počítače začnou rozhodovat za lidi. Jediné, co lidé musí udělat je říct počítači, jak se má rozhodovat, jak má postupovat a v jakém bodě určit, zda je pravděpodobnost defaultu u konkrétního úvěru již nepřijatelná.

A tím se dostávám k cíli své práce, kde cílem této práce je sestrojit model pravděpodobnosti defaultu na reálných datech. Čili budu modelovat jeden z vyžadovaných parametrů pro výpočet očekávané ztráty a tím je PD. Tento skóringový model bude založen na logistické regresi, která je stále nejpoužívanějším nástrojem v tvorbě skóringových modelů uvnitř bank po celém světe.

Logistická Regrese

Co je regrese?

Než se zaměříme přímo na logistickou regresi, je třeba si říci, co vlastně slovo regrese znamená. Regrese je jakýsi vztah mezi vysvětlovanou proměnnou a vysvětlující proměnnou[2]. Tento vztah může být různých druhů – nejznámější a pravděpodobně nejvíc využívaná je lineární regrese. Logistická regrese přímo vychází z lineární regrese, proto bude potřeba se nejdříve zaměřit a připomenout lineární regresi.

Vysvětlující proměnné v této práci budu značit pomocí písmen x_i, které pak budou ve vektoru $x = (x_1, x_2 \dots x_n)$.

Vysvětlovanou proměnnou pak budu značit písmenem y. Její odhad bude značen jako $\hat{y}$. V lineární regresi je nejlepším nezkresleným odhadem právě průměr $\bar{y}$. Tento průměr se snažím odhadnout.

Lineární regrese

Jak jsem již uvedl v úvodu této kapitoly, regrese je vztahem mezi vysvětlovanou a vysvětlující proměnnou, pokud používáme lineární regresi domníváme se, že vztah mezi vysvětlovanou a vysvětlující proměnnou je právě lineární. Kde tento lineární vztah lze popsat pomocí následující rovnice

$$\hat{y} = \beta_0 + \beta_1 x \tag{2}$$

Tuto rovnici můžu přepsat i pomocí podmíněné střední hodnoty

$$E(y|x) = \beta_0 + \beta_1 x$$

Případně pak v modelu, kde vystupuje více než jen jedna vysvětlující proměnná

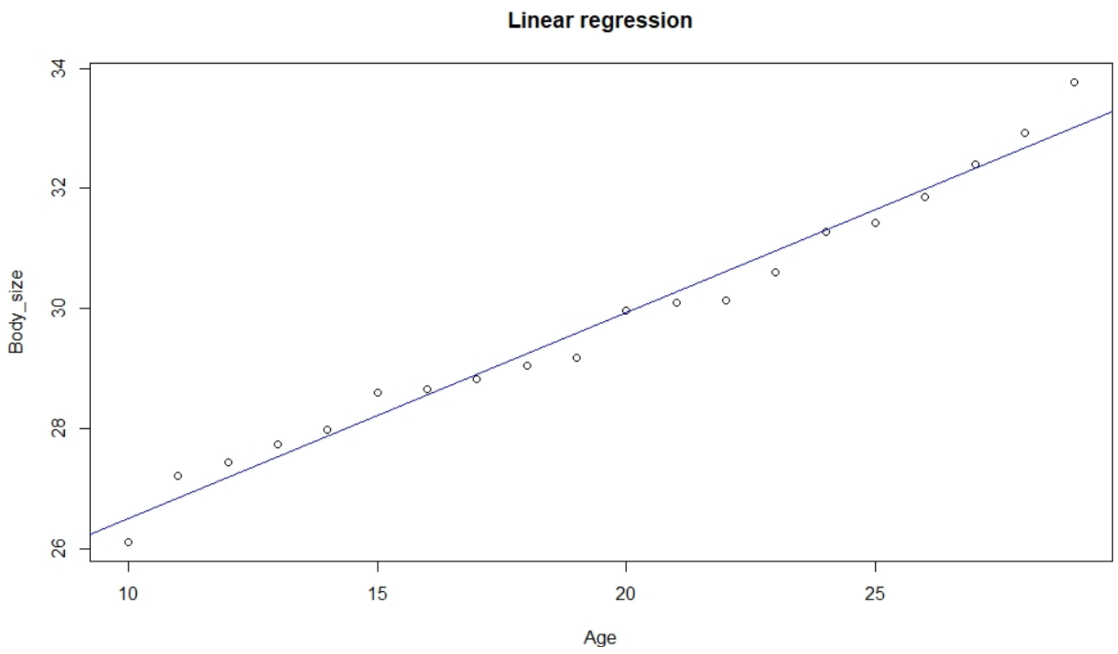
$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Kde řeckým písmenem β_i označujeme odhadované parametry, které se nazývají regresními koeficienty. Parametry se odhadující metodou nejmenších čtverců. Kde se snažíme minimalizovat následující výraz:

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)$$

Příkladem lineární regrese může být následující obrázek, kde jsem vyjádřil závislost mezi věkem ve dnech a velikostí těla milimetrech. Na první pohled je vidět, že vztah je lineární.

[2] Dá se také říci vztah mezi závislou a nezávislou proměnnou



Obrázek 1 - Lineární regrese
zdroj: vlastní zpracování

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)[3]
(Intercept) 23.06603    0.25439   90.67  < 2e-16 ***
Age          0.34347    0.01251   27.45 3.83e-16 ***
```

Z výše uvedeného výstupu vidím hodnotu regresních koeficientů a mohu zapsat regresní rovnici v následující podobě:

$$\hat{y} = 23 + 0{,}35x$$

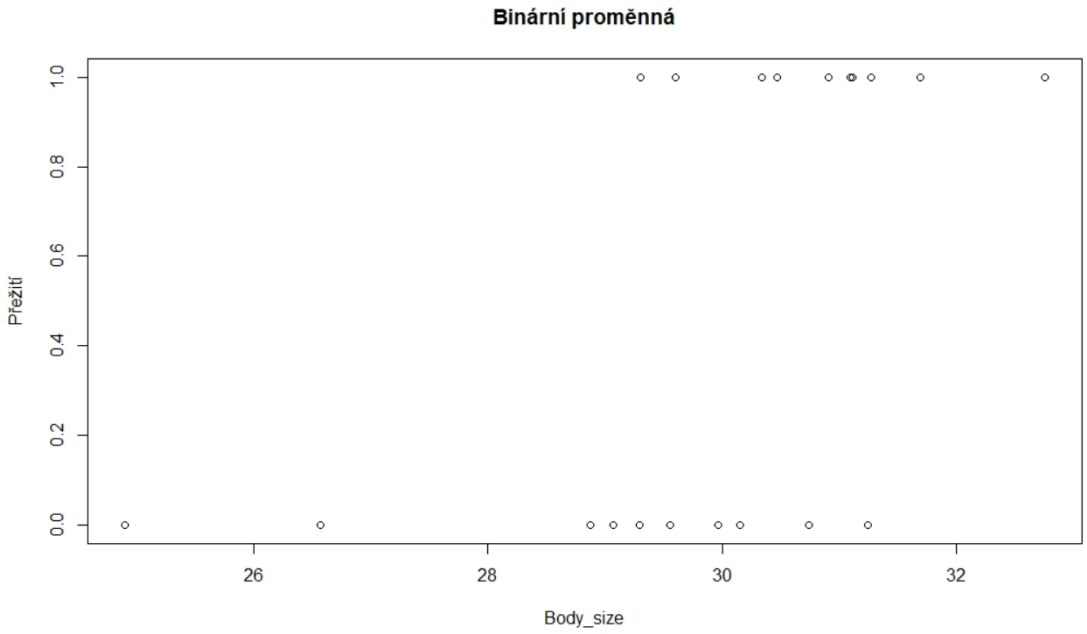
Takto vyjádřený vztah mezi věkem a velikostí těla nám může pomoci při předvídání, jak velkou velikost těla bude mít motýl, který je 35 dní starý. V tom případě pak dosadíme za x 35 a dostaneme výsledek

Lineární regrese klade určité předpoklady na zmíněnou chybovou složku:

1) Konstantní rozptyl – homoskedasticita
2) Nezávislost v čase - autokorelace
3) Normální rozdělení

Nyní uvažujme příklad binární[4] proměnné – to je taková proměnná, která má jen dva stavy nabytí – čili zemřel – nezemřel, prospěl – neprospěl, dobrý – špatný. Tato proměnná je kódována pomocí čísel 1 a 0.

[3] Čím je daná kategorie významnější, tím nižší je její p-hodnota, to je v R-studiu znázorněno také tím, že vedle p-hodnoty je vždycky patřičný počet hvězdiček. *** znamená, že je daná kategorie významná na 0,001 hladině významnosti.
[4] Také zvaná Bernoulliho proměnná

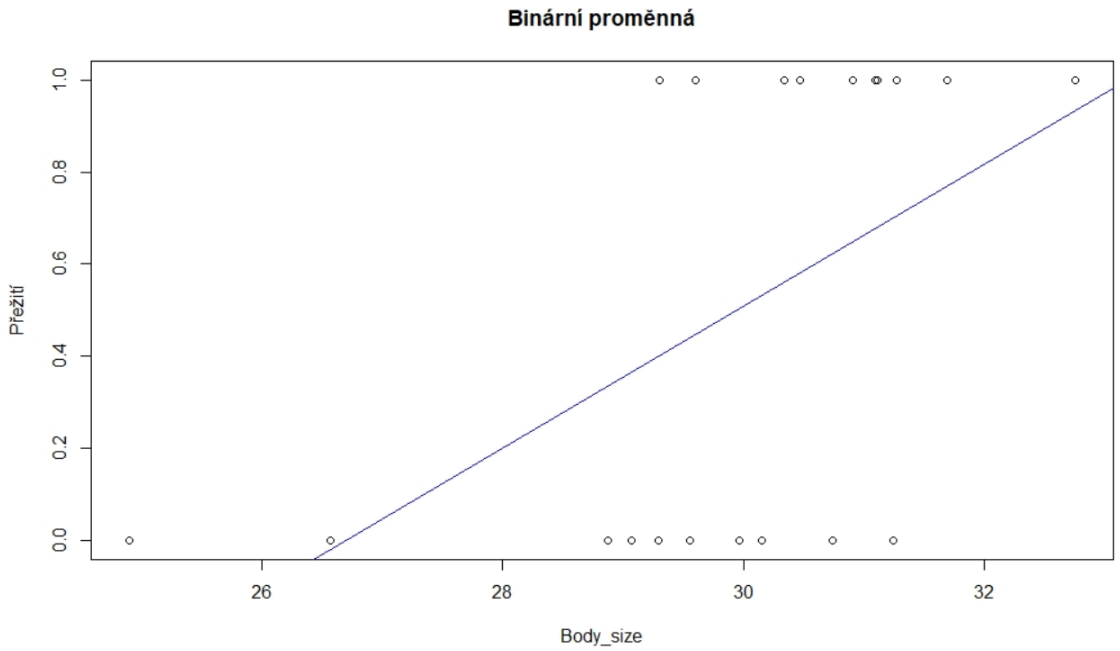


Obrázek 2 – Binární proměnná
zdroj: vlastní zpracování

Kde 1 značí nastání nějaké jevu a 0 značí nenastání. V našem případě na výše uvedeném obrázku je poupravený příklad, který jsem uvedl v části, která se zabývá lineární regresí. Zde máme opět určitý vzorek motýlů, kde část z nich přežila (kódováno pomocí 1) a naopak zemřela (kódováno 2).

V lineární regresi jsme se pokoušeli odhadovat neznámý průměr $\hat{y}$. Co vlastně značí průměr v uvedeném příkladě? Průměr se rovná počtu jedniček k počtu všech pozorování – jedná se o pravděpodobnost – pravděpodobnost, že nastane stav 1 se rovná $\pi(x)$. Pravděpodobnost, že tento stav nenastane se rovná $(1 - \pi(x))$. Rozptyl je pak roven $\pi(x) * (1 - \pi(x))$.

Pokud to výše uvedeného obrázku nafitujeme lineární regresní vztah a dostaneme následující zobrazení regresní přímky.



Obrázek 2 – Binární proměnná a lineární vztah
zdroj: vlastní zpracování

Z obrázku je patrné, že lineární vztah mezi proměnnými neplatí. V následujících řádkách shrnu ostatní rozdíly a nedostatky lineární regrese:

1) Z funkčního vztahu $E(y|x) = \beta_0 + \beta_1 x$ je patrné, že všechny možné hodnoty leží v intervalu $(-\infty; +\infty)$ – což z perspektivy pravděpodobnosti nedává žádný smysl. Odhadovaná podmíněná střední hodnota musí být v intervalu $< 0; 1 >$
2) Předpoklad konstantního rozptylu je porušen – rozptyl binární proměnné je $\pi(x) * (1 - \pi(x))$. Pokud $\pi = 0{,}5$, pak je rozptyl roven $\sigma^2 = 0{,}5 * 0{,}5 = 0{,}25$, pokud $\pi = 0{,}1$ rozptyl se bude rovnat $\sigma^2 = 0{,}1 * 0{,}9 = 0{,}09$
3) Rezidua nejsou normálně rozdělena

Logistická regrese – jedna proměnná

Z přechozí části jasně plyne nutnost pro jiný regresní vztah, a tímto jiným vztahem je konkrétně logistická regrese. Logistická regrese je regresní model s binární proměnnou, binární proto, protože může nabývat pouze dvou možných hodnot. V jednorozměrném modelu má tato vysvětlovaná proměnná alternativní[5] rozdělení s parametrem π. To značíme $Y \sim A(\pi)$. Hlavní rozdíl logistické regrese od lineární regrese je ten, že logistická regrese určuje pravděpodobnost nastání nějakého jevu – v této práci se setkáme především s defaultem – jak bylo uvedeno neschopnosti splácet úvěr. Nastání tohoto jevu se značí 1, pokud tento jev nenastane značí se 0.

V jednorozměrném modelu se pracuje pouze s jednou vysvětlující proměnnou, která má charakteristiku náhodné veličiny.

$$\boldsymbol{X} = (X_1, X_2 \dots X_n)$$

[5] Bernoulliho rozdělení

Konkrétním příkladem může být například výška klienta naší banky, který si žádá o úvěr.

Rozdělení binární proměnné v jednorozměrném modelu

Můžeme si položit otázku, jaké rozdělení má naše hledaná vysvětlovaná proměnná – například pravděpodobnost defaultu

Pravděpodobnost úspěchu je $\pi(x)$, opačný jev – pravděpodobnost neúspěchu je 1- $\pi(x)$. Tuto vlastnost binární proměnné můžeme zapsat následovně:

$$P(Y=y)=\begin{cases}\pi(x) & y=1\\ 1-\pi(x) & y=0\\ 0 & v\ jiném\ případě\end{cases}$$

Tato pravděpodobnost se dá také zapsat jako:

$$P(Y=y)=\pi(x)^{y}*(1-\pi(x))^{1-y}$$

Střední hodnota alternativního rozdělení binární vysvětlované proměnné je pak rovna $\pi(x)$ a rozptyl je roven $\pi(x)*[1-\pi(x)]$.

Odvození rovnice logistické křivky

Pojem *šance*. Šance znamená, kolikrát je pravděpodobnější nastání jevu než jeho opaku.

$$Odds(P(Y=y)=\frac{P(Y=1)}{P(Y=0)}=\frac{P(Y=1)}{1-P(Y=1)}=\frac{\pi(x)}{1-\pi(x)} \quad (2)$$

Dejme tomu, že pravděpodobnost defaultu firmy je $\pi(x)=0{,}9$ pravděpodobnost opačného jevu je $1-\pi(x)=0{,}1$ a šance nastání defaultu je tedy $\frac{0{,}9}{0{,}1}=9$ – tento výsledek můžeme interpretovat tak, že šance defaultu je 9x větší než šance opačného jevu.

V uvedeném příkladě se pohybujeme v prostoru $[0;+\infty]$ my se ale potřebujeme dostat do prostoru $[-\infty;+\infty]$, který je kompatibilní s logistickou regresí. Toto dosáhneme tím, že uvedenou šanci zlogaritmujeme. Použijeme tzv. link funkci – v našem případě je to logaritmus.

$$log-Odds=\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

Pokud si vezmeme výše zmíněnou rovnici log-oddds a dá se do rovnosti s (1) dostaneme rovnici zvanou **logit**:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1x+\epsilon \quad (3)$$

Tuto rovnici pak mohu upravít tak, abychom vyjádřili $\pi(x)$:

$$\pi(x)=\frac{1}{1+\exp\left[-(\beta_0+\beta_1x+\epsilon)\right]} \quad (4)$$

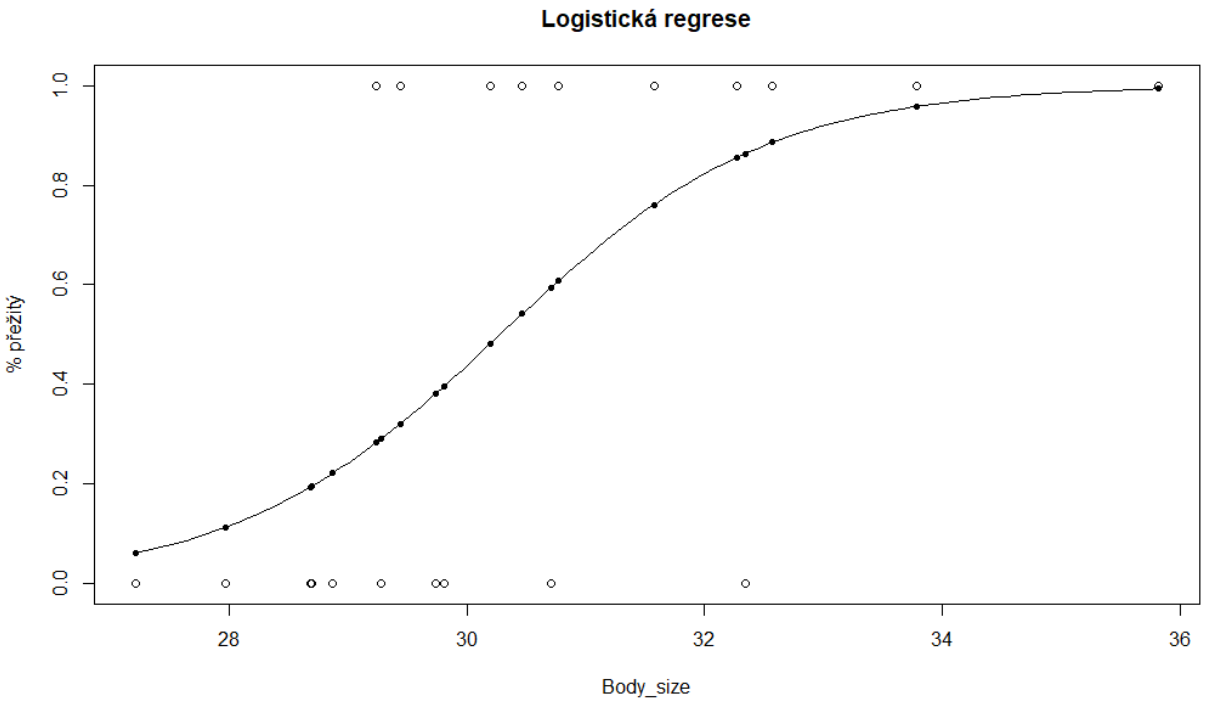
Případně pak v jiném tvaru:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp\,(\beta_0 + \beta_1 x)} \tag{5}$$

Důležitým poznatkem během odvozování vztahu pro logistickou regresi je to, že jsme vysvětlující proměnnou x nekladli žádné přímé požadavky – tento fakt dělá z logistické regrese velice univerzální nástroj a je zároveň důvodem, proč je logistická regrese nejvíc používaný nástroj v kredit-skóringu. Jako jediný předpoklad, který se dá identifikovat je lineární vztah log-odds k vysvětlované proměnné.

Důležitým rozdílem mezi těmito modely je to, že v lineární regresi má chybová složka normální rozdělení s nulovou střední hodnotou a konstantním rozptylem. Tento předpoklad neplatí u logistické regrese, kde je tomu tak, že chybová složka má nulovou střední hodnotu, ale rozptyl se dá vyjádřit jako:

$$\pi(x) * [1 - \pi(x)]$$



Obrázek 4 – Logistická křivka
zdroj: vlastní zpracování

Rovnice výše uvedené logistické křivky je následující již zmíněná rovnice (4).

$$\pi(x) = \frac{1}{1 + \exp\,[-(\beta_0 + \beta_1 x + \epsilon)]}$$

Logistická regrese patří mezi generalizované lineární modely. Rozdíl mezi lineárními a GLM modely můžeme shrnout v několika bodech:

1) Vztah mezi střední hodnotou vysvětlované proměnné Y a vysvětlujícími X můžeme zapsat jako $E(Y \mid x) = g[u(x)]$ kde g je tzv. link funkce, které spojuje lineární prediktor se sledovanou střední hodnotou vysvětlované proměnné. Tato link funkce je potřebná, protože z prostoru $[-\infty, +\infty]$, který je využíván v lineární regresi, se přesouváme do $[0,1]$ pro credit scoring nebo do prostoru $[0, +\infty]$ pro výpočty potřebné v pojištění
2) Předpoklad konstantnosti rozptylu v lineární regresi je v GLM znovu porušen, proto se se používá tzv disperzní parametr.
3) Parametry se odhadují pomocí metody největší věrohodnosti.
4) Na rezidua není kladen požadavek, aby byla normálně rozdělena. Chybová složka má alternativní rozdělení.

Odhadování parametrů v jednorozměrném modelu

V lineárním modelu se parametry odhadují pomocí metody nejmenších čtverců, to u logistické regrese, kde modelujeme binární proměnnou, nejde. V logistické regresi se používá metoda zvaná *maximální věrohodnost*.

Tato metoda nám odhadne neznáme parametry, jejichž pravděpodobnost, že budou sedět na daný vzorek dat, je největší.

Jako první budeme potřebovat věrohodnostní funkci – tato funkce vyjadřuje pravděpodobnost pozorovaných dat jako funkci neznámých parametrů. Dále se budeme snažit maximalizovat tuto věrohodnostní funkci přes sadu odhadovaných parametrů

$$\pi(x_i)^{y_i} * [1 - \pi(x_i)]^{1-y_i}$$

Pokud jsou naše pozorování vzájemně nezávislá, pak lze věrohodností funkci vypočítat jako:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} * [1 - \pi(x_i)]^{1-y_i} \tag{6}$$

Po zlogaritmování dostaneme:

$$L(\beta) = \sum_{i=1}^{n} y_i * \ln[\pi(x_i)] + (1 - y_i) * \ln\,[1 - \pi(x_i)]$$

Abychom nalezli parametry, které maximalizují výše uvedenou sumu L, věrohodnostní funkci musíme zdiferencovat a dát rovnu nule. Dostaneme dvě rovnice:

$$\sum [y_i - \pi(x_i)] = 0$$

$$\sum x_i * [y_i - \pi(x_i)] = 0$$

Po vyřešení těchto rovnic dostaneme odhad parametrů $\beta_0\ a\ \beta_1$, které se značí $\widehat{\beta_0}\ a\ \widehat{\beta_1}$.
V lineární regresi jsme při odhadování parametrů pomocí SSE dostali lineární rovnici

s neznámými parametry, které je jednoduché vyřešit. Při minimalizaci log likelihood funkce dostaneme rovnice, které nejsou lineární, je nutno použít numerických technik k vyřešení.

Jakmile určíme koeficienty naší regrese musíme zjistit, zda jsou proměnné v modelu dostatečně významné. V další sekci mé práce se budeme ptát na otázku – říká nám model více pokud zahrnuje danou proměnnou nebo nám řekne více, pokud v modelu proměnná nebude vystupovat?

Logistická regrese – více proměnných

Model, který by vysvětloval vztah mezi jednou vysvětlující a jednou vysvětlovanou proměnnou by nám byl k ničemu. Síla modelování leží právě v možnosti modelovat několik vysvětlujících proměnných najednou.

Některé proměnné v modelovaných datech mohou být nespojité, spojité nebo kategorické. Příkladem nespojité proměnné je počet dětí, příkladem spojité proměnné je měsíční mzda, příkladem kategorické proměnné může být rodinný stav. Rodinný stav může mít dále několik podúrovní. V nejjednodušším případě se bavíme pouze o dvou možnostech – ženatý, svobodný. Pokud přidáme další úroveň, můžeme mít – ženatý, svobodný, vdovec. Tyto proměnné se do modelu vloží tak, že se použijí tzv. *dummy* proměnné:

- $D1 = 1$ pokud je klient ženatý, jinak $D1 = 0$
- $D2 = 1$ pokud je klient svobodný, jinak $D2 = 0$
- $D3 = 1$ pokud je klient vdovec, jinak $D3 = 0$

Problém nastává v proměnné $D3$, protože tato proměnná je lineární kombinací předchozích dvou proměnných. Pokud máme v modelu lineárně závislé proměnné, nastává problém multikolinearity, která nám pak znevažuje odhady parametrů. Proto se tyto kategorické proměnné kódují jiným způsobem.

	D1	D2
svobodný	1	0
Ženatý	0	1
Vdovec	0	0

Tabulka 1- Dummy proměnné

Kategorie vdovec by ve výše uvedeném případě byla nazývána referenční kategorií. Referenční kategorie není přímo reprezentována ve výpisu modelu, ale je nepřímo zastoupena v *konstantě*. Pro každou úroveň dummy proměnné se odhadne vlastní parametr.

Pokud bude daný žadatel o úvěr vdovec a zároveň tento fakt bude považován jako významný pro stanovení log šance nesplacení úvěrů můžu toto zapsat pomocí následující rovnice.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 * 0 + \beta_2 * 0$$

Pro ženatého bude rovnice následovná:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 * 0 + \beta_2 * 1$$

A konečně pro svobodného:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 * 1 + \beta_2 * 0$$

Můžeme shrnout, že pro *dummy* proměnnou, která má k úrovní, potřebuju pouze $k-1$ dummy proměnných.

Vrátí se zpět k logistickému modelu s více proměnnými. Mějme vektor n nezávislých proměnných, které označíme $x = (x_1, x_2, \dots, x_n)$. Zároveň označíme pravděpodobnost nabytí stavu 1 pomocí: $P(y = 1|x) = \pi(x)$.

Stejně jako v (3) můžu vyjádřit rovnici pro logit u multi-variate logistické regrese pomocí:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 * x_1 + \beta_1 * x_2 \dots \beta_n * x_n \tag{7}$$

Pro zkrácení můžeme psát:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = g(x)$$

A rovnice multi-variate logistické regrese je dána podobně jako v (5):

$$\pi(x) = \frac{\exp(g(x))}{1+\exp\left(g(x)\right)} \tag{8}$$

Nyní zakomponuji poznatky z kategorických proměnných do rovnice pro logit. Mějme i-tou kategorickou proměnnou s k_j úrovněmi, tato proměnná bude zakódována pomocí $k_j - 1$ úrovní. Tyto proměnné můžeme označit jako D_{il} a odhadované koeficienty pro tyto proměnné označíme pomocí β_{il}, kde $l = 1,2, \dots, k_i - 1$. Logit pro model s n proměnnými a j-tou kategorickou proměnnou bude:

$$g(x) = \beta_0 + \beta_1 * x_1 + \beta_1 * x_2 + \cdots + \sum_{l=1}^{k_j} D_{il} * \beta_{il} + \cdots + \beta_n * x_n \tag{9}$$

Fitování multi-variate logistického modelu probíhá stejně jako u jednorozměrného modleu – tedy pomocí maximum likelihood. Rozdíl mezi těmito dvěma modely je patrný. Nicméně je nutné uvést pár pojmů.

Mějme n nezávislých pozorování pro vysvětlovanou proměnou y - $(x_i, y_i), i = 1,2 \dots, n$. Vektor odhadnutých parametrů je $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_n)$.

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} * [1 - \pi(x_i)]^{1-y_i} \tag{10}$$

Kde pravděpodobnost nabytí sledovaného stavu vysvětlovanou proměnnou je definována jako $\pi(x_i)$:

$$\pi(x_i) = \frac{\exp(g_i)}{1 + \exp\,(g_i)} \tag{11}$$

Pokud vyřešíme (10) dostaneme $n + 1$ rovnic s $n + 1$ parametry. Stejně jako v případě s jednou proměnnou musíme použít numerické metody pro výpočet těchto parametrů.

Mějme tedy vektor odhadnutých parametrů značený $\hat{\beta}$. Metoda odhadování rozpylů a kovariancí odhadovaných parametrů vychází z matice druhých parciálních derivací log-likelihood funkce.

Tyto druhé derivace vynásobíme (-1) dáme do matice o rozměrech $(p + 1) * (p + 1)$. Tato matice se značí jako $I(\beta)$ a nazývá se *informační maticí*. Matici rozptylů a kovariancí získáme z inverze této matice:

$$var(\hat{\beta}) = I^{-1}(\beta)$$

Ve většině případů není možné napsat přesné vyjádření prvků této matice. $var(\beta_i)$ je rozptyl prvku na i-té diagonále. $Cov(\beta_i, \beta_j)$ udává kovarianci mezi prvky i a j.

Vztah pro výše zmíněnou *informační matici*:

$$\hat{I}(\hat{\beta}) = X' * V * X$$

Kde:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{12} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{12} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

A kde:

$$V = \begin{bmatrix} \hat{\pi}_1 * (1 - \hat{\pi}_1) & 0 & 0 & \dots & 0 \\ 0 & \hat{\pi}_2 * (1 - \hat{\pi}_2) & 0 & \dots & 0 \\ \vdots & \vdots & \hat{\pi}_2 * (1 - \hat{\pi}_3) & \dots & \vdots \\ 0 & 0 & 0 & \dots & \hat{\pi}_m * (1 - \hat{\pi}_m) \end{bmatrix}$$

Kde připomenu že vztah pro výpočet $\pi(x_i)$ je uveden výše pod (11).

Tvorba modelu – teoretická část

Modelování skóringového modelu je z části umění a z části statistika (Bolton, 2009). Statistickou část jsem nastínil v minulé kapitole, díky nimž můžeme dostat nějaké číslo – skóre, díky němuž můžeme činit nějaká rozhodnutí. Umění leží v tom, že developer modelu má jistou volnost při tvorbě. Do modelu může zahrnout proměnné, které se jeví statisticky nevýznamné, o kterých ze zkušenosti ví, že do modelu patřit budou. Je také na developerovi, jak vyřešit chybějící data – je vhodná interpolace, nebo je nutné tato data vyřadit. Jak řešit odlehlá pozorování atd.

Mým cílem je modelovat vztah mezi pravděpodobností defaultu (konkrétně jde o pravděpodobnost defaultu do doby splatnosti) a řadou vysvětlujících proměnných. Může nastat otázkou jakou konkrétní pravděpodobnost defaultu budu modelovat. Celkově se rozeznávají dva druhy (Hull, J. 2011):

1) Reálné pravděpodobnosti defaultu – získávají se z historických pozorování. Používají se pro analýzu scénářů a pro výpočet Credit VAR
2) Rizikově-neutrální pravděpodobnosti defaultu – získávají se z cen korporátních dluhopisů nebo z cen akcií firmy. Tyto pravděpodobnosti se používají převážně pro výpočet cen různých derivátů (nejrozšířenějším kreditním derivátem je CDS- credit default swap)

V této práci budu používat reálné historické pravděpodobnosti. K získání bezrizikových pravděpodobností defaultu bych potřeboval, aby firmy v mém data-setu byly obchodované na burze nebo aby alespoň získávali kapitál prostřednictvím dluhopisů. Ani jedno z těchto kritérií není běžné pro malé firmy, na kterých je postaven data-set. Nicméně se prokázalo, že bezrizikové pravděpodobnosti jsou až $6x$ větší[6] než reálné pravděpodobnosti defaultu (John C. Hull 2018).

Výběr proměnných do modelu

V této kapitole se seznámíme se strategiemi, které nám poradí, které proměnné zahrnout do modelu a které naopak vypustit.

Proměnné v modelu můžeme dělit na diskrétní a spojité. Diskrétní proměnné můžeme charakterizovat jako jakýkoliv slovný popisek např. barva auta, pohlaví atd. V těchto proměnných se nejčastěji využívá již uvedených dummy proměnných, které pak odkazují na určenou referenční kategorii. Spojité pak, jak již název naznačuje, vyjadřují nějaké číslo na daném intervalu, může jít např o výši mzdy, peníze na běžném bankovním účtu atd. V tom případě nám může vzniknout problém, kde v seznamu deseti-tisíců klientů má každý jinou mzdu. Takto nám vznikne deset tisíc kategorií s vlastním odhadnutým parametrem a p-hodnotou, což je samozřejmě nesmysl.

Jako první se začíná s **univariate analysis** – prohlédneme si každou proměnnou zvlášť, vytvořím základní popisné statistiky, vytvoříme a zhodnotíme grafy. Jedná se o počáteční fázi datové analýzy, vyřadím ty proměnné, které z věcného hlediska nebudou mít žádnou schopnost ovlivnit vysvětlovanou proměnnou – na barvu očí nebude mít vliv jméno

[6] Platí pro dluhopisy s nejvyšším ratingem. Pro dluhopisy s nejnižším ratingem je pak vztah téměř rovný.

pozorovaného klienta. Do této počáteční analýzy také patří statistiky zvané Weight of Evidence (WoE), Information Value (IV).

Další část datové analýzy je **univariate analysis** V této části datové analýzy prozkoumám statistiky. V této analýze již nafituji první logistický model, v kterém bude vystupovat pouze jedna proměnná a budu zkoumat její vliv na vysvětlovanou proměnnou, v této části budou rozhodující p-hodnoty jednotlivých proměnných.

Podle Christine Bolton je nejdůležitější věcí při tvorbě skóringového modelu porozumět konkrétnímu problému, proč skóringový model vlastně developer dělá. Sama dává na data určité subjektivní požadavky:

1) Proměnné by měly dávat smysl – nezařazovat zbytečné proměnné do modelu, čím jednodušší – tím lepší.
2) Proměnné by měly mít predikční sílu.
3) Vyhnout se multikolinearitě.
4) Data by měla být lehce dostupná a stabilní v čase.
5) Proměnné by měly být „legální“ – v tom smyslu, že v dnešní době nelze diskriminovat na základě barvy kůže nebo pohlaví.
6) Proměnné by měly souviset s cílem developera modelu.
7) Minimální ztráta informací – tento bod souvisí s vyřazováním proměnných z modelu. Je důležité vyřazovat „špatně“ proměnné, a ještě důležitější ponechávat „důležité“ proměnné. Bolton rozeznává celkem 3 druhy nepřesností – první nepřesnost vyplývá z nezahrnutí důležité proměnné do modelu, druhá nepřesnost vychází ze statistických chyb odhadovaných koeficientů a třetí nepřesnost vychází z nesprávně zvoleného datového vzorku, který nesouvisí s modelovaným problémem – vzorek musí být dostatečně reprezentativní.

Zmíněnou predikční sílu proměnných určuji na základě následujících statistik:

Weight of Evidence

WoE určuje predikční sílu[7] daných kategorií. Každé rozhodnutí, které model udělá, je založené na nějaké pravděpodobnosti, že se stane daná událost (default v případě této práce). Tato statistika „konvertuje“ riziko, které se vztahuje k jednotlivým rozhodnutím, do lineární podoby, která je snáz pochopitelná.

$$WoE(c) = ln\frac{\Pr\,[c|good]}{\Pr\,[c|bad]} \tag{12}$$

Kde c značí zkoumanou kategorii v modelu (například pohlaví)[8]. WoE můžeme chápat jako podíl všech dobrých klientů v dané proměnné ke všem špatným klientům v téže proměnné.

Jmenovatel i čitatel vzorce se může pohybovat pouze v [0;1]. Pokud převládají dobří klienti WoE bude kladné, pokud převládají špatní klienti WoE bude záporné. (Witzany, 2017)

[7] Prediction power

[8] Tento výpočet je založen na předpokladu, že proměnná byla již kategorizována. WoE lze vypočítat i pro spojité proměnné, tato statistika se pak nazývá *modifikovaná WoE*.

Information Value

IV nám pomáhá vybrat do modelu důležité proměnné a budeme s ní nadále pracovat jako s hlavním nástrojem v univariate analýze.

Definována jako:

$$IV = \sum_{c=1}^{N} (\Pr[c|good] - \Pr[c|bad]) * WoE(c) \tag{13}$$

Můžeme ji tedy definovat jako průměrný vážený WoE, kde sumace jde přes všechny kategorie c až do N. Váhami pak jsou pravděpodobnosti good a bad v dané proměnné. Vysoká hodnota této statistiky říká, že daná proměnná je dobře rozdělena na stavy 0 a 1, čili je jednoznačné, jaký má tato proměnná vliv na vysvětlovanou proměnnou.

Proměnné s $IV < 0.1$ jsou považovány za slabé, proměnné, které mají $IV > 0.3$ jsou považovány za silné (Bolton, 2009).

Tato statistika je vždy nezáporná, protože WoE bude mít vždy stejné znaménko jako rozdíl distribucí. (Witzany, 2017)

Tvorba logistického modelu

Existují dvě základní možnosti, jak vytvořit model logistické regrese (Hosmer, 2000):

1) Stepwise logistic regression
2) Best subset logistic regression

Obě dvě možnosti se využívají na kredit skóring a je jen na developerovi, kterou techniku zvolí. Já jsem v mé práci zvolil stepwise regresi. Proto o ní v několika následujících řádkách budu hovořit.

V stepwise logistické regresi se proměnné vkládají do modelu jednu za druhou a rozhoduje se, zda daná proměnná zůstane v modelu pouze na základě statistických kritérií. Tato technika se dá praktikovat dvěma způsoby:

1) Forward selection – začneme s prázdným modelem a postupně přidáváme proměnné jednu za druhou. Sledujeme, jak se mění relevantní statistiky a rozhodujeme, zda přidanou proměnnou ponecháme nebo naopak vyndáme z modelu.
2) Backward selection – v této možnosti začnu s plným modelem a postupně odstraňuji proměnné a sleduji, jak se mění statistiky na základě kterých rozhoduji, zda danou proměnnou doopravdy odstraním, nebo ji v modelu ponechám. V podstatě se jedná o iterativní proceduru, kdy v každém kroku přidáváme do modelu proměnnou, která co nejvíce zvýší prediktivní sílu modelu.

Zmíněné relevantní statistiky nám udávají, jak „důležitá“ daná proměnná je, v lineární regresi jsme mohli sledovat F-test, kde se předpokládalo, že chyby budou normálně rozdělené, v logistické regresi má chybová složka binomialní rozdělení a důležitost proměnných je

odhadována pomocí likelihood ratio chi-square testu. Čím větší vliv má proměnná na log-likelihood modelu, tím je považována za důležitější.

Likelihood ratio test, o kterým mluvím v odstavci výše, se dá určit:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2(L_0 - L_1) \tag{14}$$

Kde L_0 je maximální hodnota likelihood funkce modelu, který obsahuje pouze konstantu a L_1 je maximální hodnota likelihood funkce modelu, který obsahuje proměnou x_1. Pokud je tento rozdíl roven nule, znamená to, že není žádný rozdíl mezi základním modelem a modelem s přidanou proměnnou = vypovídající schopnost modelu se nijak nezměnila.

Krok 0

Předpokládejme, že máme k dispozici celkem p možných proměnných. Krok číslo 0 začne tím, že nafituje pouze konstantu modelu a vyhodnotí její log-likelihood L_0. Na to navazuje fitování všech p- možných modelů, které obsahují pouze jednu proměnnou kalkulování jejich log-likelihood funkce- L_j^0 je hodnota log-likelihood funkce pro $j - tou$ proměnnou v nultém kroku. Hodnota likelihood ratio testu, pro model obsahující pouze konstantu a model obsahující proměnnou x_j se rovná $G_j^0 = -2 * (L_0 - L_j^0)$. Teď se konečně dostáváme k pravděpodobně nejdůležitější statistice a tou je $P - honodta$. $P - honodtu$ proměnné x_j budu značit jako p_j^0 a dá se vypočítat:

$$P\left[(\chi^2(v) > G_j^0\right] = p_j^0 \tag{15}$$

Kde $v = 1$ (stupně volnosti), pokud x_j je spojitá proměnná a $k - 1$ je počet kategorií, pokud x_j je kategorická proměnná. Nejdůležitější proměnná je ta, která má nejmenší p-hodnotu. Otázkou je, jak musí být p-hodnota velká, aby proměnná byla považována za nedůležitou? Tomuto se říká α hladina významnosti a je většinou 5%. Proměnná s nejmenší p-hodnotou je tedy přidána do modelu. Nazveme ji x_{k1}, protože tato proměnná je kandidátem pro vstup do modelu v příštím kroku s číslem 1.

Krok 1

Do modelu se přidá výše zmíněná proměnná x_{k1}, dostáváme log-likelihood modelu s touto proměnnou, který označím jako L_{k1}^1. Dále musí program rozhodnout, zda zbývající proměnné jsou důležité, to znamená, že program musí *nafitovat* $p - 1$ logistických regresí, které obsahují x_{k1} a x_j, model, který obsahuje tyto dvě proměnné má log-likelihood rovný L_{k1j}^1. Model pak vypočítá likelihood ratio test $G_j^1 = -2 * (L_{k1}^1 - L_{k1j}^1)$ Znovu se určí proměnná s nejmenší p-hodnotou p_j^1, tentokrát s názvem x_{k2}, pokud je tato p-hodnota menší než námi zvolená α pokračujeme ve vybírání proměnných, pokud je větší proces se zastaví.

Krok 2

V tomto kroku máme v modelu již dvě proměnné x_{k1} a x_{k2}. Může se ale stát, že jakmile do modelu vstoupila druhá proměnná, první proměnná se stala nedůležitou, proto musí procedura obsahovat nějakou kontrolu pro *backward eliminaci.* Vzniká tedy nový log-likelihood modelu bez proměnné x_{kj} a to L_{-kj}^2. Likelihood ratio test plného modelu versus

modelu s odstraněnou proměnnou je $G_{kj}^2 = -2 * (L_{-j}^1 - L_{k1k2}^1)$. Aby program rozhodl, jakou proměnnou vyřadit, vyřadí tu, která má největší p-hodnotu.

Poté program pokračuje s forward vybíráním a najde proměnnou x_{k3}, najde k ní p-hodnotu. Nafituje $p - 2$ logistických regresí, které obsahují již vybrané proměnné. Tento proces se opakuje tak dlouho dokud jsme nevyčerpali všechny proměnné nebo do té doby, kdy již není možné přidat žádnou proměnnou, která by zvýšila statistiku likelihood na dané α hladině významnosti.

Hodnocení kvality modelu

Co je vlastně cílem modelu? Model musí být dostatečně jednoduchý, ale zároveň musí co nejlépe vystihovat modelovaný problém. V modelu se snažíme zahrnout co nejméně vysvětlujících proměnných, ale zároveň se snažíme nevypustit ty, které jsou důležité. Čím více máme proměnných, tím větší standardní chyba bude a tím více bude model závislý na konkrétních podkladových datech – My model potřebujeme, aby vysvětloval danou problematiku, ne konkrétní problém.

Špatný vs dobrý klient

V našem vytvořeném modelu se budeme snažit na základě konkrétních informacích rozhodnout, zda klient je špatný nebo dobrý – dobrému klientovi banka úvěr poskytne, špatnému ne. Proto je nesmírně důležité si hned na začátku vymezit tyto dva pojmy.

Rozdělíme si klienty do 4 skupin. Dvě jsem již zmínil nahoře. Skutečně dobří klienti – ty označíme N_{good}, a skutečně špatní klienti - N_{bad}. Kdybychom věděli, kdo je kdo, tak nám nevzniknou žádné ztráty a zároveň zanikne důvod, proč psát tuto diplomovou práci. Musíme použít model, který ke klientovi přiřadí, zda je dobrý, či špatný. Mohou vniknout 4 situace. Dobří klienti, kteří jsou modelem označeni jako dobří, označíme $N_{good|good}$. Dobří klienti, kteří jsou modelem označeni jako špatní klienti - $N_{good|bad}$ o opačně špatní klienti, které model označí jako dobré klienty $N_{bad|good}$ a špatní klienti označení jako špatní $N_{bad|bad}$. To co jsem právě popsal se nazývá *Confusion Matrix.*

Confusion Matrix		**predikované hodnoty**	
		Good	Bad
skutečné hodnoty	Good	$N_{good\|good}$	$N_{good\|bad}$
	Bad	$N_{bad\|good}$	$N_{bad\|bad}$

Tabulka 2- Confusion matrix

Pokud se model strefí (možnosti $N_{good|good}$ a$N_{bad|bad}$) pak nevnikají žádné ztráty. Nicméně pokud model určí možnost $N_{bad|good}$ dostane úvěr někdo, kdo ho nesplatí – vniknou ztráty. Pokud model označí dobrého klienta špatným $N_{good|bad}$, pak banka ztrácí potencionální úrokové výnosy. Proto se snažíme tyto dvě kategorie minimalizovat.

Ujasním pár vztahů:

$$N_{good|good} + N_{good|bad} + N_{bad|good} + N_{bad|bad} = N_{all}$$

První vztah je zcela jasný, součtem všech zúčastněných osob je roven celému vzorku.

$$specificity = \frac{N_{bad|bad}}{N_{bad|good} + N_{bad|bad}} \tag{16}$$

Specificita, také zvaný jako *true negative rate*. Je podíl skutečných špatných kientů ke všem špatným klientům. Dokonalý model by měl specificitu rovnu 1.

$$1 - specificity = \frac{N_{bad|good}}{N_{bad|good} + N_{bad|bad}}$$

Také zvaný *false positivite rate*. Je podíl špatných klientu, kteří byli označeni jako dobří ke všem špatným klientům. Dokonalý model by měl tento ukazatel roven 0. S tímto ukazatelem je zároveň spojena tzv. Type 2 error– uznání špatného klienta. (Witzany, 2017).

$$sensitivity = \frac{N_{good|good}}{N_{good|good} + N_{good|bad}} \tag{17}$$

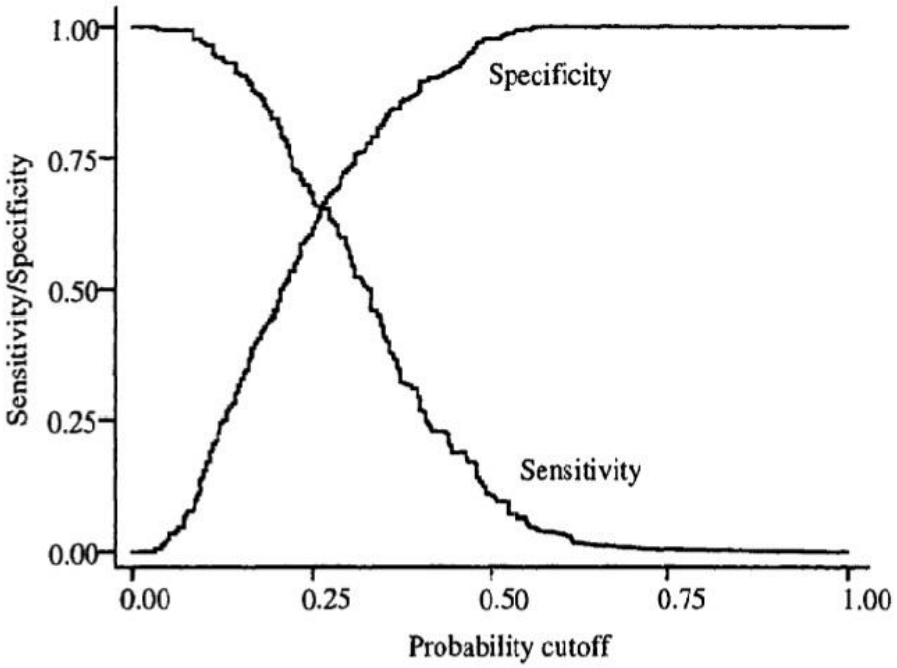
Zvaný *true positive rate*, je podílem skutečně dobrých klientů a všech dobře označených klientů. Opět dokonalý model by měl hodnotu 1.

$$1 - sensitivity = \frac{N_{good|bad}}{N_{good|good} + N_{good|bad}}$$

Také zvaný jako *false negative rate*. Chyba zvaná type 1 error- zamítnutí dobrého klienta. V dokonalém modelu by tento ukazatel byl roven 0. (Witzany, 2017)

Cut-Off point

Mnou vytvořený model odhadne pravděpodobnosti mezi $[0,1]$ – jakékoliv číslo. Nicméně mně zajímá, zda daného klienta budeme považovat za špatného či dobrého. K tomu slouží takzvaný Cut-Off point. Pokud tento bod určíme jako $c = 0.5$, pak všichni klienti, kteří mají modelem odhadnutou šanci defaultu > 0.5 jsou označeni jako 1 – jim úvěr nedáme. Pokud < 0.5, pak jsou označeni jako 0 a úvěr dostanou.



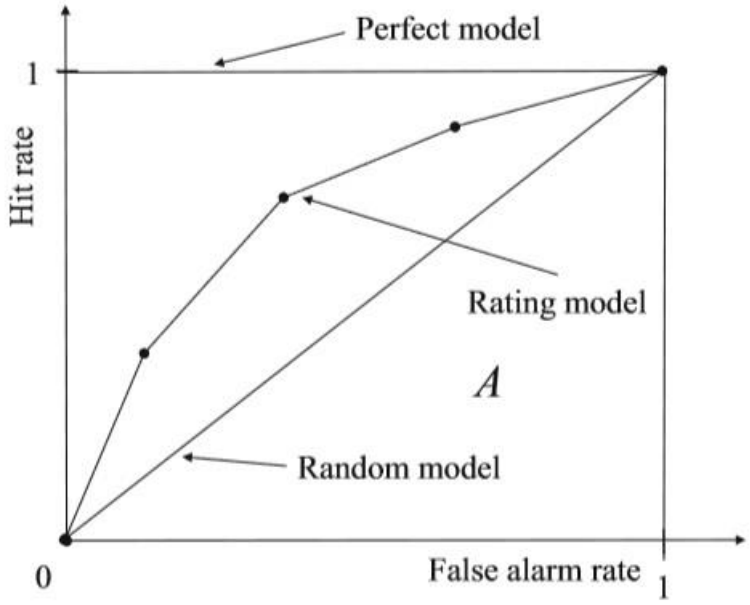
Obrázek 2 - optimální cut-off bod, zdroj: (Hosmer, 2000)

Obrázek shrnuje, kde by měl optimální Cut-off point ležet. Je poznat, že jde o jakousi výměnu mezi specificitou a sensitivitou. Čím více lidí pošleme do defaultní kategorie (označení 1 pro default) tím větší šance je, že nám zbyli pouze zdraví klienti. Dává smysl, že sensitivita dosahuje hodnoty 1 v bodě $c = 0$. V tomto bodě banka poskytne úvěr všem lidem nezávisle na tom, jaké skóre jim model přiřadil. Proto „trefí“ 100% dobrých klientů. Zároveň dává smysl, že sensitivita je v tomto samém bodě nula, jelikož jsme dali úvěr všem žadatelům, model „netrefil“ ani jednoho špatného klienta.

ROC

v celém znění zvaná Reciever Operator Characteristic Curve a s ní spjatá oblast pod křivkou - $AUROC$ - která nám říká, jak úspěšně model dokázal diskriminovat mezi špatnými a dobrými klienty. Znovu chceme tutu statistiku co nejblíž k číslu 1.

Jedním ze způsobů, jak rozumět ROC křivce je, na horizontální křivce máme false positive rate – čili klienti, kteří byli určeny modelem jako dobří, ale ve skutečnosti jsou špatní. Na Vertikální křivce máme True positive rate – zdraví klienti, správně určení modelem.



Obrázek 3 - ROC curve, zdroj: (Witzany, 2017)

Hosmer a Lemeshaw ve své knize uvádějí jednoduchý benchmark jak poznat, zda vytvořený model je kvalitní:

$AUROC < 0.5$ znamená, že dobrým klientům nedáváme půjčky a špatným klientům půjčky poskytujeme. Tato situace by neměla být možná, pokud nechceme zbankrotovat banku.

$AUROC = 0.5$ znamená, že model nerozlišuje mezi dobrými a špatnými klienty, je to jako kdybychom házeli mincí a podle toho rozdávali půjčky – ROC křivka v tomto případě má formu diagonály.

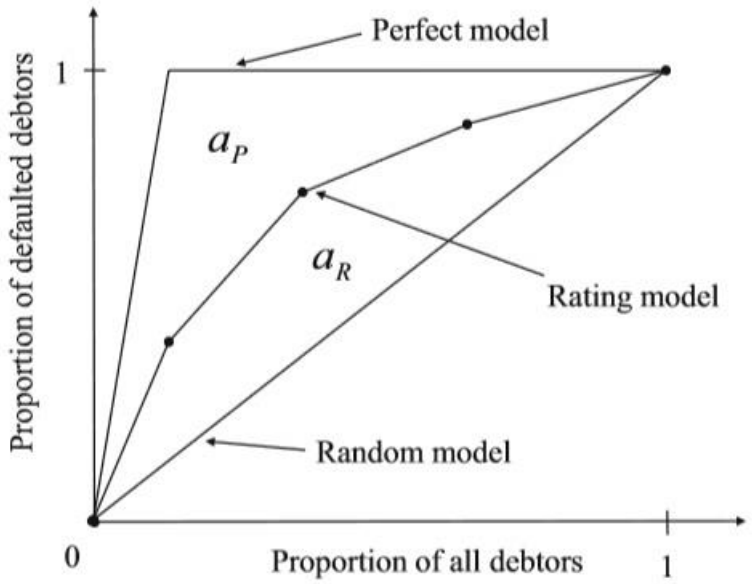
$0{,}7 < AUROC < 0{,}8$ – je považováno za přijatelný model.

$0{,}8 < AUROC < 0{,}9$ – považováno za velice dobrý model.

$AUROC > 0{,}9$ – je považováno za excelentní model. V praxi nevyskytuje, tzv *too good to be true*

CAP

je křivka kde na horizontální ose máme všechny klienty F_{ALL} a na vertikální ose jsou všichni špatní klienti F_{bad}, pomocí CAP vypočítáme statistiku zvanou AR – accuracy rate.



Obrázek 4 - cumulative accuracy profile, zdroj: (Witzany, 2017)

$$AR = \frac{Plocha\ mezi\ Cap\ křivkou\ a\ diagonální\ křivkou}{Plocha\ mezi\ "dokonalým"\ Cap\ modelem\ a\ diagonální\ křivkou} \quad (4)$$

[9]

Zároveň je z tohoto grafu možno vypočítat statistiku zvanou **gini**. Tato statistika se rovná poměrem plochy pod CAP křivkou k k dokonalému modelu. Čím blíže je tato statistika k 1 tím je model dokonalejší.

$$Gini = \frac{Plocha\ pod\ Cap\ křivkou}{Plocha\ pod\ "dokonalým"\ Cap\ modelem} \quad (19)$$

Na první pohled je vidět, že obě dvě křivky jsou si velmi podobné. Vztah mezi AR a AUROC můžeme zapsat jako $AR = 2 * AUROC - 1$

Kolmogorov-Smirnov Statistic

KS statistika porovnává vzdálenost mezi empirickou a teoretickou distribuční funkcí. Platí, že čím menší vzdálenost, tím je model přesnější. Obecně Vypočítá se jako:

$$KS = \sup\left(F_n(x) - F(x)\right) \quad (5)$$

F(x) je teoretická distribuční funkce, $F_n(x)$ je empirická distribuční funkce vytvořená z n pozorování a sup značí supremum. V kontextu tvorby skóringových modelů pak KS statistiku můžu definovat jako maximální vzdálenost ROC křivky od diagonály.

[9] Tím rozumíme křivku, která má 45° a začíná v bodě [0,0]

AIC

Akaikeho kritérium se používá pro relativní srovnání mezi modely pro jedna konkrétní data. Platí, že čím menší číslo, tím je model kvalitnější. AIC se vypočítá jako:

$$AIC = 2k - 2\ln(\tilde{L}) \tag{20}$$

Kde k je počet parametrů a $\tilde{L}$ je maximum likelihood-funkce. Jak je vidět, AIC je penalizační kritérium. Čím více parametrů v modelu, tím větší hodnota AIC, tím horší model.

BIC

Bayseovské informační kritérium. Je velice podobné výše uvedenému AIC, znovu vybereme ten model, jehož BIC je nejmenší.

$$BIC = \ln(n)k - 2\ln(\tilde{L}) \tag{21}$$

Toto kritérium penalizuje méně v případě, že model má menší počet parametrů, je však daleko přísnější při vyšším počtu parametrů.

Tvorba modelu – praktická část

V praktické části vytvořím model logistické regrese na reálných datech. Kde se budu snažit modelovat vztah mezi řadou vysvětlujících proměnných a jednou vysvětlovanou – která je konkrétně pravděpodobnost defaultu do splacení[10]. Tato data mají název National SBA – Small Busines Administration. Jedná se o data od roku 1962 do roku 2014 v USA. SBA je vládní agentura, která podporuje malé firmy v USA. Jedná se o data, která byla zaznamenána od roku 1962 do roku 2014. Jedná se celkem o 899 164 pozorování. Tato pozorování jsou z celkem 51 států USA[11].

SBA podporuje malé firmy tak, že poskytuje federální garanci na poskytnutý úvěr. To znamená, že pokud dojde k defaultu, SBA zaplatí předem danou část. Tato garance zvyšuje šanci, že banka poskytne úvěr malé firmě.

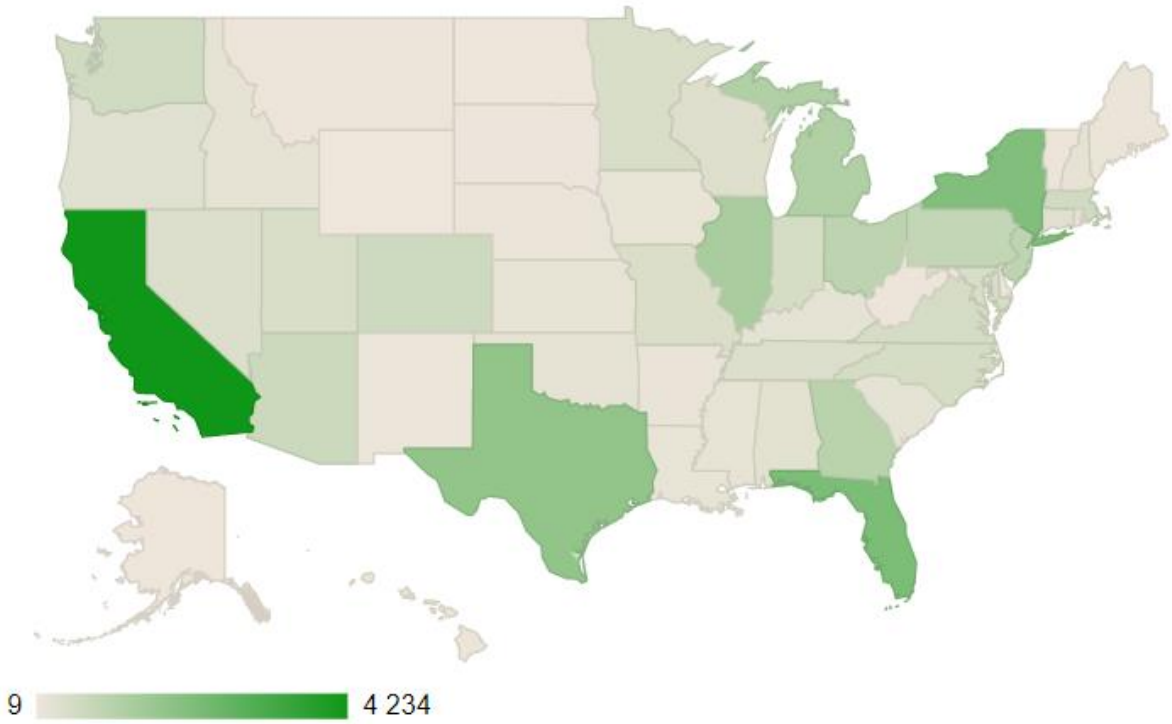
Před tím, ne začnu s analýzou dat je nutné charakterizovat úvěry pro malé firmy. Z tohoto data setu budeme schopni vyvodit nějaké společné charakteristiky pro úvěry pro malé firmy. Průměrný doba, na který je úvěr sjednán je 111 měsíců (10 let) a průměrná doba do defaultu je (pokud k defaultu dojde) 4,5 roku.

V této práci se zaměřím na konkrétní období od roku 2006, jako důvod uvádím to, že se nechci vyhnout období krize, kde se dá očekávat jak zvýšenou intenzitu defaultů, tak zároveň snížený počet poskytnutých úvěrů. Bude zajímavé se pokusit toto období, kde je relativně zvýšená míra defaultu, nějak vyřešit. Zároveň musím přihlédnout k uvedeným společným vlastnostem SBA úvěru. Tzn nemá cenu mít v datech rok 2011 když průměrná doba do defaultu je 4,5 roku!

V první části zaměřím pouze na jeden konkrétní stát z USA a to Texas. A budu se snažit odpovědět na otázku, zda poskytnou úvěr náhodnému zákazníkovi s předem danými charakteristikami – vžiju se tedy do pracovníka banky k ruce budu mít pouze mnou vytvořený model a nějaká data poskytnutá potenciálním klientem, který bude z Texasu. V druhé části se pak pokusím tento rozhodovací proces rozšířit na celé USA.

Nejdřív se podívám na všechna data z roku 2006 a kolik defaultů se odehrálo v jednotlivých státech USA. K tomu poslouží mapa USA členěná na jednotlivé státy. Tmavší odstíny zelené označují vyšší počet defaultů v daném státě.

[10] Dále je také velice časté modelovat PD v prvním roce, kde se očekává, že největší počet defaultu nastane právě v první roce po vzniku úvěru.
[11] 50 států + DC

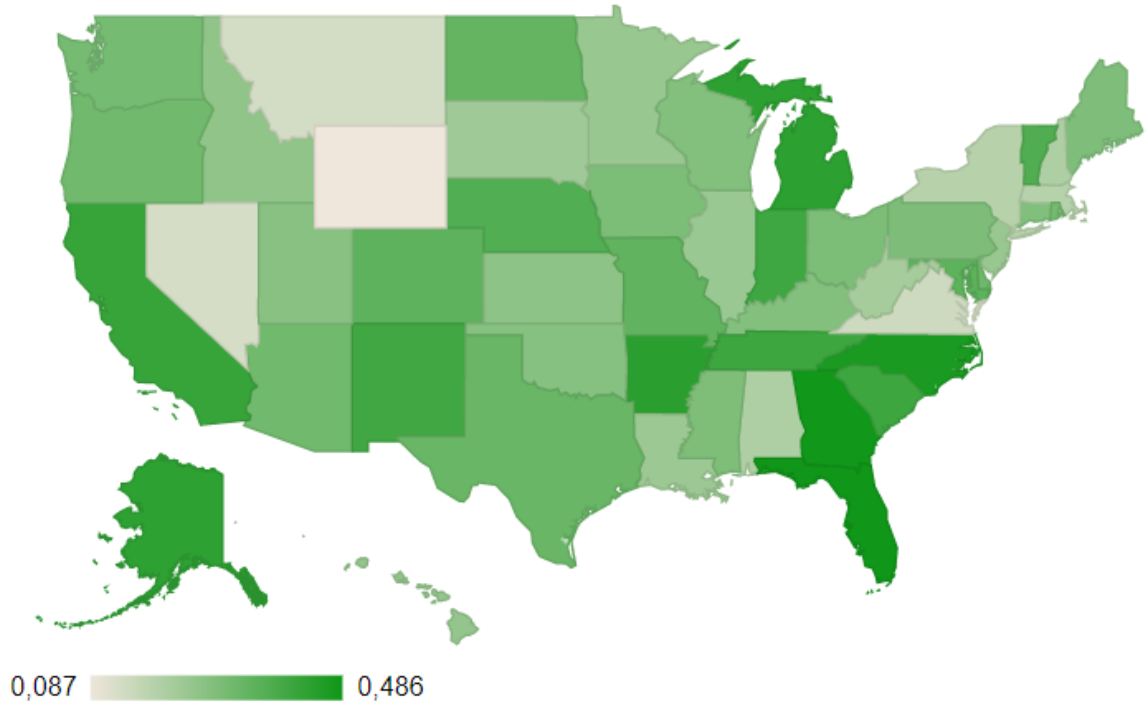


Obrázek 5 - heat map- počet defaultů, , zdroj: Vlastní zpracování v R

Od roku 2006 došlo k největšímu počtu defaultů v Kalifornii. Druhým nejhorším státem je v tomto smyslu New York. Tento fakt je způsoben hlavně díky ekonomické krizi, která začala v roce 2009. Kdy finanční a technologické odvětví byla zasažena nejvíc ze všech. Texas se umístil na až na čtvrtém místě s 1819 defaulty[12].

Nicméně suma defaultů není vypovídající, populace států USA je velice různorodá, zároveň ekonomické podmínky v jednotlivých státech jsou odlišné – typicky venkovský jih je více zaměřen na zemědělství. Kalifornie je známá velkými technologickými giganty. Texas je pak známý ropným průmyslem. Wall Street v New Yorku pak přitahuje množství finančních institucí. Je nutné zobrazit, jak se liší míra defaultu napříč americkými státy. Následující mapa zobrazuje míru defaultu v jednotlivých státech. Tmavší zelená značí vyšší míru defaultu.

[12] Toto číslo se postupně měnilo – začátek byl 1761 defaultů za sledované období a americký stát. Podrobnosti jsou v příloze.



Obrázek 6- Heat map - míra defaultu v USA, zdroj: Vlastní zpracování v R

Z obrázku[13] je patrné, že bude záviset na tom, kde daný klient sídlí. Kde nejhůře dopadl stát Florida s téměř 50% defaultujících úvěrů. Nejlépe dopadl Wyoming s pouhými 9%.

V data setu vystupuje celkem 27 proměnných – z toho 1 je vysvětlovaná a 26 vysvětlujících. Data obsahují chybějící položky, s kterými bude třeba se vypořádat.

[13] Tato mapa je zvaná Heat Map

Následující tabulka shrnuje všechny proměnné, které vstupovaly do počáteční univariate analýzy.

Variable Name	Data Type	Description of variable
LoanNr_ChkDgt	Text	Id klienta
Name	Text	Jméno klienta
City	Text	Město klienta
State	Text	Stát klienta
Zip	Text	Zip Code klienta
Bank	Text	Jméno banky
BankState	Text	Domovský stát banky
NAICS	Text	Americká značka průmyslu
ApprovalDate	Datum	Datum poskytnutí úvěru
ApprovalFY	Text	Rok poskytnutí úvěru
Term	Číslo	Doba úvěru
NoEmp	Číslo	Počet zaměstnanců
NewExist	Text	1 = již existující Business, 2 = nový Business
CreateJob	Číslo	Počet vytvořených prací
RetainedJob	Číslo	Počet zachráněných prací
FranchiseCode	Text	Franchise Code 00000 or 00001 = No Franchise
UrbanRural	Text	1= Město, 2= venkov,
RevLineCr	Text	Revolving Line of Credit
LowDoc	Text	LowDoc Loan Program
ChgOffDate	Datum	Datum, kdy došlo k defaultu
DisbursementDate	Datum	Datum, kdy přišly peníze na účet
DisbursementGross	Měna	Množsví peněz, které přišlo na účet
BalanceGross	Měna	Peníze, které klient dluží
MIS_Status	**Text**	**Stav úvěru – vysvětlovaná proměnná**
ChgOffPrinGr	Měna	Odepsaná částka
GrAppv	Měna	Velikost úvěru
SBA_Appv	Měna	Výše Garance

Tabulka 3 - Proměnné 1

Podrobnou počáteční analýzu kompletním popisem všech proměnných, kódem a rozhodovacím procesem proč vyřadit či ponechat jednotlivé proměnné můžete nalézt v příloze číslo 1 této kapitoly.

Proměnné jsem odstraňoval především na základě „pocitu“. Kdy jsem se snažil argumentovat, bez jakékoliv statistické podpory, že například jméno klienta nijak neovlivní schopnost splácet úvěr. Proměnné A1 a A2 byly vytvořeny mnou jako pomocné proměnné. Proměnná A1 vyjadřuje procentuální část úvěru, která je garantována SBA. Proměnná A2 pak vyjadřuje, jestli je úvěr kryt nějakou nemovitostí.

Z 26 + 1 původních proměnných v datech zbylo celkem 10 + 1 proměnných. Následující seznam ukazuje zbylé proměnné, které budu dál analyzovat a z kterých vytvořím svůj skóringový model.

Název	Typ
MIS_Status	vysvětlovaná proměnná, 1 = default, 0 = paid off
NAICS	Spojitá proměnná. Značka průmyslového odvětví
Term	Spojitá proměnná. Doba úvěru
NoEmp	Spojitá proměnná. Počet zaměstanců
DisbursementGross	Spojitá proměnná.Peníze připsané na účet.
NewExist	Faktor. Nové podnikaní = 1, existující = 0.
FranchiseCode	Faktor. Franšíza = 1, běžný podnik = 0.
UrbanRural	Faktor. Město = 1, venkov = 0.
RevLineCr	Faktor. Revolvingový úvěr = 1, normální úvěr = 0.
A1	Faktor. 4 úrovně. % zajištění SBA
A2	Faktor. 1 = zajištění pomocí nemovitosti, 0 = nezajištěný

Tabulka 4 - Proměnné pro univariate analysis

Faktorové proměnné, kterých je 6, jsou již připraveny vstoupit do modelu bez dalšího upravování. Problém bude ve spojitých proměnných, kterých je celkem 4, které bude potřeba faktorizovat – to znamená je rozdělit na určitý počet úrovní.

Univariate Analysis

V této druhé analýze se už podívám na jednotlivé proměnné v souvislosti s vysvětlovanou proměnnou. Bude mě zajímat distribuce špatných (označených pomocí čísla 1) úvěrů mezi kategoriemi jednotlivých proměnných. Spočítám statistiky uvedené v teoretické části, a to hlavně IV a WoE. Zároveň nafituji první logistický model, který bude mít pouze dvě proměnné – vysvětlující a vysvětlovanou. Musím se vypořádat se spojitými proměnnými a udělat ze všech proměnných faktory, kde cílem není nejmenší počet kategorií, ale to, aby rozdíly v defaultních mírách byli statisticky významné

NAICS

Proměnná ukazující odvětví průmyslu, v kterém žadatel o úvěr plánuje podnikat. Tato proměnná obsahuje celkem 23 různých odvětví. Tyto proměnné rozdělím na základě jejich míry defaultu do celkem 3 skupin. Pokud dané odvětví má 0-33% defaultu, bude v první skupině – celkem 1847. Pokud dané odvětví má 33%-36,4% míry defaultu, bude ve druhé skupině – celkem 1991 firem. Pokud bude v 36,4-67% půjde do třetí skupiny, celkem 1508 malých firem.

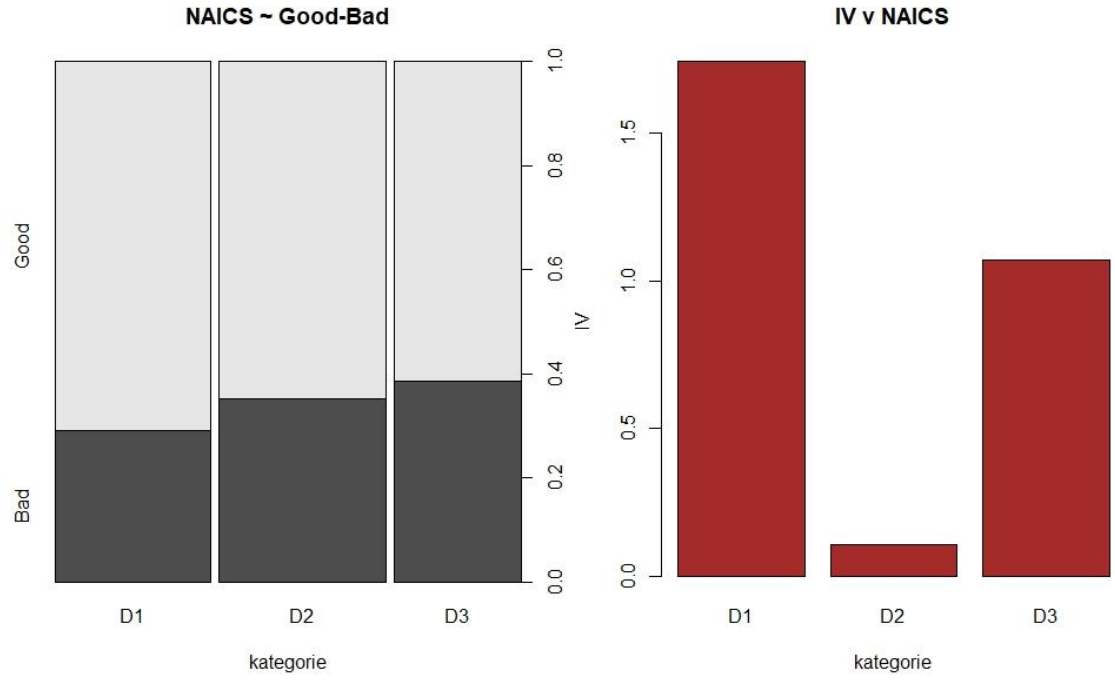
NAICS	# Defaultů	# firem	DR	Skupina
55	2	3	0.6666667	3
31	37	80	0.4625000	3
52	42	100	0.4200000	3
48	94	230	0.4086957	3
51	43	110	0.3909091	3
56	114	307	0.3713355	3
44	248	678	0.3657817	3
42	107	294	0.3639456	2
45	101	279	0.3620072	2
23	190	540	0.3518519	2
71	34	97	0.3505155	2
72	204	586	0.3481229	2
53	42	122	0.3442623	2
61	21	61	0.3442623	2
11	4	12	0.3333333	1
54	178	534	0.3333333	1
32	31	100	0.3100000	1
49	8	27	0.2962963	1
81	171	581	0.2943201	1
33	51	196	0.2602041	1
62	98	380	0.2578947	1
21	0	26	0.0000000	1
22	0	2	0.0000000	1

Tabulka 5 - NAICS

Names	0	1	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV

D1	537	1309	29.52	37.12	1846	34.54	29.1	2.29	1.74040
D2	702	1289	38.59	36.56	1991	37.25	35.26	-0.54	0.10962
D3	580	928	31.89	26.32	1508	28.21	38.46	-1.92	1.06944

Tabulka 6 - WOE a IV pro proměnnou NAICS



Obrázek 7 - Bivariate analysis proměnné NAICS, zdroj: vlastní zpracování

Z horního obrázku lze velice dobře vidět, že separace do skupin podle defaultních měr byla dobrým krokem – skupina D1 má nejmenší počet defaultů a to konkrétně 29,1 procent všech klientů. Druhá defaultní skupina pak má defaultní míru 35,26% a třetí nejvyšší míru 38,46%.

Takto faktorizovanou proměnnou dám do modelu a zjistím, že všechny faktorizace této proměnné jsou významné na jakékoliv hladině významnosti.

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.89102    0.05125 -17.387  < 2e-16 ***
   D2        0.28333    0.06947   4.078 4.54e-05 ***
   D3        0.42102    0.07367   5.715 1.10e-08 ***
```

Tato proměnná bude vystupovat ve výsledném modelu. Bude zde záležet, do jaké defaultní skupiny daná malá firma spadne. D1 skupina je nejméně riziková, D3 je pak nejvíce riziková.

Rovnice pro tento zmenšený model vypadá následovně:

$$\log\left(\frac{p}{1-p}\right) = -0{,}89 + 0{,}28 * D2 + 0{,}42 * D3$$

Tyto odhady parametrů mi udávají vztah mezi závislou proměnnou (default – ano/ne) a mezi nezávislou proměnnou, která je v tomto prvním případě NAICS. Tyto koeficienty jsou na logit měřítku. V tomto případě je interpretace následovná – pokud se klient dostane do D2

skupiny očekáváme +0,283 nárůst v log-odds[14]. Pokud pochází klient ze skupiny D3, nárůst v log-odds bude ještě větší a to +0,42, toto souhlasí s mým záměrem rozdělit klienty do tří skupiny od nejnižší po nejvyšší míru defaultu. Z log-odds můžu udělat „pouze“ odds jednoduchou úpravou.

$$exp\left[\log\left(\frac{p}{1-p}\right)\right] = \exp\left(-0{,}89 + 0{,}28 * D2 + 0{,}42 * D3\right)$$

Pro klienty v první skupině platí rovnice:

$$\frac{p}{1-p} = \exp\left(-0{,}89\right)$$

$$\frac{p}{1-p} = 0.41$$

$$p = 0.29$$

Pravděpodobnost defaultu klienta v první skupině je $0{,}29\%$. To jsou zhruba 3 malé firmy z 10. Odds by v tomto případě byly $3/7$.

Pro klienty v druhé skupině platí rovnice:

$$\frac{p}{1-p} = \exp(-0{,}89 + 0{,}28 * 1)$$

$$\frac{p}{1-p} = 0.54$$

$$p = 0.35$$

Pravděpodobnost defaultu klienta v druhé skupině je $0{,}35\%$. To jsou zhruba 3,5 malých firem z 10. Odds by v tomto případě byly $3{,}5/6{,}5$.

A pro klienty ve třetí skupině platí rovnice:

$$\frac{p}{1-p} = \exp\left(-0{,}89 + 0{,}42 * 1\right)$$

$$\frac{p}{1-p} = 0.625$$

$$p = 0.38$$

Pravděpodobnost defaultu klienta v první skupině je $0{,}38\%$. To jsou zhruba 4 malé firmy z 10. Odds by v tomto případě byly $4/6$.

TERM

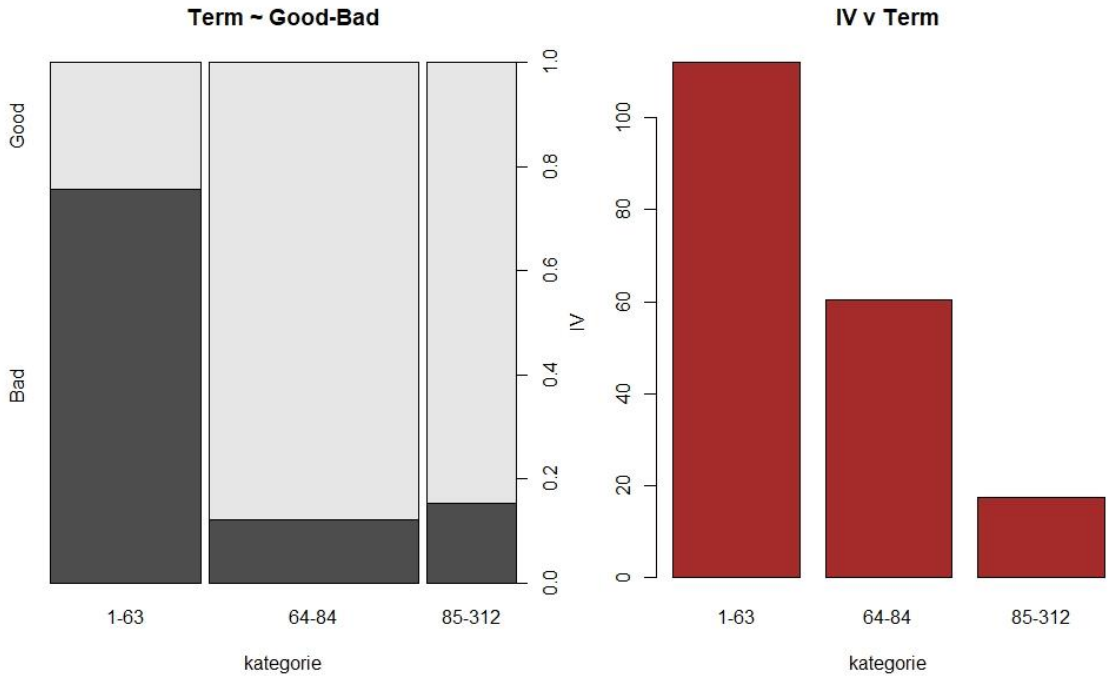
Spojitá proměnná s minimální hodnotou 1 měsíc a s maximální hodnotou 312 měsíců. Tuto proměnnou rozdělím na celkem 3 skupiny podle toho, na jak dlouhou si daná malá firma

[14] Předpokladem je, že ostatní proměnné zůstávají konstantní – to zde není problém – máme jen jednu nezávislou proměnnou

úvěr sjednala. Pokud je úvěr sjednán na kratší dobu než 63 měsíců, je zařazen do první skupiny, pokud je úvěr sjednán na dobu mezi 64 až 84 měsíců, je zařazen do druhé skupiny, pokud je sjednán na delší dobu než na 84 měsíců, je zařazen do třetí skupiny.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
1-63	1357	435	74,6	12,34	1792	33,53	75,73	-17,99	112,0057
64-84	301	2194	16,55	62,22	2495	46,68	12,06	13,24	60,46708
85-312	161	897	8,85	25,44	1058	19,79	15,22	10,56	17,51904

Tabulka 7 - WOE a IV pro proměnnou Term



Obrázek 8- Bivariate analysis proměnné Term, zdroj: vlastní zpracování

Toto rozdělení nám odhalí poměrně zajímavé zjištění a to je to, že větší šanci na nesplacení úvěru mají úvěr sjednané na kratší dobu, téměř 75% všech defaultních úvěrů zapadá do první skupiny, která je 63 měsíců max.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.13769    0.05510   20.65   <2e-16 ***
Term2       -3.12406    0.08255  -37.85   <2e-16 ***
Term3       -2.85534    0.10179  -28.05   <2e-16 ***
```

Všechny podkategorie jsou významné na jakékoliv hladině významnosti. Nejmenší rizikovost z pohledu šance na default má skupina číslo dvě (64-84 měsíců), nejrizikovější je pak skupina číslo jedna (1-63 měsíců).

Rovnice pro výpočet šance pro jednotlivé podkategorie je následující:

$$exp\left[\log\left(\frac{p}{1-p}\right)\right] = \exp\left(1.13 - 3.12 * Term2 - 2.85 * Term3\right)$$

Zde uvedu pouze výpočet pro skupinu číslo jedna (1-63 měsíců):

$$\frac{p}{1-p} = 3.09$$

$$p = 0.75$$

Pro úvěry, které jsou sjednané na krátkou dobu platí, že téměř 75% z nich dosáhne defaultu. Šanci v tomto případě můžu interpretovat jakože u každých tří malých firem ze čtyř dojde k defaultu.

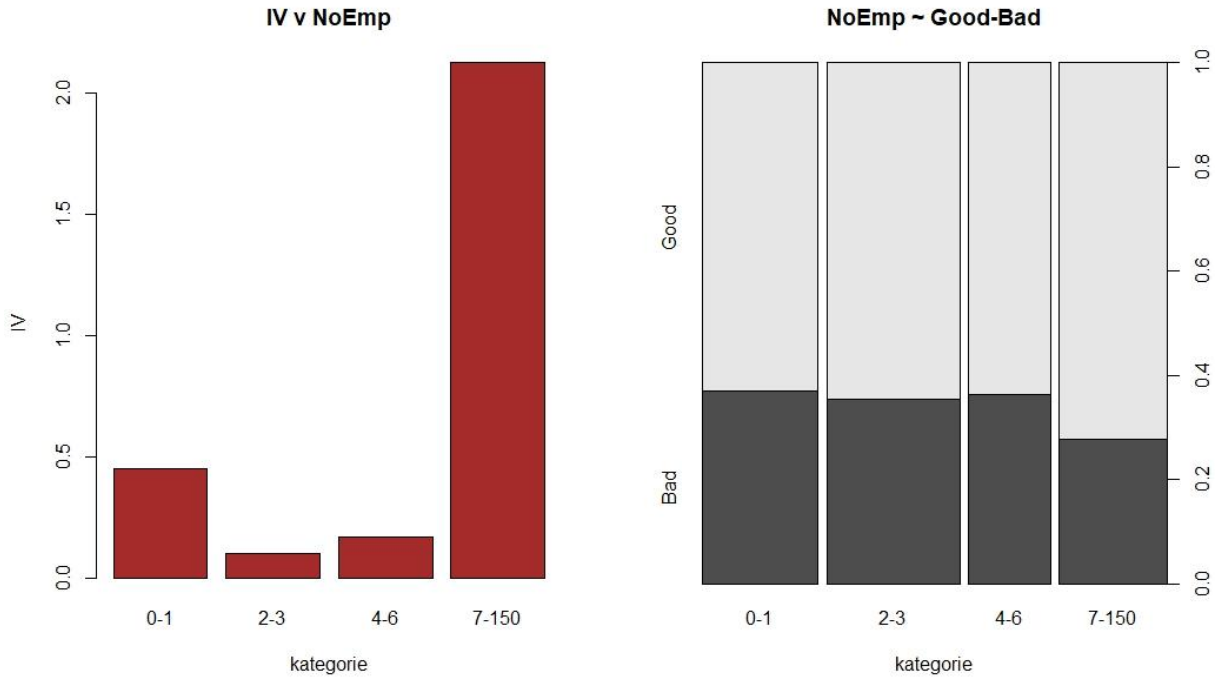
Celkově bych tuto proměnnou interpretoval tak, že čím delší je doba, na kterou je úvěr splácen, tím nižší je pravděpodobnost, že dojde k defaultu.

NoEmp

Většina firem (zhruba 75%) má méně než 6 zaměstnanců. Tuto spojitou proměnnou jsem rozdělil na celkem 4 skupiny. V první skupině (0-1 zaměstanec) je celkem 1411 malých firem. V druhé skupině (2-3 zaměstnanci) je 1619 malých firem, ve třetí skupině (4-6 zaměstnanců) je 1000 firem a v poslední čtvrté skupině je 1315 firem.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
0-1	522	889	28,7	25,21	1411	26,4	37	-1,3	0,4537
2-3	572	1047	31,45	29,69	1619	30,29	35,33	-0,58	0,10208
4-6	362	638	19,9	18,09	1000	18,71	36,2	-0,95	0,17195
7-150	363	952	19,96	27	1315	24,6	27,6	3,02	2,12608

Tabulka 8 - WOE a IV pro proměnnou NoEmp



Obrázek 9- Bivariate analysis proměnné NoEmp, zdroj: vlastní zpracování

Na první pohled v této proměnné nedojde k dostatečné separaci. První, druhá a třetí skupiny jsou si až moc podobné. Zkusím proměnnou zařadit do modelu.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.53243    0.05514  -9.656  < 2e-16 ***
NoEmp2      -0.07212    0.07579  -0.952    0.341
NoEmp3      -0.03426    0.08585  -0.399    0.690
NoEmp4      -0.43173    0.08274  -5.218 1.81e-07 ***
```

Dvě prostřední podskupiny jsou nevýznamné. Zkusím sloučit první tři skupiny.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.56977    0.03279 -17.375  < 2e-16 ***
NoEmp4      -0.39439    0.06986  -5.645 1.65e-08 ***
```

Tato úprava mi umožní celou proměnnou zařadit do modelu. Rovnice pro šance a interpretace koeficientů je následovná:

$$exp\,[\log\left(\frac{p}{1-p}\right)] = \exp\,(-0.57 - 0.39 * NoEmp4)$$

Šance pro první skupinu „menších firem“ a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0{,}57$$

$$p = 0.36$$

Dá se říci, že zhruba jedna ze tří malých firem, které mají méně než 7 zaměstnanců nebude schopna splatit svůj úvěr.

A pro druhou skupiny „větších firem“ je:

$$\frac{p}{1-p} = 0{,}38$$

$$p = 0.28$$

Zde můžu říci, že zhruba jedna ze čtyř firem, které mají počet zaměstnanců vyšší než 7, nebude schopna dostát svým závazkům.

Pro tuto proměnnou bude platit, že čím více zaměstnanců firma má, tím jsou šance na default nižší.

DisbursementGross

Je spojitou proměnnou v rozsahu od 5000$ do 2 688 052 $. První myšlenkou této proměnné je to, že čím větší velikost úvěru je na účet malé firmy připsána, tím hůře se bude splácet. Tuto proměnnou je třeba zkategorizovat. Nejdřív si vypočítám 25% kvantily.

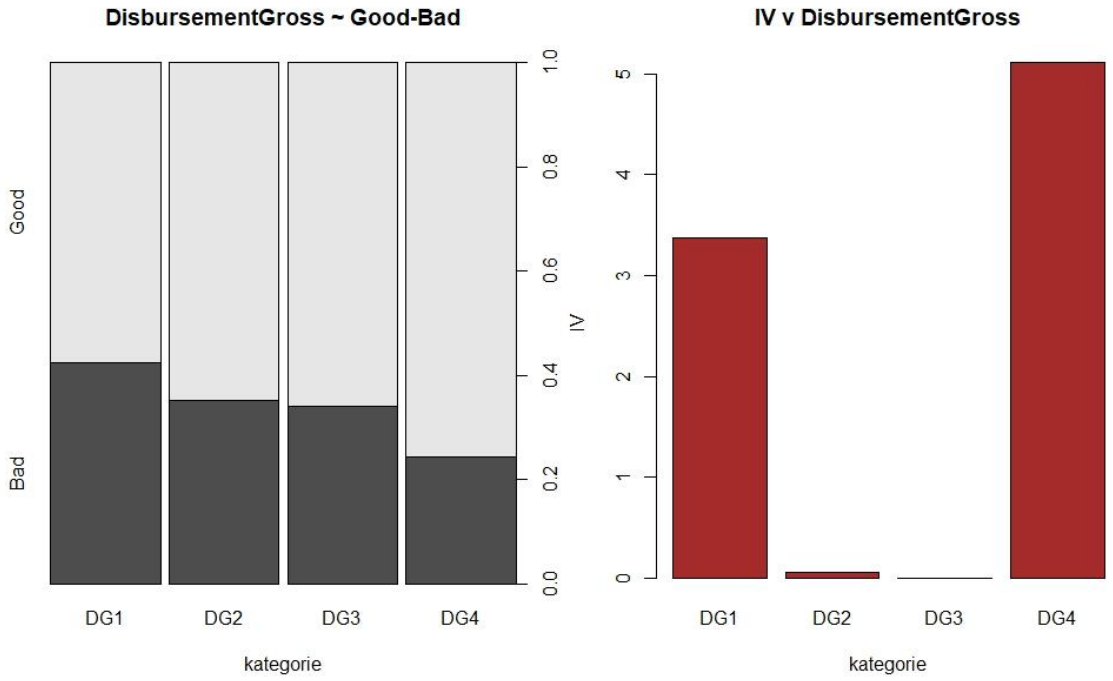
kvantil	hodnota
25%	25150
50%	66642
75%	173200
100%	2688052

Tabulka 9- kvantily proměnné Dgross

Přesně takto rozdělím tuto numerickou proměnnou na kategorizovanou proměnnou a dále s ní budu pracovat.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
DG1	569	771	31,28	21,87	1340	25,07	42,46	-3,58	3,36878
DG2	469	864	25,78	24,5	1333	24,94	35,18	-0,51	0,06528
DG3	456	880	25,07	24,96	1336	25	34,13	-0,04	0,00044
DG4	325	1011	17,87	28,67	1336	25	24,33	4,73	5,1084

Tabulka 10 - WOE a IV pro proměnnou DGross



Obrázek 10- Bivariate analysis proměnné DisbursementGross, zdroj: vlastní zpracování

Kategorie jsou rovnoměrně rozdělené a je vidět sestupný trend, tedy že čím vyšší připsaná částka na účet tím nižší je míra defaultu. Mezi druhou a třetí kategorií není téměř žádný rozdíl, při pohledu na IV graf mají obě kategorie téměř stejnou hodnotu informací, proto tyto dvě kategorie sloučím do jedné.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.30381    0.05527  -5.497 3.86e-08 ***
DG2              -0.33033    0.06862  -4.814 1.48e-06 ***
DG4              -0.83106    0.08438  -9.849  < 2e-16 ***
```

Všechny kategorie jsou významné na jakékoliv hladině významnosti.

interpretace koeficientů je následovná:

$$exp\left[\log\left(\frac{p}{1-p}\right)\right] = \exp\left(-0.30 - 0.33 * DG2 - 0.83 * DG4\right)$$

Šance pro první skupinu firem s menší velikostí úvěru (5000$ -25150$) a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0{,}74$$

$$p = 0.42$$

Dá se říci, že zhruba tři ze sedmi malých firem, kterým bude připsán úvěr ve velikosti v rozmezí 5000$ - 25000$ nebude schopna splatit tento úvěr.

A pro druhou skupiny, které žádají o „středně velký úvěr“ jsou šance a pravděpodobnost defaultu v pozorovaném vzorku následující:

$$\frac{p}{1-p} = 0{,}53$$

$$p = 0.34$$

Zhruba 2 ze tří firem, kterým bude na účet připsán „středně velký úvěr“ bude schopna úvěr splatit.

Pro poslední skupinu firem, kterým bude připsán na účet „velký úvěr“ platí následující:

$$\frac{p}{1-p} = 0{,}23$$

$$p = 0.19$$

Čtyři z pěti těchto firem budou schopni úvěr splatit.

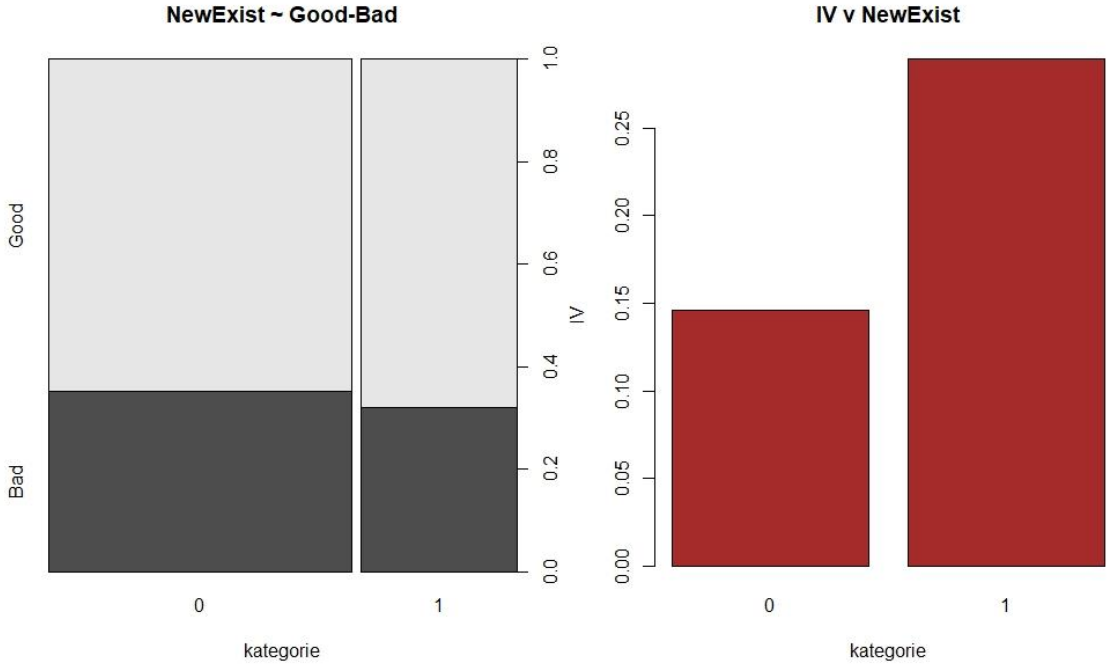
Původní myšlenka, že čím větší velikost úvěru, tím větší šance selhání se nepotvrdila. Je tomu evidentně naopak.

NewExist

Proměnná, která již je faktorizovaná, určuje, zda firma, která si zřizuje úvěr je novou firmou nebo naopak již zaběhlou firmou.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
0	1240	2294	68,17	65,06	3534	66,12	35,09	-0,47	0,14617
1	579	1232	31,83	34,94	1811	33,88	31,97	0,93	0,28923

Tabulka 11 - WOE a IV pro proměnnou NewExist



Obrázek 11- Bivariate analysis proměnné NewExist, zdroj: vlastní zpracování

Obě kategorie nejsou příliš odlišné. V první je míra defaultu 35% a v druhé 31%. Dá se tedy říct, že novější firmy mají o 4% vyšší míru defaultu než již zaběhnuté firmy.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.61519    0.03525  17.453   <2e-16 ***
NewExist1   -0.13991    0.06149   2.275   0.0229 *
```

Proměnná je významná na 5% hladině významnosti.

Rovnice pro šanci je následující:

$$exp\,[\log\left(\frac{p}{1-p}\right)] = \exp\,(-0.62 - 0.14 * NeWExist1)$$

Šance pro druhou skupinu „již zaběhnutých firem“ a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0.46$$

$$p = 0.31$$

Šanci můžu interpretovat jako, že u jedné ze tří zaběhlých firem dojde k defaultu (1/2). Pravděpodobnost defaultu se snižuje, pokud daná firma je již zaběhlá.

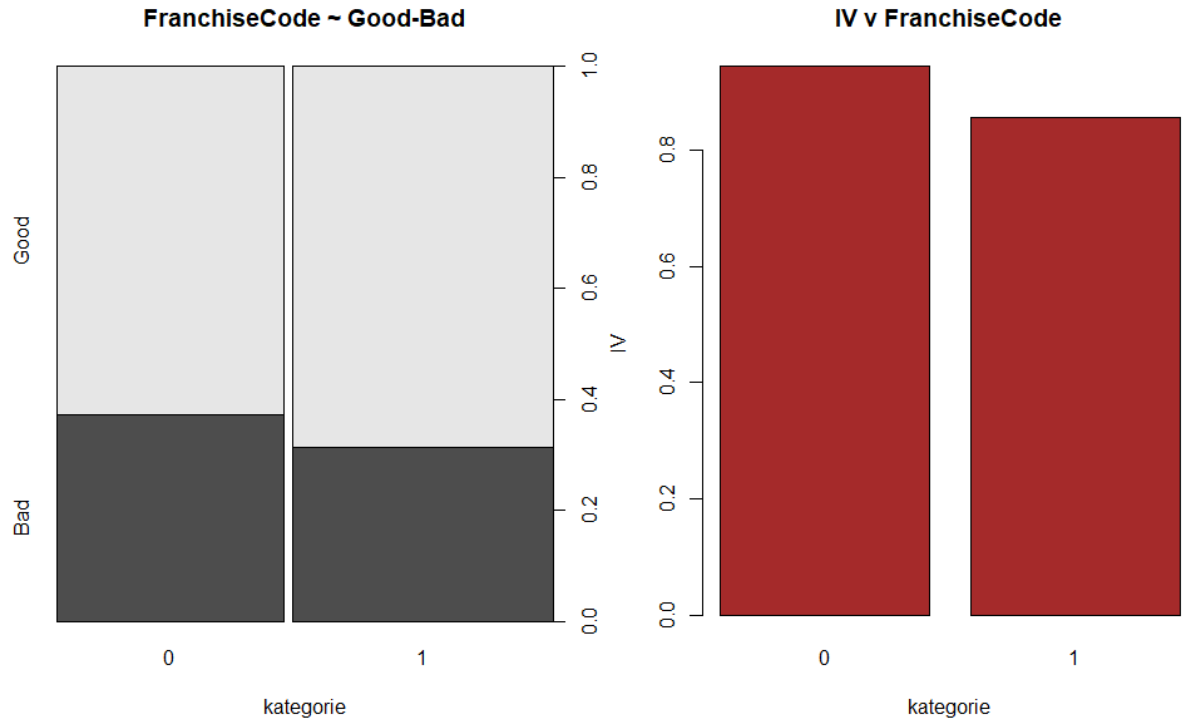
FranchiseCode

Proměnná, která říká, zda je žádající malá firma franšízou nebo nikoliv. Myšlenka ze začátku bude, že pokud je firma franšízou, bude mít menší šanci defaultovat. Jako franšíza dostane již

úspěšný „nápad“ a nemusí tolik riskovat, jako když neznámá firma přijde s novým plánem na trh.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
0	927	1561	50,96	44,27	2488	46,55	37,26	-1,41	0,94329
1	892	1965	49,04	55,73	2857	53,45	31,22	1,28	0,85632

Tabulka 12 - WOE a IV pro proměnnou FranchiseCode



Obrázek 12-univariate analysis proměnné FranchiseCode, zdroj: vlastní zpracování

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.52113    0.04147  12.568  < 2e-16 ***
FranchiseCode1 -0.26865    0.05787   4.642 3.45e-06 ***
```

Celá proměnná je významná na jakékoliv hladině významnosti. Můžu tedy potvrdit to, že pokud je daná firma franšízou, je u ní menší šance, že dojde k defaultu.

Rovnice pro šanci je následující:

$$exp\left[\log\left(\frac{p}{1-p}\right)\right] = \exp\left(-0.52 - 0.27 * FranchiseCode1\right)$$

Šance pro první skupinu firem, které nejsou franšízou a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0.59$$

$$p = 0.37$$

Šance pro druhou skupinu firem, které jsou franšízou a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0.45$$

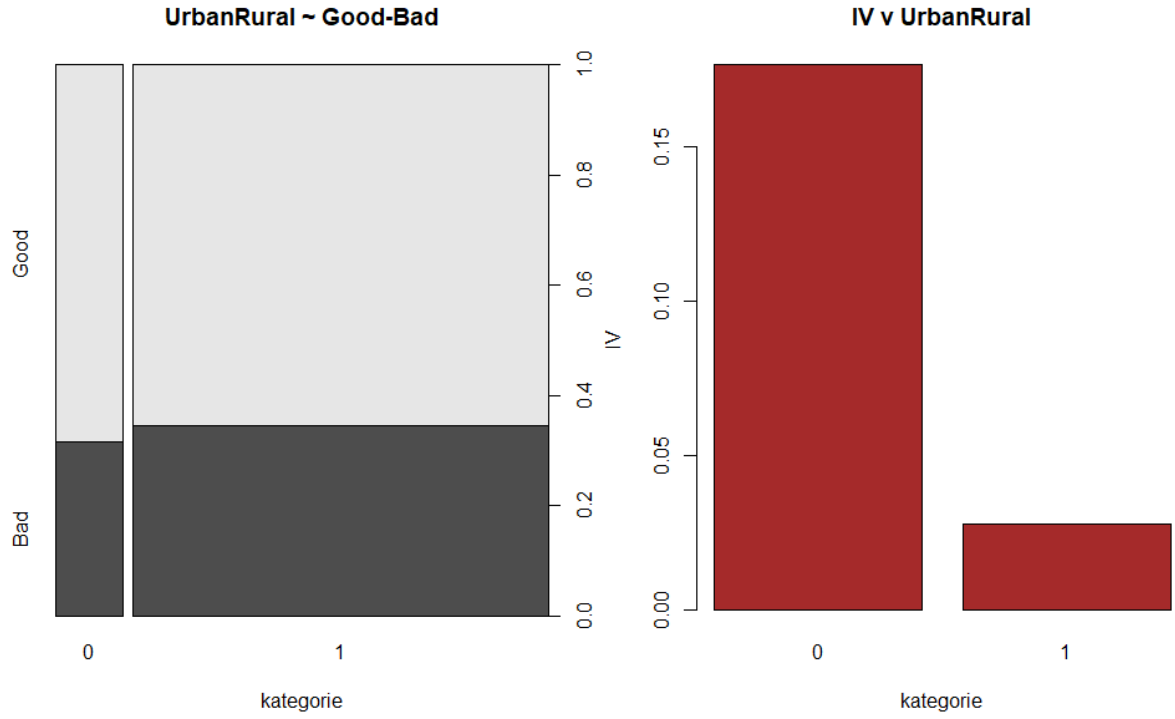
$$p = 0.31$$

UrbanRural

Proměnná, která říká, zda je firma z venkova nebo z města. Na jedné straně máme vyšší HDP a vyšší koncentraci lidí, na druhé straně ve městech jistě bude větší konkurence. Valná většina úvěrů pochází od lidí z města (86%).

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
0	234	508	12,86	14,41	742	13,88	31,54	1,14	0,1767
1	1586	3017	87,14	85,59	4603	86,12	34,46	-0,18	0,0279

Tabulka 13 - WOE a IV pro proměnnou UrbanRural



Obrázek 13- univariate analysis proměnné UrbanRural, zdroj: vlastní zpracování

Z obrázku je patrné, že nedojde k dostatečně velké separaci. U firem, které pocházejí z vesnice, je průměrná míra defaultu 31% a u firem z města je tomu 34%.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.77516    0.07901   9.811   <2e-16 ***
UrbanRural1 -0.13211    0.08488  -1.557     0.12
```

Tato proměnná není významná ani na 10% hladině významnosti a z hlediska P hodnoty se jeví jako nedůležitá. Proto ji z modelu odstraním.

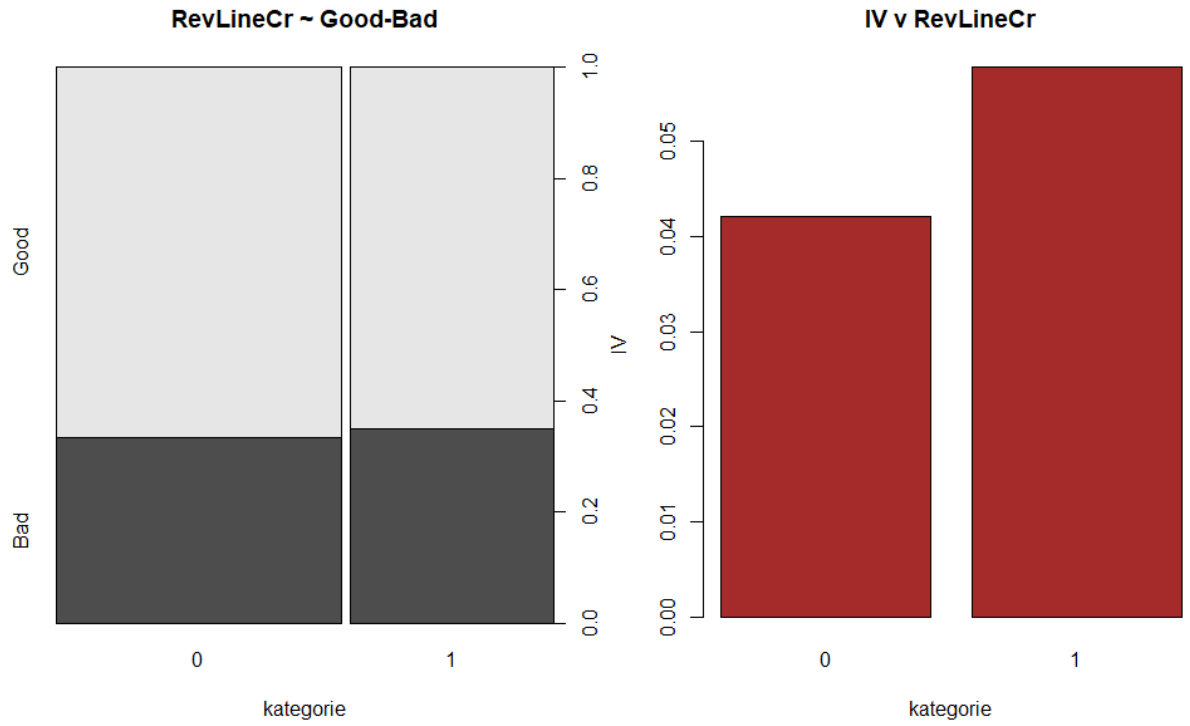
RevLineCr

Proměnná, která vyjadřuje, zda daný klient má *revolving line of credit.* Nula značí, že nemá, jednička pak že má. Následující tabulka ukazuje hodnoty IV a WOE.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
0	1044	2077	57,36	58,92	3121	58,39	33,45	0,27	0,04212
1	776	1448	42,64	41,08	2224	41,61	34,89	-0,37	0,05772

Tabulka 14 - WOE a IV pro proměnnou RevLineCR

Malé hodnoty IV napovídají, že tato proměnná nebude významná.



Obrázek 14- univarite analysis proměnné RevLineCR, zdroj: vlastní zpracování

Je patrné, že nedošlo k dostatečné separaci a proměnná pravděpodobně nebude vysvětlovat šance defaultu dostatečně době

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.68787    0.03794  18.131   <2e-16 ***
RevLineCr1  -0.06408    0.05847  -1.096    0.273
```

Tato proměnná se z hlediska P-hodnoty jeví jako nedůležitá. Z modelu ji proto odstraním.

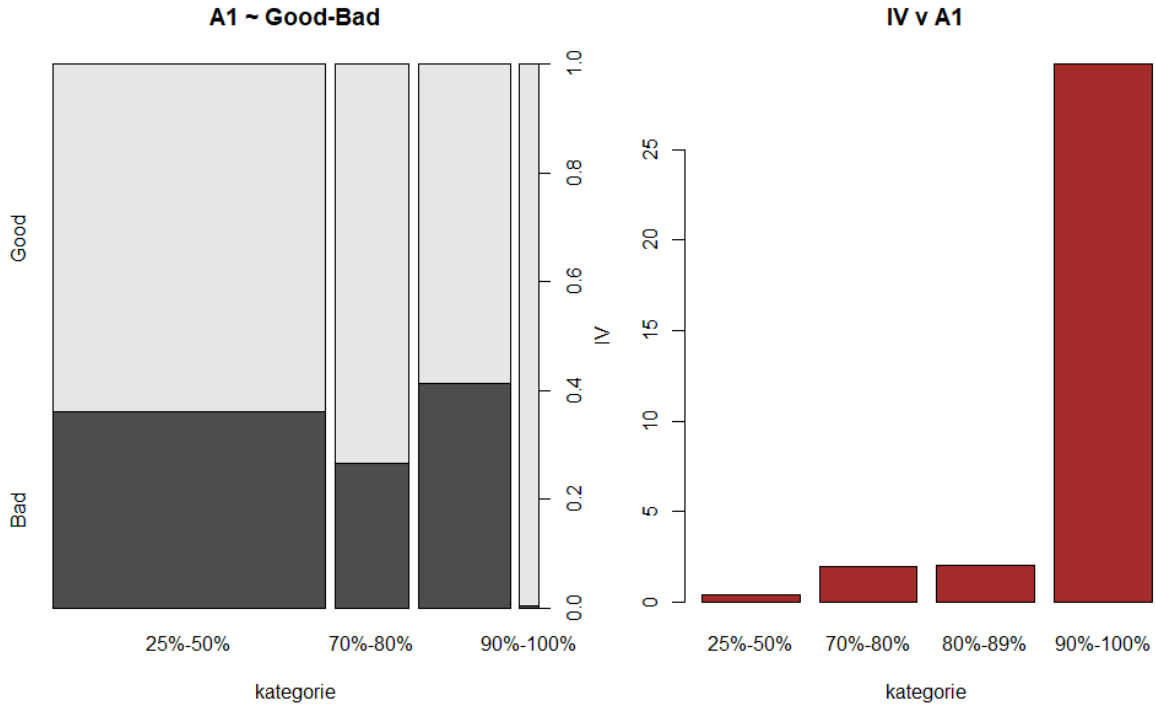
A1

Proměnná, která vyjadřuje, kolik procent z výše dluhu je pokryto garancí. Je kategorickou proměnnou se 4 kategoriemi.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
25%-50%	1143	2035	62,8	57,73	3178	59,46	35,97	-0,84	0,42588
70%-80%	230	639	12,64	18,13	869	16,26	26,47	3,61	1,98189
80%-89%	446	632	24,51	17,93	1078	20,17	41,37	-3,13	2,05954
90%-100%	1	219	0,05	6,21	220	4,12	0,45	48,22	29,70352

Tabulka 15 - WOE a IV pro proměnnou A1

V poslední kategorii, která je kryta z 90%-100% garancí původně existovalo 0 špatných úvěrů, IV a WoE pak vychází nekonečně. Proto je nutné V tomto případě alespoň jeden špatný úvěr vytvořit.



Obrázek 15- univariate analysis proměnné A1, zdroj: vlastní zpracování

Z tabulky není vidět jasný trend – předpokládal jsem, že čím vyšší krytí úvěru, tím menší procento špatných úvěrů. Proměnná se chová poněkud zvláštně. Musím se podívat na p-hodnoty.

Z obrázku se zdá, že by bylo možné sloučit dvě prostřední kategorie do jedné. Toto možné není, je to kvůli tomu, že mají odlišná znamínka u WoE. Zároveň, jak se ukáže v ve výpisu, mají parametry těchto kategorií opačné znaménko, proto není možné kategorie sloučit.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept) -0.57684    0.03696  15.606  < 2e-16 ***
A11          -0.44499    0.08532   5.216 1.83e-07 ***
A12           0.22827    0.07205  -3.168  0.00153 **
```

```
A13          -4.81223    1.00173   4.804 1.56e-06 ***
```

Všechny proměnné vycházejí důležité. Referenční kategorie je kategorie 25% - 50%, která je nejobsáhlejší. Z výpisu je vidět, že pravděpodobnost splácet je vyšší u každé kategorie kromě 80% - 89%. Tento fakt je poněkud problématický, protože interpretace této proměnné není jednoznačná. Vysvětlit, proč dochází u kategorie krytí 80% - 89% k vyšší míře defaultu než u ostatních kategorií je nemožné.

Rovnice pro šanci je následující:

$$exp\left[\log\left(\frac{p}{1-p}\right)\right] = \exp\left(-0.58 - 0.44 * A11 + 0.23 * A12 - 4.81 * A13\right)$$

Šance pro první skupinu firem, jejichž úvěry jsou kryty z 25% až 50% a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0.56$$

$$p = 0.36$$

Šance pro poslední skupinu firem, jejichž úvěry jsou kryty z 90% až 100% a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0.0046$$

$$p = 0.0046$$

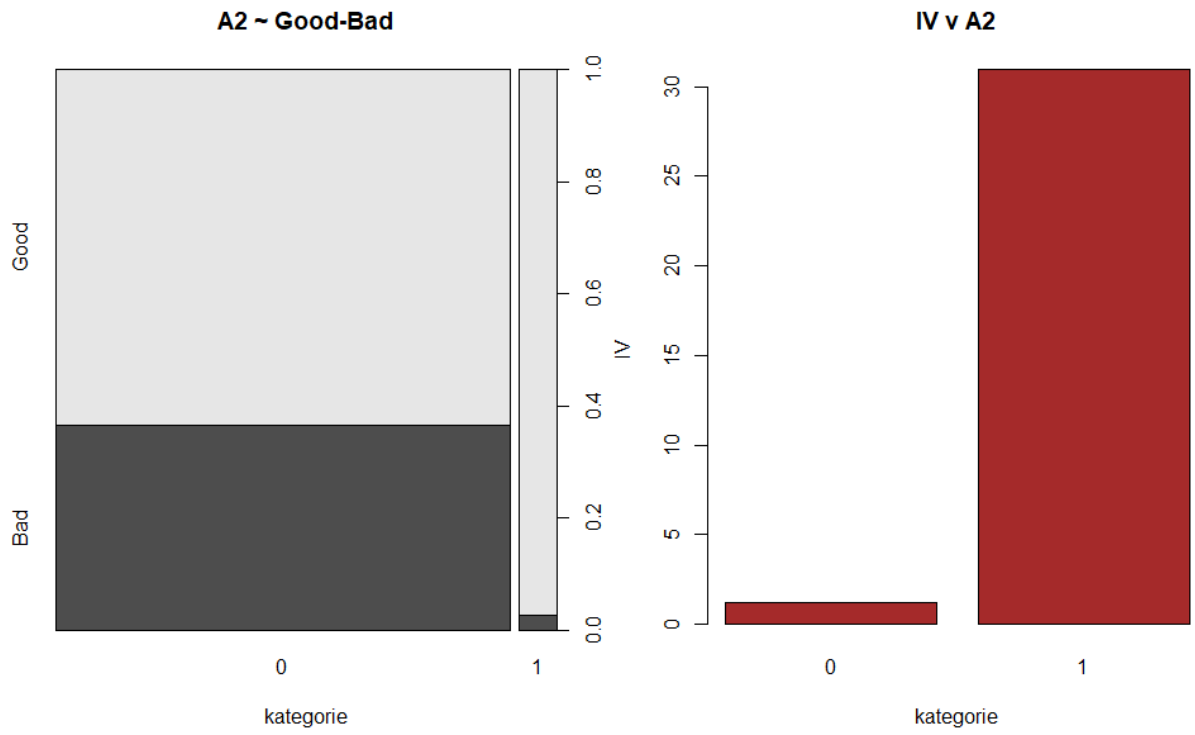
V poslední skupině nedošlo pouze k jednomu defaultu, který jsem musel přidat, aby mi vyšly rozumné IV a WoE statistiky.

A2

Proměnná zobrazující, jestli je úvěr kryt nemovitostí či nikoliv. Myšlenka je taková, že pokud bude proměnná kryta nemovitostí, pravděpodobnost splacení úvěru bude vyšší.

Names	Bad	Good	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
0	1809	3131	99,4	88,82	4940	92,42	36,62	-1,13	1,19554
1	11	394	0,6	11,18	405	7,58	2,72	29,25	30,9465

Tabulka 16 - WOE a IV pro proměnnou A2



Obrázek 16- univariate analysis proměnné A2, zdroj: vlastní zpracování

Z obrázku je patrné, že tato proměnná vysvětluje míru defaultu v jednotlivých faktorech velice dobře. V první kategorii, která není kryta nemovitostí dosahuje míra defaultu 36%. V druhé kategorii je míra defaultu pouze 2%.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.54858    0.02953  18.575   <2e-16 ***
A21          -3.02988    0.30711   9.866   <2e-16 ***
```

Rovnice pro šanci je následující:

$$exp\,[\log\left(\frac{p}{1-p}\right)] = \exp\,(-0.55 - 3.03 * A21)$$

Šance pro první skupinu firem, které nemají úvěr kryt nemovitostí a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0.58$$

$$p = 0.36$$

U každé třetí firmy, která nemá úvěr pokrytý nemovitostí dojde k defaultu.

Šance pro druhou skupinu firem, které mají úvěr krytý nemovitostí a pravděpodobnost defaultu v pozorovaném vzorku:

$$\frac{p}{1-p} = 0.028$$

$$p = 0.027$$

Pouze u tří firem ze sta, které mají úvěr pokryt nemovitostí, dojde k defaultu.

Model 1

Do prvního modelu mi vstoupí všechny proměnné, které vstupovaly do předchozí kapitoly mimo dvou – UrbanRural a Revline.

Následující tabulka shrnuje závěry univariate analýzy.

Název	Typ	významnost
MIS_Status	vysvětlovaná proměnná, 1 = default, 0 = paid off	
NAICS	Faktor. 3 úrovně. Značka průmyslového odvětví.	***
Term	Faktor. 3 úrovně. Doba úvěru.	***
NoEmp	Faktor. 3 úrovně. Počet zaměstnanců.	***
DisbursementGross	Faktor. 3 úrovně. Peníze připsané na účet.	***
NewExist	Faktor. Nové podnikaní = 1, existující = 0.	*
FranchiseCode	Faktor. Franšíza = 1, běžný podnik = 0.	***
A1	Faktor. 4 úrovně. % zajištění SBA.	***
A2	Faktor. 1 = zajištění pomocí nemovitosti, 0 = nezajištěný	***

Tabulka 17 - zbylé proměnné a jejich významnost

Všechny proměnné až na proměnnou NewExist vyšly významné na jakékoliv hladině významnosti. Nicméně dvě další proměnné – Term a A1 – jsou hůře interpretovatelné. Všechny proměnné jsou již faktorizovány a nebrání mi nic v tom, abych vytvořil první model, který bude obsahovat všechny tyto proměnné.

Jako první si všechna svoje data rozdělím na training část – zde bude obsaženo 65% všech dat a na testovací vzorek, kde bude obsaženo zbylých 35% dat.

V training sample mám celkem 2282 dobrých a 1192 špatných úvěrů (34,31% defaultní míra), test sample obsahuje 1244 dobrých a 627 špatných úvěrů (33,51% default rate).

První model na **testovacím** vzorku se všemi proměnnými má následující výstup:

```
Coefficients M1:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.62736    0.12200  13.339  < 2e-16 ***
NAICSD2               0.38448    0.09166   4.194 2.74e-05 ***
NAICSD3               0.33178    0.09766   3.397 0.000680 ***
Term64-84            -3.42528    0.09203 -37.219  < 2e-16 ***
Term85-312           -2.76337    0.14261 -19.377  < 2e-16 ***
NoEmp7-150           -0.42685    0.09645  -4.426 9.62e-06 ***
DisbursementGrossDG2 -0.50359    0.09750  -5.165 2.40e-07 ***
DisbursementGrossDG4 -0.63576    0.16460  -3.863 0.000112 ***
NewExist1            -0.18945    0.08827  -2.146 0.031864 *
FranchiseCode1       -0.65268    0.08668  -7.530 5.07e-14 ***
A170%-80%             1.23064    0.18297   6.726 1.74e-11 ***
A180%-89%             0.77232    0.11232   6.876 6.16e-12 ***
A190%-100%          -14.66384  245.33749  -0.060 0.952339
A21                  -1.83692    0.34225  -5.367 8.00e-08 ***
AIC: 4414.6
```

Většina proměnných je významná na jakékoliv hladině významnosti, až na proměnnou NewExist a poslední úroveň proměnné A1. AIC v plném modelu je 4414.6. AUC[15] je 88.05, KS je 64.72 a Gini je 76.1.

Jako první se pokusím odstranit proměnnou NewExist a sledovat, jak se změní statistiky. Po odstranění této proměnné se AIC rovná 4417.3 – vypovídací schopnost modelu se zmenšila, nicméně rozdíl v AIC je minimální.

```
Coefficients M2:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.56299    0.11790  13.257  < 2e-16 ***
D2                   0.37577    0.09154   4.105 4.05e-05 ***
D3                   0.33040    0.09758   3.386 0.000709 ***
Term2               -3.43414    0.09201 -37.323  < 2e-16 ***
Term3               -2.77805    0.14239 -19.510  < 2e-16 ***
NoEmp2              -0.39533    0.09538  -4.145 3.40e-05 ***
DG2                 -0.48988    0.09724  -5.038 4.71e-07 ***
DG4                 -0.61521    0.16402  -3.751 0.000176 ***
FC1                 -0.62519    0.08569  -7.296 2.96e-13 ***
A12                  1.17017    0.18029   6.490 8.56e-11 ***
A13                  0.70601    0.10782   6.548 5.83e-11 ***
A14                -14.69920  245.78088  -0.060 0.952310
A21                 -1.80458    0.34180  -5.280 1.29e-07 ***
AIC: 4417.3
```

AUC: 87.9, KS: 64.25, Gini: 75.8. I ostatní statistiky potvrdily, že odstraněním proměnné NewExist dojde pouze k minimální ztrátě vypovídající schopnosti tohoto modelu.

Při pohledu na proměnnou A1 je jasné, že poslední kategorie, s krytím úvěru od 90%-100% (značená jako A14), nemá žádnou vypovídací schopnost, proto tuto kategorii sloučíme s kategorií s krytím od 80% do 89% (značená jako A13). Tím vytvořím novou kategorii, která má krytí od 80% do 100%.

```
Coefficients M3:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.59480    0.11777  13.542  < 2e-16 ***
D2                  0.37222    0.09123   4.080 4.51e-05 ***
D3                  0.34055    0.09746   3.494 0.000475 ***
Term2              -3.42854    0.09193 -37.297  < 2e-16 ***
Term3              -2.78335    0.14206 -19.592  < 2e-16 ***
NoEmp2             -0.40832    0.09500  -4.298 1.72e-05 ***
DG2                -0.50597    0.09720  -5.206 1.93e-07 ***
DG4                -0.79304    0.15918  -4.982 6.29e-07 ***
FC1                -0.63395    0.08585  -7.385 1.53e-13 ***
A12                 1.34239    0.17495   7.673 1.68e-14 ***
A13                 0.62599    0.10649   5.878 4.15e-09 ***
A21                -2.23770    0.33798  -6.621 3.57e-11 ***
AIC: 4445
```

V tomto modelu došlo k nepatrnému navýšení AIC. Ostatní statistiky jsou - AUC: 87.78, KS: 64.01, Gini: 75.56, opět nedošlo k markantnímu zhoršení třetího modelu.

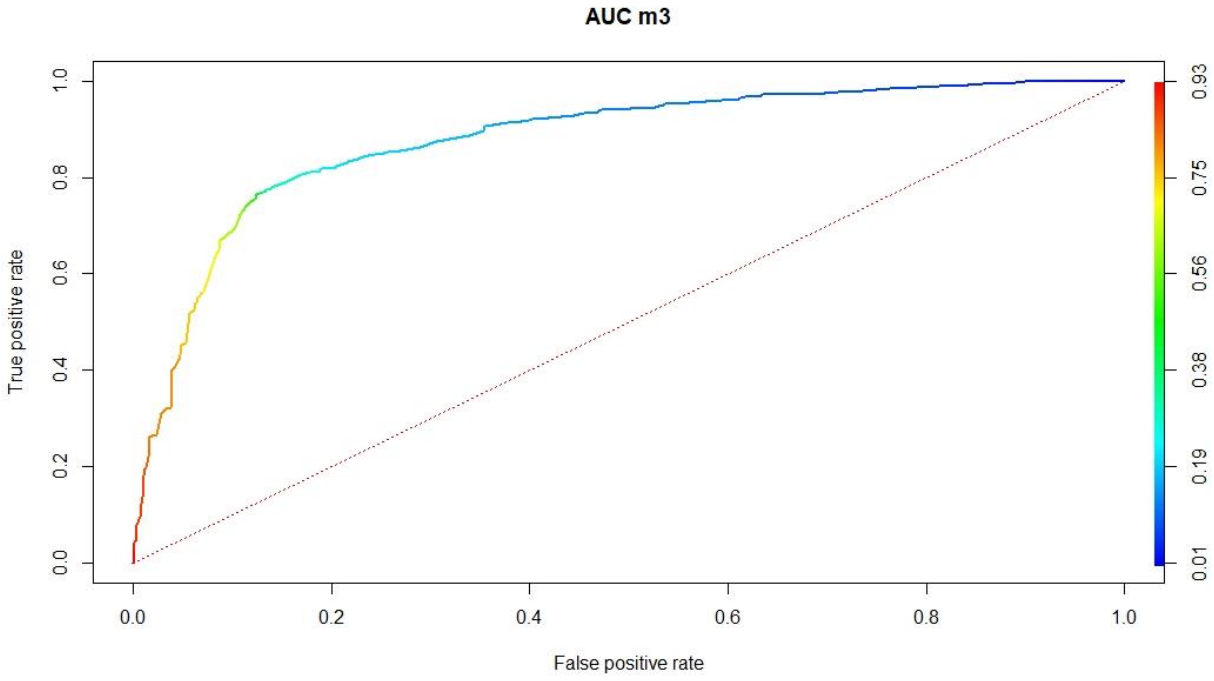
[15] Všimněte si, že platí vztah AR=2*AUC -1, uvedený v teoretické části.

Všechny proměnné a jejich úrovně jsou již významné na libovolné běžně používané hladině významnosti. Proto tento model můžu považovat za finální.

Rovnice modelu pro i-tého klienta je:

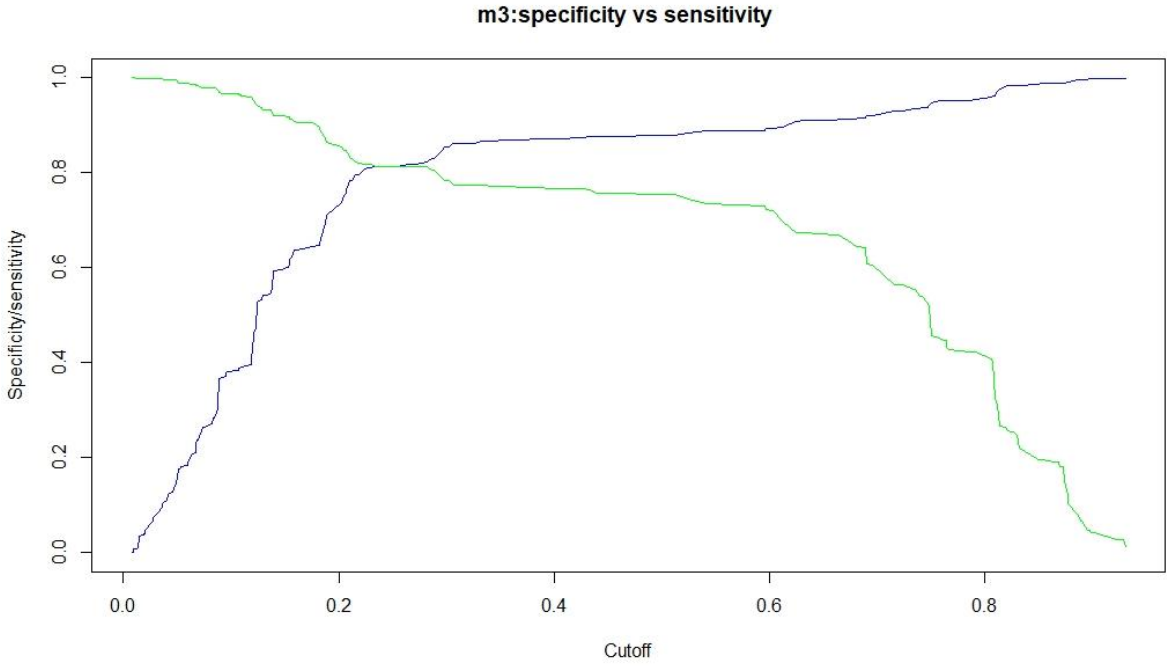
$$\log\left(\frac{p_i}{1-p_i}\right) = 1.59 + 0{,}37 * D2 + 0{,}34 * D3 - 3{,}43 * Term2 - 2.78 * Term3 - 0{,}4 * NoEmp2 - 0.5 * DG2 - 0.79 * DG4 - 0.63 * FC1 + 1.34 * A12 + 0.63 * A13 - 2.24 * A21$$

Výše uvedenou statistiku AUC, která se u toho modelu rovná 87.78, můžeme zobrazit na grafu následovně.



Obrázek 17- AUC závěrečného modelu fitovaného na populaci Texasu, zdroj: vlastní zpracování

Jako další zajímavá a důležitá křivka je křivka sensitivity a specificity, která mi pomůže zjistit, kde bych měl udělat cut-off, abych maximalizoval přesnost mého modelu.



Obrázek 18- optimal cut-off point finálního modelu fitovaného na populaci Texasu, zdroj: vlastní zpracování

Horní obrázek ukazuje, jak se budou měnit ukazatele specificity a sensitivity v závislosti na cut-off úrovni. Zároveň v bodě, kde se obě křivky protínají je optimální cut-off bod. V tomto bodě je vzdálenost obou křivek minimální. Tzn že obě dvě křivky dosahují svého společného maxima právě v tomto bodě. Tento bod leží na úrovni cut-off $= 0.288$, v tomto bodě je jak specificita tak sensitivita 0.8. Rád bych také upozornil, že úrovně těchto dvou statistik jsou poměrně blízko sebe až do úrovně cut-off $= 0.6$.

V optimálním bodě bychom poskytli úvěr pouze 1155 firmám z celkových 1871 (61%). Pokud by politika banky byla poněkud více liberální, zvolila by cut-off v bodě 0,6. V tomto případě by banka poskytla úvěr 1285 (68%) všech firem.

Pokud bych zvolil cut-off jako 0,9, pak bych už zařazoval velký počet špatných úvěrů do kategorie dobrých úvěrů – bance by z tohoto rozhodnutí plynuly zbytečné ztráty, naopak pokud bych dal cut-off level jako 0,1, pak bych zařadil až příliš dobrých úvěrů do špatné kategorie a banka by přicházela o zisky. Já zvolím jako výchozí hodnotu cut-off úrovně 0,6.

Confusion Matrix pro cut-off 0.6		**predikované hodnoty**	
		Good	Bad
skutečné hodnoty	Good	1110	134
	Bad	175	452

Tabulka 18 - Confusion matrix pro model 1

Při této úrovni cut-off můj model odhadl správně (hit rate) 1562 klientů. Z toho bylo celkem správně 1110 označeno jako dobří klienti a 452 označeno správně jako špatní klienti. Odhadl špatně celkem 309 firem – z toho je 175 špatných klientů určil jako dobrých klientů – to

znamená, že na základě tohoto modelu bych poskytl 175ti klientům úvěr, který by pak nesplatili – type 1 error. Model označil 134 dobrých klientů jako špatné – type 2 error.

Teď můžu vypočítat specificitu a sensitivitu, o kterých jsem se zmiňoval v teoretické části:

$$specificity = \frac{452}{175 + 452} = 0{,}72$$

Můj model dokáže se 72% úspěšností rozlišit špatné žadatele o úvěr. To znamená, že 28% špatných firem úvěr poskytne.

$$sensitivity = \frac{1110}{1110 + 134} = 0{,}89$$

Můj model dokáže s 89% označit dobrého klienta. To znamená, že celkem 11% dobrých klientů skončí ve špatné skupině.

Následující tabulka ukazuje, jakých hodnot nepřesnosti dosahujeme při použití různých úrovní cut-off.

cut off	*Nbg*	*Ngb*	*Ngg*	*Nbb*	*Spec*	*Sens*
0	0	1244	0	627	1	0
0.1	21	767	477	606	0,967	0,383
0.2	92	327	535	917	0,909	0,621
0.3	136	181	1063	491	0,783	0,855
0.4	146	159	1085	481	0,767	0,872
0.5	154	152	1092	473	0,754	0,878
0.6	175	134	1110	452	0,721	0,892
0.7	274	88	1156	353	0,563	0,929
0.8	372	51	1193	255	0,407	0,959
0.9	601	2	1242	26	0,041	0,998
1	627	0	1244	0	0	1

Tabulka 19 - vývoj specificity a sensitivity na různé úrovně cut-off bodu

Zajímavé je, že k minimalizaci chyby 1. a 2. typu dochází v bodě cut-off = 0,4.

Takže na otázku, kterou jsem vztyčil na začátku této kapitoly – jestli dát úvěr klientovi, který pochází z Texasu, nyní dokážu odpovědět.

Dejme tomu, že malá firma, která se jmenuje *TX*, která *není franšízou*, si chce vypůjčit *25 000$* na *60 měsíců.* Dále jsme zjistili, že firma má 3 zaměstance, a chce si založit *malý obchod se smíšeným zbožím (retail trade NAICS #45)*, zároveň úvěr není kryt nemovitostí. SBA přislíbila 50% krytí v případě defaultu.

Rovnice pro tohoto TX klienta je následující:

$$\log\left(\frac{p_{TX}}{1 - p_{TX}}\right) = 1{,}59 + 0{,}37 * D2$$

$$\frac{p_{TX}}{1 - p_{TX}} = \exp\left(1{,}59 + 0{,}37\right)$$

$$\frac{p_{TX}}{1-p_{TX}} = 7{,}09$$

Šance na nesplacení úvěry u TX firmy jsou $7x$ větší než na splacení úvěrů. Pravděpodobnost nesplacení se pak vypočte jako:

$$p_{TX} = \frac{7{,}09}{8{,}09} = 0{,}87 = 87\%$$

Na základě mnou vytvořeného modelu jsem určil pravděpodobnost nesplacení úvěru jako 87%. Této firmě bych rozhodně za těchto okolností nepůjčil. Nicméně pokud by vlastník této firmy souhlasil s tím, že banka použije jeho nemovitost jako garanci k splacení úvěru, pak by se jeho šance značně zvedly:

$$\frac{p_{TX}}{1-p_{TX}} = \exp\left(1{,}59 + 0{,}37 - 2{,}24\right)$$

$$\frac{p_{TX}}{1-p_{TX}} = 0{,}76$$

$$p_{TX} = 0{,}43 = 43\%$$

Pokud by firmy zastavila svoji nemovitost, aby dostala tento úvěr, pak by se už dalo o poskytnutí úvěru uvažovat. Záleželo by na *cut-off* úrovni, která by stanovila banka na SBA úvěry. Pokud bychom nastavili úroveň cut-off na 0,6, pak by tato firma úvěr dostala. Pokud by banka byla konzervativní a ustanovila cut-off bod na úrovni optima, pak by tato firma úvěr nedostala.

Jako poslední uvedu tabulku, kde shrnuji již uvedené statistiky v trénovacím a testovacím vzorku u tohoto finálního modelu. Pokud by došlo k výraznému zhoršení při přechodu z trénovacího vzorku do testovacího vzorku, vzniklo by podezření, že je model přefitovaný.

	Train	Test
AUC	88.61	**87.78**
KS	67.11	**64.01**
GINI	77.22	**75.56**

Tabulka 20- srovnání

Model 2

Druhý model se pokusím sestavit na celkovou populaci USA (všech 50 států). Budu zde mít o jednu proměnnou více – proměnnou State. Kterou se pokusím do modelu zapracovat tak, aby dokázala vysvětlit rozdíly v defaultní míře, která je mezi státy patrná.

Před tím, než začnu model sestavovat, se podívám, jak dobře či špatně si vede model #1 na tomto vzorku dat. Očekával bych, že si model povede hůř, jelikož jak jsem na začátku praktické části ukázal, defaultní míra napříč státy USA je značně rozdílná. Data jsou znovu rozdělena na traning (35,94% defaultní míra) a test části (36,29%).

```
Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)        0.72719    0.02636   27.591  < 2e-16 ***
D2                 0.33553    0.02288   14.664  < 2e-16 ***
D3                 0.31139    0.02476   12.577  < 2e-16 ***
Term2             -2.94338    0.02271 -129.629  < 2e-16 ***
Term3             -2.56345    0.04065  -63.060  < 2e-16 ***
NoEmp2            -0.35836    0.02395  -14.964  < 2e-16 ***
FC                -0.57387    0.02173  -26.413  < 2e-16 ***
DG2                0.02227    0.02339    0.952    0.341
DG4               -0.25967    0.03854   -6.737 1.61e-11 ***
A12                1.37390    0.04679   29.362  < 2e-16 ***
A13                0.83800    0.03126   26.812  < 2e-16 ***
A2                -1.36297    0.07174  -19.000  < 2e-16 ***
AIC: 67727
```

Model se zdá být v pořádku. Ostatní statistiky říkají, že došlo k nemalému poklesu vypovídací schopnosti tohoto modelu AUC=84.13, KS = 58.91, Gini = 68.26. Předpokládám, že toto zhoršení bylo způsobeno tím, že jsem do dat přidal novou proměnnou *State.* Proto bude třeba tuto proměnnou přidat i do modelu.

Do modelu zahrnu všechny proměnné, které jsem zahrnul do finální modelu M3 – jedná se tedy o NAICS, DisbursementGross, Term, NoEmp, FranchiseCode, A1, A2. Jediná proměnná, kterou budu muset upravovat je proměnná *State.*

Po všech úpravách, které jsem provedl při tvorbě prvního modelu, mi zbylo celkem 73 117 pozorování. V tomto vzorku je celkem 26 368 defaultů – 36,06% je průměrná míra defaultu u úvěrů s garancí SBA, které vznikly v roce 2006.

Proměnná State

Tato proměnná ukazuje, v jakém státě žadatel o úvěr bydlí, nebo v jakém státě má sídlo. Jedná se celkem o 50 států USA[16]. Hodnota defaultu se v jednotlivých státech pohybuje od 9,5% pro Wyoming do 50% pro Floridu. Kompletní seznam států s počty úvěrů, počty defaultů a defaultní mírou naleznete v příloze č. 3.

% DR	# States	# loans	Group
0-0.2	7	2070	1
0.2-0.3	18	19733	2
0.3-0.4	14	22850	3
0.4+	11	28464	4

Tabulka 21 - rozdělení State

Horní tabulka četností ukazuje kolik států spadá do určitých defaultních pásem. Podle těchto pásem státy rozdělím a dám je do patřičných skupin, skupina 1 je nejméně riziková, skupina čtyři je pak nejvíce riziková.

Pro tuto proměnnou vykonám stejnou bi-variate analýzu jako tomu bylo v předchozí části. Začnu tabulkou, která mi ukazuje IV a WoE statistiky.

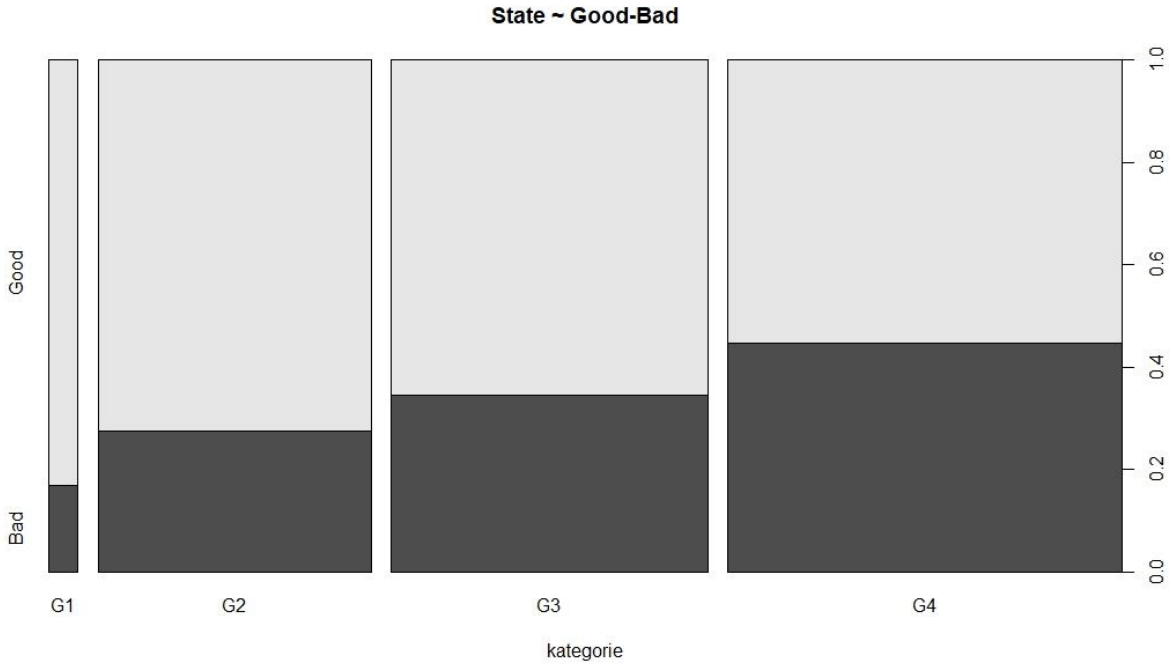
Names	0	1	Bad_pct	Good_pct	Total	Total_Pct	Bad_Rate	WOE	IV
G1	1722	348	3,68	1,32	2070	2,83	16,8	-10,25	2,419
G2	14298	5435	30,58	20,61	19733	26,99	27,5	-3,95	3,93815
G3	14972	7878	32,03	29,88	22850	31,25	34,4	-0,69	0,14835
G4	15757	12707	33,71	48,19	28464	38,93	44,64	3,57	5,16936

Tabulka 22 - IV a WOE proměnné State

Vidím, že IV i Woe jsou odlišné pro všechny kategorie, nebude proto možno žádnou kategorii sloučit.

[16] DC není v datech zahrnováno.

Jako další se podívám na graf, který ukazuje, jak jsou jednotlivé kategorie (G1- G4) zastoupeny v počtu špatných a dobrých úvěrů.



Obrázek 19- rozložení good-bad úvěrů v upravené proměnné State, zdroj: vlastní zpracování

Zdá se, že jednotlivé kategorie jsou od sebe dostatečně odlišeny. Jako poslední mi zbývá dát pouze proměnnou state do modelu a zjistit, jak dobře mojí vysvětlovanou proměnnou (MIS_Status) tato proměnná vysvětluje.

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.59904    0.05877  -27.21   <2e-16 ***
StateG2      0.63178    0.06090   10.38   <2e-16 ***
StateG3      0.95693    0.06040   15.84   <2e-16 ***
StateG4      1.38391    0.05997   23.08   <2e-16 ***
```

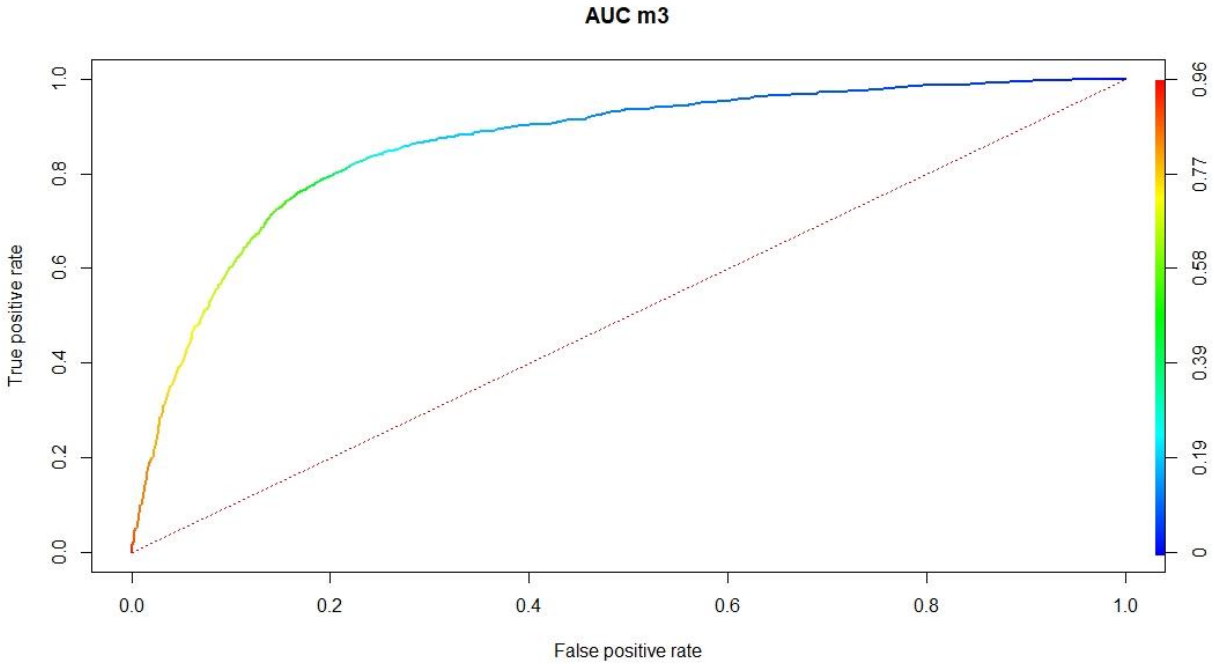
Zjistil jsem, že všechny kategorie jsou významné na jakékoliv hladině významnosti.

Nyní už mi nic nebrání v tom, abych to upraveného modelu zahrnul proměnnou State.

```
Coefficients:
                 Estimate Std. Error  z value Pr(>|z|)
(Intercept)      -0.78075    0.07082  -11.025  < 2e-16 ***
State2            0.93194    0.06860   13.585  < 2e-16 ***
State3            1.43421    0.06825   21.016  < 2e-16 ***
State4            2.07117    0.06804   30.441  < 2e-16 ***
D2                0.35570    0.02344   15.172  < 2e-16 ***
D3                0.31068    0.02536   12.253  < 2e-16 ***
Term2            -3.06640    0.02360 -129.920  < 2e-16 ***
Term3            -2.66560    0.04229  -63.033  < 2e-16 ***
NoEmp2           -0.34234    0.02454  -13.950  < 2e-16 ***
FranchiseCode    -0.56464    0.02236  -25.247  < 2e-16 ***
DG2               0.05793    0.02392    2.422   0.0154 *
DG4              -0.28496    0.03969   -7.179 7.02e-13 ***
A12               1.43302    0.04800   29.852  < 2e-16 ***
A13               0.88925    0.03200   27.791  < 2e-16 ***
A2               -1.51237    0.07270  -20.802  < 2e-16 ***
```

AIC:64914

Model je rozhodně lepší než přechodzí, statistika AIC je nižší a dokonce kategorie DG2 se stala významnou na 5% hladině významnosti. Ostatní statistiky potvrdily, že tento model vysvětluje nová data lépe – AUC = 86.1, KS=59.69, Gini = 72.2.



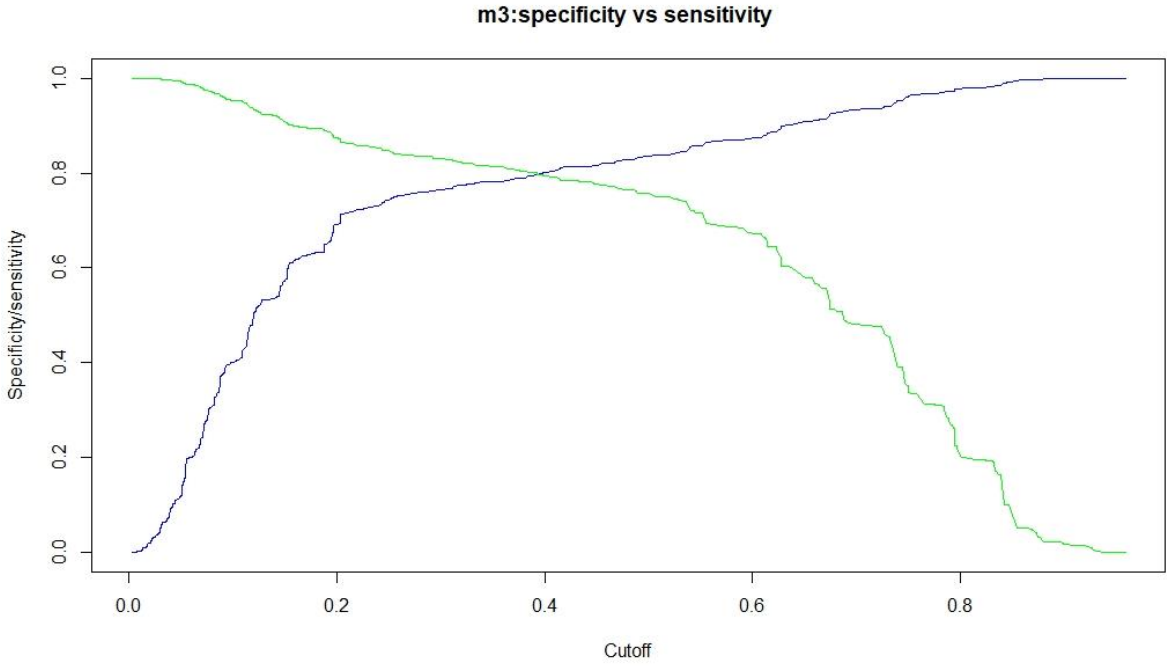
Obrázek 20- AUC finálního modelu fitovaného na celou populaci USA, zdroj: vlastní zpracování

Rovnice modelu pro i-tého klienta je:

$$\log\left(\frac{p_i}{1-p_i}\right) = -0.78 + 0.93 * State2 + 1.43 * State3 + 2.07 * Sate4 + 0{,}36 * D2 + 0{,}31 * D3 - 3{,}07 * Term2 - 2.67 * Term3 - 0{,}34 * NoEmp2 - 0.56 * DG2 - 0.28 * DG4 - 0.56 * FC1 + 1.43 * A12 + 0.89 * A13 - 1.51 * A21$$

Z výpisu lze vidět, že nově přidaná proměnná *State* má největší vliv na zvýšení log-odds modelu (konkrétně State4 zvýší log-odds nesplacení o 2.07).

Optimální cut-off bod je u tohoto modelu v $c = 0{,}4$, kde sensitivita a specificita dosahují hodnoty 0,8.



Obrázek 21- Cut-off point finálního modelu fitovaného na celou populaci USA, zdroj: vlastní zpracování

V tomto bodě by můj model zamítl 10606 žádostí o úvěr a přijal celkem 14985 žádostí o úvěr.

Confusion Matrix pro cut-off 0.4		**predikované hodnoty**	
		Good	Bad
skutečné hodnoty	Good	13068	3236
	Bad	1917	7370

Tabulka 23 - Confusion matrix pro model 2

Test část obsahuje 9287 špatných úvěrů a 16304 dobrých úvěrů.

Při této úrovni cut-off můj model odhadl správně (hit rate) 20438 klientů. Z toho bylo celkem správně 14985 označeno jako dobří klienti a 10606 označeno správně jako špatní klienti. Odhadl špatně (misklasifikoval) celkem 5153 firem – z toho je 1917 špatných klientů určil jako dobrých klientů – to znamená, že na základě tohoto modelu bych poskytl 1917ti klientům úvěr, který by pak nesplatili – type 1 error. Model označil 3236 dobrých klientů jako špatné – type 2 error.

Teď můžu vypočítat specificitu a sensitivitu, o kterých jsem se zmiňoval v teoretické části:

$$specificity = \frac{7370}{7370 + 1917} = 0{,}8$$

Tento model dokáže se 80% úspěšností rozlišit špatné žadatele o úvěr. To znamená, že 20% špatných firem úvěr dostane. V původním test vzorku bylo celkem 9287, tento model označil jako špatnou firmu správně 7370 firem – to je 80%.

$$sensitivity = \frac{13068}{13068 + 3236} = 0{,}8$$

Můj model dokáže s 80% označit dobrého klienta. To znamená, že celkem 20% dobrých klientů skončí ve špatné skupině. V původním test vzorku bylo celkem 16304 dobrých úvěrů. Sensitivita na úrovni 0,8 způsobila, že bylo správně označeno 13068 firem – toto je přesně 80%.

Následující tabulka ukazuje, jakých hodnot nepřesnosti dosahujeme při použití různých úrovní cut-off.

cut off	*Nbg*	*Ngb*	*Ngg*	*Nbb*	*Spec*	*Sens*
0	0	16304	0	9287	1	0
0.1	433	9778	6526	8854	0,953	0,400
0.2	1169	5032	11272	8118	0,874	0,691
0.3	1537	3837	12467	7714	0,834	0,765
0.4	*1917*	*3236*	*13068*	*7370*	*0,794*	*0,802*
0.5	2254	2702	13602	7033	0,757	0,834
0.6	3047	2047	14257	6240	0,672	0,874
0.7	4833	1064	15240	4454	0,480	0,935
0.8	7414	315	15947	1873	0,202	0,981
0.9	9125	23	16281	162	0,017	0,999
1	9287	0	16304	0	0	1

Tabulka 24 - vývoj specificity a sensitivity na různé úrovně cut-off bodu modelu 2

Stejně jako u minulého modelu je nutno uvést statistiky v testovacím a trénovacím vzorku.

	Train	Test
AUC	87.3	**86.1**
KS	63.02	**59.69**
GINI	73.6	**72.2**

Tabulka 25- srovnání

Opět jako u v minulé části došlo k jemnému zhoršení predikční síly modelu. Nicméně zhoršení není nijak závažné, můžu tedy tvrdit, že model není přefitovaný.

Závěr

V teoretické části jsem položil základy nutné teorie, která je potřebná k budování skóringového modelu založeného na logistické regresi. Tato část je rozdělena na část, která popisuje logistickou regresi a na část, která popisuje samotné modelování.

V praktické části jsem vybudoval dva logistické modely na historických datech – modeloval jsem tedy reálné pravděpodobnosti defaultu. Tato reálná data jsou data SBA garantovaných úvěrů. První model byl založený na vzorku užší populace – pouze stát Texas a pouze úvěry vzniklé v roce 2006. V druhém modelu jsem se snažil vyvinout skóringový logistický model pro celou populaci USA z roku 2006 – tohoto jsem docílil tak, že jsem použil předchozí model pro jeden stát a vzniklé rozdíly jsem vysvětlil pomocí nové proměnné, která měla zachytit různorodé ekonomické, kulturní a sociální podmínky států USA. Otázkou je, co by bylo užitečnější – vytvořit 50 modelů pro každý stát zvlášť, nebo vytvořit pouze jeden model. Já osobně se přikláním k druhé možnosti, kdy za použití pouze jedné proměnné navíc (která má celkem 4 úrovně) jsem dokázal vysvětlit 15x větší datový soubor.

Oba dva modely, by podle pánu Hosmera a Lemeshowa, byly hodnoceny jako velice dobré. Fakt, který jsem doložil na třech statistikách – AUC, Gini a KS.

Do přílohy jsem pak přesunul celou prvotní analýzu dat, která je sice integrální částí tvorby jakéhokoliv modelu, nicméně časté opakování a poměrně jednoduchá popisová statistika hodná maximálně bakalářské práce vysloužily místo maximálně v příloze. Tato kapitola je doprovozena kódem z programu R a mými komentáři.

Pro ty čtenáře, kteří si čtou pouze závěr, co si z diplomové práce odnést?

V úvodu této práce jsem mluvil o regulaci a dozoru v podobě Basel II. Je nutné si uvědomit, že skóringové a případně jiné risk manažerské modely nevznikají jen kvůli dohledu. Banka by stála sama proti sobě, kdyby nevěděla, jak se její portfólio úvěrů bude vyvíjet v budoucnu, s jakými ztrátami může počítat.

Proč zrovna logistická regrese? Hlavní důvod, proč je logistická regrese nejpoužívanější metodou modelování binární vysvětlované proměnné spatřuji v tom, že výsledný vektor kreditního skóre je velice lehce interpretovatelný a zároveň v tom, že logistická regrese nemá žádné požadavky na proměnné vstupující do modelu.

Na začátku této diplomové práce jsem mluvil o *krednitním skóre*, které je přiřazováno všem žadatelům o úvěr. Toto skóre je pravděpodobnost nesplacení úvěru, kterou odhadne bankou vytvořený model. Nejlépe hodnocený žadatel by dostal skóre 0 – není vůbec žádná šance, že by nesplatil úvěr. Nejhůře hodnocený žadatel by dostal skóre 1 – je 100% šance, že úvěr nesplatí.

Jak má banka zjistit, kde udělat pomyslnou čáru, kterou když skóre jednotlivého žadatele přesáhne, jeho žádost o úvěr bude automaticky zamítnuta? Tento bod se dá najít vícero způsoby, v této práci jsem volil bod, kde se specificita rovná sensitivitě. Zároveň se dá

považovat za rozumné tento bod hledat v místě, kde chyba typu 1 a chyba typu 2 je minimální.

A jako poslední bych rád zmínil, že v dnešní době je absolutně nutné umět ovládat nějaký statistický program – Python, R, Matlab. Časy papírů a tužek jsou dávno pryč a časy excelu skončily na střední škole.

Příloha 1 – popisná část univariate analysis

V této části uvádím všechny proměnné, které se vyskytovaly v původních datech, jak jsem se s nimi vypořádal a přidávám napsaný kód – který bude zapsán jiným písmem, a to Times New Roman. Začnu nahráním dat do programu R.

```
data <- read.csv("C:/ ")
```

Úvodní manipulace dat

Jako první věc, co musím udělat je změnit to, jak je default definován v datech. Pokud dojde k defaultu (původně značeno CHGOFF) označím to jako 1, pokud dojde k plnému zaplacení (značeno Paid) označím jako 0.

```
data$MIS_Status <-as.factor(ifelse(data$MIS_Status =="CHGOFF",'1',"0"))
```

Z původních mohu zjistit nějaké zajímavé charakteristiky, které mají všechny tyto úvěry společné. Tyto charakteristiky mi dále pomohou rozhodnout, jaký časový úsek zahrnout do dat, s kterými budu vytvářet model.

Jedná se hlavně o průměrnou dobu, na jakou je úvěr sjednán.

```
summary(data$Term)
```

výše uvedená řádka nám řekne, že průměrný počet měsíců, na který je úvěr sjednán je 111 – tedy téměř 10 let.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	60.0	84.0	110.8	120.0	569.0

Tato skutečnost mi říká, že nemůžu zařadit do dat pro modelování „nové“ úvěry, jelikož nejsou dostatečně staré na to, aby buď dosáhly defaultu nebo naopak bylo jasné, že se úvěr splatí.

Další důležitou charakteristikou je průměrná doba do defaultu. Nejdříve je potřeba z původních dat vyfiltrovat jen ta, u kterých došlo k defaultu.

```
dataD =data[data$MIS_Status == c("1"), ]
```

Dále si vyberu pouze ty sloupce, které k této analýze potřebuji

```
dataD2 = select(dataD, ApprovalDate, ChgOffDate)
```

v těchto datech je datum zaznamenáno jako 01-JAN-99 – což znamená 1.1.1999. Program R s tímto formátem neumí pracovat, zároveň je nemožné změnit pouze část buňky, proto se toto datum musí upravit externě na formát 01-01-1999, s kterým již program R umí pracovat. Proto si vyvedu tato data z Rka do .xls.

```
write.xlsx(dataD2, "c:/ ")
```

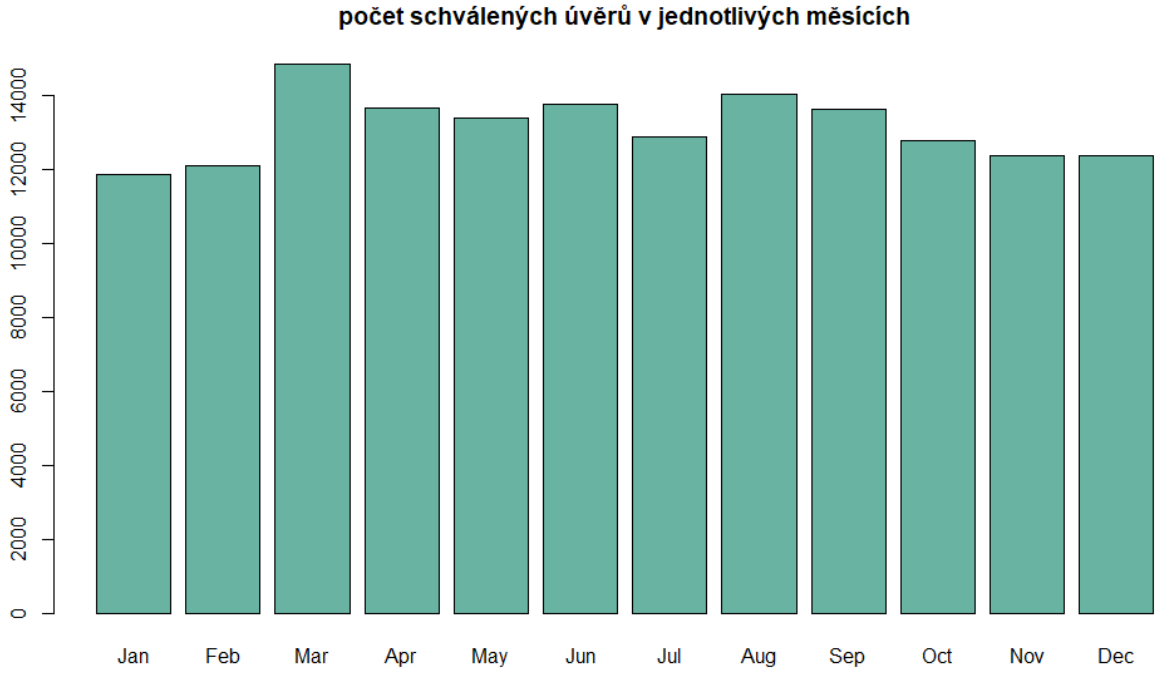
Následující tabulka shrnuje poznatky z výsledného upraveného souboru.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	1019	1471	1642	2069	12930

Čili průměrná doba, za kterou úvěr defaultuje je 1642 dní. To je zhruba 4,5 roku.

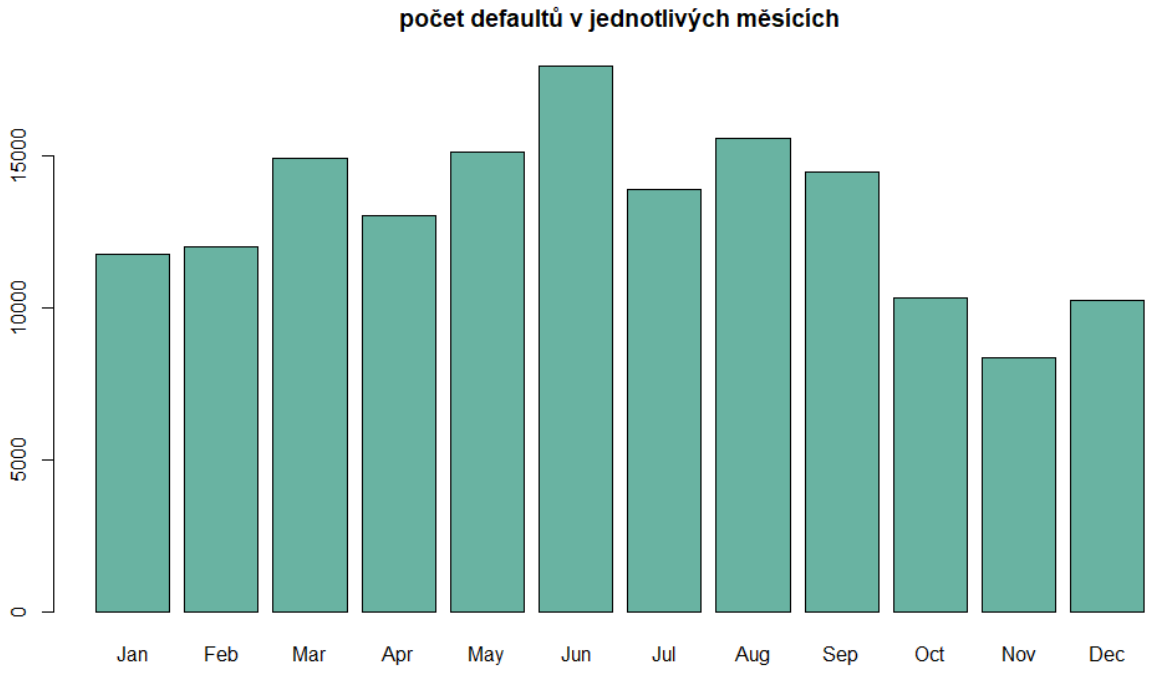
Následující dva grafy (pro zajímavost) znázorňují, v jakém měsíci nejčastěji docházelo k defaultu a v jakém měsíci si klienti nejčastěji brali úvěr.

```
barplot(approval$count, main = 'počet schválených úvěrů v jednotlivých měsících',names.arg = approval$month , col="#69b3a2",horiz = FALSE)
```



Obrázek 22 - distribuce schálených úvěrů v jednotlivých měsících

První graf znázorňuje, v jakém měsíci si klienti brali nejčastěji úvěry – není překvapivé, že během roku nedochází k nijak zvláštní variaci.



Obrázek 23 - distribuce defaultů v jednotlivých měsících

Tento graf je již vykazuje jistou varianci v počtu defaultů v jednotlivých měsících, co je zajímavé, že nejčastěji k defaultu dochází v polovině roku, naopak nejméně defaultů je zaznamenáno na konci roku kolem v měsících listopadu a prosinci a zároveň na začátku roku v lednu a únoru.

Z původních 877 500 dat vyfiltruji pouze ty, jejich proměnná *ApprovalFY* je 2006.

```
data1 =SBAnational[SBAnational$ApprovalFY == c("2006"), ]
```

V tom pod-vzorku mám celkem 75 756 dat – 26 517 defaultů a 49 239 zaplacených úvěrů – celkem tedy 35% šance defaultu. Tato hodnota je průměrnou hodnotou defaultu na úvěrech pro malé firmy, které jsou garantované státem.

Dále mě bude zajímat, jak je míra defaultu rozložena po jednotlivých Státech USA – k tomu slouží již uvedená „Heat Map“.

Na tuto mapu potřebuji pouze proměnné míry defaultu a státu

```
newd = select(data1, MIS_Status, State)
```

Dále musíme zagregovat takto vyselektovaná data podle států – tím se nám vytvoří 50 řádek statů a k ním patřičný počet defaultů za sledovaný rok 2006

```
newd2= aggregate(newd[,"MIS_Status"], by=list(newd$State), FUN =mean, na.rm= TRUE)
```

Stáhnu balíček googleVis a udělám mapu podle následujícího kódu.

```
library(googleVis)

st= read_excel("C:/Users/JIRINA/Desktop/DIMPLO/states.xlsx")

GeoStates <- gvisGeoChart(st, "state.name", "default %",
```

```
                options=list(region="US",
                             displayMode="regions",
                             resolution="provinces",
                             width=600, height=400))
plot(GeoStates)
```

Dále vyfiltruji jen ta data, která pochází z Texasu.

```
data2 =data1[SBAnational$State == c("TX"), ]
```

tento pod-vzorek obsahuje celkem 5380 dat – z toho je 1761 defaultů a 3619 úspěšně splacených úvěrů – což mi dává průměrnou míru defaultu za rok 2006 v Texasu 32%. Což nám dává podprůměrnou míru defaultu měřeno k celku USA v roce 2006.

Takto připravená data budu nadále upravovat po jednotlivých proměnných. Po tom, co prohlédnu všechny proměnné modelu a vyřadím ty, které nebudou důležité, rozdělím data na 2 části – trénovací a validační.

Na trénovacím vzorku naučím model a na validačním vzorku ho pak vyzkouším a zjistím, jak dobrou má klasifikační schopnost.

LoanNr_ChkDgt

První proměnná, která přiřazuje ke každému žadateli o úvěr vlastní číslo. Rovnou je jasné, že takovouto proměnnou v modelu nebudu zahrnovat, proto ji z dat odstraníme.

```
data2 = data2[,-1]
```

Name

Proměnná, která udává jméno klienta, které jsou v absolutní většina jména firem, je jasné, že tato proměnná nebude nijak ovlivňovat šanci splatit úvěr. Proto ji z dat odstraníme.

```
data2 = data2[,-1]
```

City

Zajímavější proměnná než dvě předchozí, nicméně na první pohled je jasné, že nijak nebudu ovlivňovat pravděpodobnost splacení.

Tato proměnná obsahuje 1100 různých měst v Texasu. Pouze v 6 z nich s schválilo více než 100 úvěrů a to konkrétně:

Město	# úvěrů	# obyvatel k 2006[17]	% populace	# defaultu	% defaultu
Houston	876	2099451	0,04%	325	37,10%
Dallas	349	1197816	0,03%	121	34,67%
San Antonio	278	1327407	0,02%	99	35,61%
Austin	232	790390	0,03%	71	30,60%
El Paso	175	649121	0,03%	45	25,71%
Fort Worth	138	741206	0,02%	35	25,36%
Arlington	128	365438	0,04%	54	42,19%
Plano	128	259841	0,05%	52	40,63%

Tabulka 26 - největší města v Texasu a jejich míra defaultu

Poměr počtu úvěrů k celkové populaci města se u všech měst pohybuje okolo 0,04%. Zároveň ani nemůžu říci, že % defaultu je větší/menší než průměr (připomínám, že průměr míry defaultu v Texasu je 32%.)

```
data2 = data2[,-1]
```

Nicméně si myslím, že bude vhodné tuto proměnnou prozkoumat tak, že ji rozdělím na dvě skupiny:

1) Velká města – města s počtem obyvatel > 100 000
2) Malá města – města s počtem obyvatel < 100 000

Dá se říci, že obyvatelé ve velkých městech disponují větším disponibilním příjmem, proto je možné, že budou schopni splácet úvěr lépe než obyvatelé z venkova. Nicméně tento poznatek je již zobrazen v další proměnné a to UrbanRural

State

Proměnnou státu nemusíme řešit, jelikož máme všechny poskytnuté úvěry v jediném státě a to Texasu. Nicméně z nad-státního pohledu je to velice důležitá proměnná, což potvrzuje výše uvedená mapa.

Tato proměnná je důležitá hlavně z ekonomický důvodů. HDP na hlavu je v USA značně variablitní napříč všemi státy. Texas měl v tomto roce HDP per capita 46 680 USA a byl 13. nejbohatším státem v USA. Nejbohatší stát měl HDP per capita ve výši 72 204 USD a nejchudší 31 658 USD

```
data2 = data2[,-1]
```

ZIP

Proměnná obsahuje 1000 různých ZIP čísel, tato proměnná může být důležitá – malá firma založená v chudé čtvrti nebude tak prosperovat jako firma založená v bohatší čtvrti. Nicméně

[17] Dostupné z https://www.tsl.texas.gov/ref/abouttx/popcity32010.html

analýza této proměnné by byla extrémně náročná – ZIP čísla bych musel nastudovat u celého Texasu a pak zároveň nějakým způsobem zjistit, zda je čtvrť považována za chudinskou/nebezpečnou.

```
data2 = data2[,-1]
```

Bank Name

Název banky poskytující úvěr samozřejmě nebude mít žádný vliv na pravděpodobnost splácet.

JPMorgan Chase Bank	1340
Bank Of America	686
Wells Fargo Bank	479
BBCN Bank	323
Compass Bank	270
Capital One	66
Business Loan Center	120
United Central Bank	108

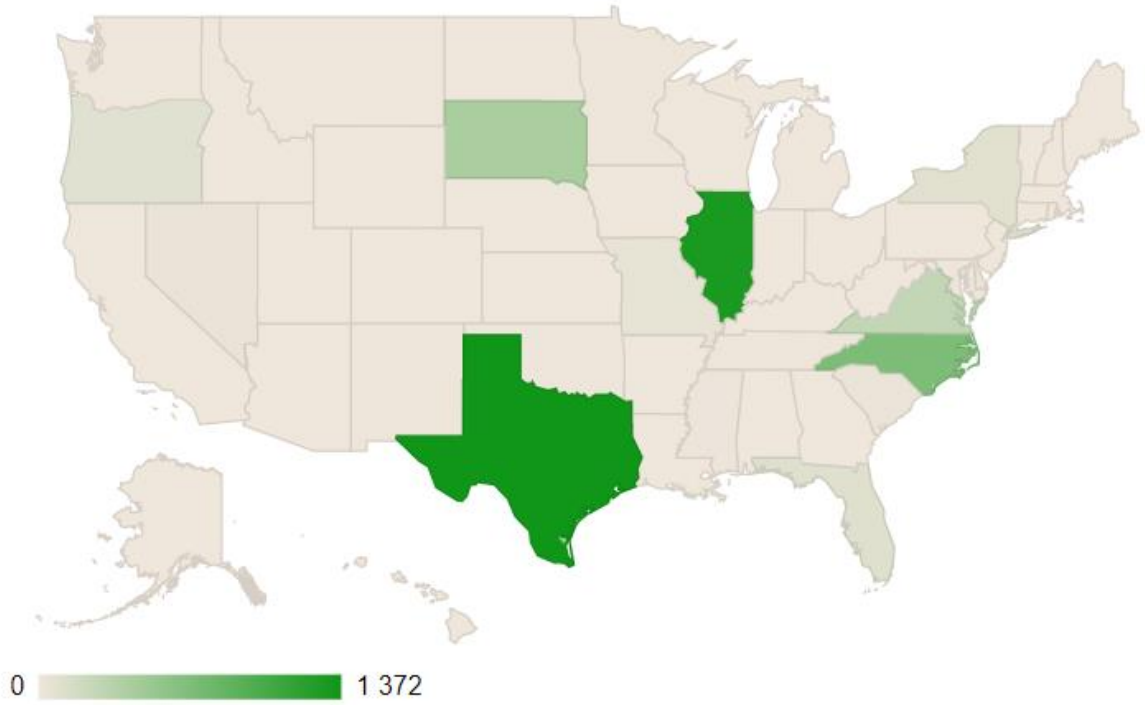
Tabulka 27 - Nejvýznamější banky Texasu

Pouze uvedu banky, které poskytly více než 100 úvěrů – Těchto osm bank dohromady poskytlo 3392 úvěrů což tvoří kolem 60% poskytnutých úvěrů v roce 2006

Bank_State

Jedná se o proměnou, která znázorňuje, z jakého státu si podniky z Texasu nejčastěji půjčují. Nepřekvapivě vede Texas s 1372 úvěry, druhá už je pak poměrně vzdáleny stát Illinois a třetí Severní Karolína se 701 úvěry.

Tuto proměnnou bychom mohli překlasifikovat na domací banku = 0 a mimostátní banku = 1



Obrázek 24 - domácí banka vs mimostátní banka

NAICS[18]

Tato zkratka znamená North American Industry Classification systém. Jde o to, že každý průmysl dostane svojí vlastní číselnou značku. Například pokud firma, která chce podnikat ve financích, zažádá o úvěr, který bude garantován SBA, tak dostanou značku 48 nebo 49[19].

Jak je znázorněno na spodní tabulce, tak druh průmyslu samozřejmě ovlivňuje pravděpodobnost defualtu. Nejnižší míry defaultu se ukazují u manuálních pracích – práce v dolech, lovení zvěře. Nejvyšší míry defaultu pak dosahují firmy, které chtějí podnikat v realitách.

[18] Více informací o této značce na https://www.naics.com/search/

[19] Pokud jsou dvě značky vedle sebe, jsou sloučené dvě velice podobně odvětví – pojišťovnictví a bankovnictví je v tomto seznamu uvedeno pod jedním druhem průmyslu a to Finance and insurance.

číselná značka	druh průmyslu	Míra defaultu %
21	Mining, quarrying, and oil and gas extraction	8
11	Agriculture, forestry, fishing and hunting	9
55	Management of companies and enterprises	10
62	Health care and social assistance	10
22	Utilities	14
92	Public administration	15
54	Professional, scientific, and technical services	19
42	Wholesale trade	19
31-33	Manufacturing	16%
81	Other services (except public administration)	20
71	Arts, entertainment, and recreation	21
72	Accommodation and food services	22
44;45	Retail trade	22,5%
23	Construction	23
56	Administrative/support & waste management/remediation Service	24
61	Educational services	24
51	Information	25
48;49	Transportation and warehousing	25%
52	Finance and insurance	28
53	Real estate and rental and leasing	29

Tabulka 28 - Zkratky v NAICS a jejich význam

Tato proměnná bude jistě ponechána v modelu. Nicméně ji je nutno mírně upravit a to tak, že je nutné zachovat pouze dvě první číslice. Abychom vymazali posledních pět číslic a dostali tak pouze dvě číslice použijeme package jménem *stringr*.

```
library(stringr)

data2$NAICS =str_sub(data2$NAICS, end=-5)
```

Dále budeme chtít zobrazit tabulku četností jednotlivých průmyslů, v kterých uchazeči o úvěr plánují podnikat.

```
xxx= select(data2, NAICS)

xxx$n = replicate(5381,1)

xxx2= aggregate(xxx[,"n"], by=list(xxx$NAICS), FUN =sum, na.rm= TRUE)
```

Z těchto poznatků pak víme, že žadatelé o garantovaný úvěr nejvíce chtějí podnikat v průmyslu č. 44 – Retail trade (maloobchodní služby). Druhý nejpopulárnější sektor je s číslem 72 – ubytování a jídlo. Nejméně populární je 22 a 55.

Approval Date

Proměnná, která značí přesné datum odsouhlasení úvěru. Tato proměnná je pro všechna data stejná, ustanovil jsem ji na začátku praktické části jako rok 2006.

ApprovalFY

Proměnná, která značí hospodářský rok, do kterého spadá daný úvěr – může sem zapadnout i několik úvěrů, které jsou odsouhlaseny na konci předchozího roku. Toto je plně v kompetenci jednotlivých bank, které si samy určují, od jakého měsíce začíná hospodářský rok.

Term

Další zajímává proměnná, která značí, na jak dlouho byl úvěr sjednán (v měsících).

```
Summary(data2$Term)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   57.00   84.00   85.05   84.00  312.00
```

Průměrná doba, na kterou je úvěr sjednán je 85 měsíců a medián je 84 měsíců. Zároveň je z horního výpisu vidět, že proměnná obsahuje nějaké nulové hodnoty, to samozřejmě nedáv á žádný smysl, tyto hodnoty musí být odstraněny.

```
data=data[!(data$Term ==0),]
```

NoEmp

Proměnná, která značí, počet zaměstnanců daného podniku.

```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.

   0.0       1.0         3.0       7.441    6.0         4000
```

V této proměnné se budu muset vypořádat se dvěma odlehlými pozorování[20]. Jeden je vidět z horní tabulky, podnik, který má 4000 zaměstanců a druhý je podnik, který má 300 zaměstanců.

Pouze 19 podniků má více jak 100 zaměstancnů. A více než 20 zaměstanců má pouze 330 podniků z celého vzorku – 6,1%. Proto si zvolím libovolně číslo – v mém případě 150 – a podniky, které mají více než 150 zaměstnanců vyřadím ze vzorku. Jedná se celkem o 3 firmy.

```
data2=data2[!(data2$NoEmp>150),]
```

Velikost firmy jistě bude mít na vliv na schopnost splácet.

NewExist

Je proměnná, která označuje, zda žadatel již má založenou firmu nebo ji plánuje založit s pomocí úvěru.

nový/st.	Počet
0	3
1	3557
2	1821

Tabulka 29- rozdělení NewExist

Číslice 0 v tomto případě znamená, že kolonka pravděpodobně nebyla vyplněna, proto se bude muset odstranit.

```
data2=data2[!(data2$NewExist ==0),]
```

[20] Tzv. outliers

Zároveň bude nutné tuto proměnnou překódovat na jiné úrovně. Kategorie s číslem 2 značí nový podnik, kategorie s číslem 1 značí již existující podnik. Je nutné tuto kategorie přeměnit na číslo 1 pro nový podnik a číslo 0 pro již existující.

```
data2$NewExist<-as.factor(ifelse(data2$NewExist ==2,'1', '0'))
```

a dostaneme novou tabulku, která už neobsahuje 3 neznámá pozorování.

0	3554
1	1821

Tabulka 30- rozdělení NewExist 2

CreateJob

Tato proměnná říká, kolik bylo vytvořeno pracovních míst. Na první pohled zbytečná proměnná.

```
data2 =data2[,-5]
```

RetainedJob

Stejně jako předchozí proměnná tato proměnná nemá žádnou vypovídací schopnost a do modelu mi nebude vstupovat.

```
data2 =data2[,-5]
```

FranchiseCode

Znamená, jestli je daný podnik franšízou[21] nebo nikoliv. 0 znamená, že franšízou podnik není. 1 pak znamená, že franšízou je. Problém u této proměnné nastal, že je zde několik desítek vyplněných polí s jinou hodnotou. Je patrné, že lidé při vyplňování informací zadávali nějaký jiný kód. Proto se tato proměnná upraví tak, že pokud bude hodnota vyšší než 1, dostanu zpět hodnotu 1.

```
data2$FranchiseCode<-as.factor(ifelse(data2$FranchiseCode >=1,'1', '0'))
```

Moje data teď obsahují celkem 2495 podniků, které nejsou franšízou, a 2880 podniků, které franšízou jsou.

UrbanRural

O této proměnné jsem se již zmínil v souvislosti s vyřazeným ZIPcode. Proměnná nám říká, jestli je žadatel z venkova nebo z města.

Proměnná má celkem 3 úrovně. 0 je opět chybějící vyplnění, 1 značí město a 2 značí venkov. Většina žadatelů je z města (85%). Proto nová úroveň 1 bude značit žadatele z města a 0 bude značit žadatele z venkova.

0	21
1	4611
2	743

Tabulka 31- rozdělení UrbanRural

[21] Např. Starbucks, KFC

Chybějící pozorování opět vymažu.

```
data2=data2[!(data2$UrbanRural ==0),]
```

RevLineC

Tato proměnná říká, jestli má podnik zřízen *revolving line of credit*. Pokud ano pak je proměnná Y, pokud ne proměnná je N. Opět musíme pouze zaměnit písmena za číslice Y=1

```
data2$RevLineCr<-as.factor(ifelse(data2$RevLineCr =='Y','1', '0'))
```

Výsledkem je, že celkem 2228 žadatelů o úvěr si žádá o tuto službu. 3126 žadatelů se pak obejde bez.

LowDoc

Je způsob, jak si zažádat o úvěr pouze krátkým formulářem. Tento způsob měl ulevit pracovník administrativy. Nicméně tento způsob se v mnou sledovaném období již nepoužívá. Proto celou proměnnou vymažu.

ChgOffDate

Jedná se o proměnnou, která zaznamenáva, kdy došlo k defaultu. Pokud k defaultu nedošlo, pole je prázdné – toto se musí opravit a nahradit chybějící hodnotu 0.

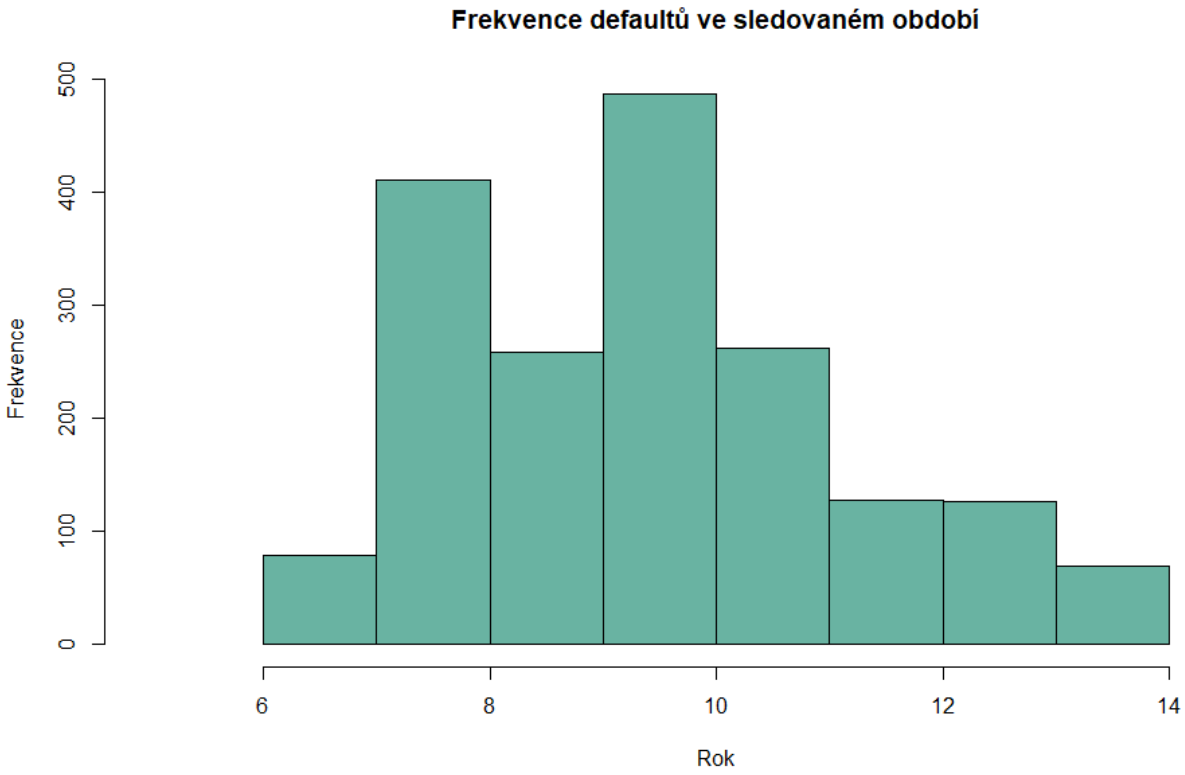
```
count=0
for (i in data$ChgOffDate) {
  if(i == '')  count = count+1
}
```

Výše uvedené podmínka mi vyšetří, kolik je nulových řádek. Je jich celkem 3526 a nenulových je 1819.

Prázdné řádky musíme nahradit nulou.

```
data$ChgOffDate <- sub("", "0", data$ChgOffDate)
```

Samotná proměnná nebude mít samozřejmě žádnou vypovídající hodnotu. Nikdy nevíme dopředu, kdy k defaultu dojde. Nicméně bude zajímavé spočítat počet defaultů v jednotlivých letech



Obrázek 25 - frekvence defaultů v jednotlivých letech

Na histogramu je vidět podezřelý trend, kdy v rokáczh07, 08, 09, 10, 11 je pozorována zvýšena frekvence defaultů – toto je samozřejmě způsobeno tím, že se světová ekonomika v těchto letech potýkala s ekonomickou krizí. Pro úplnost přikládám tabulku s počtem defaultů v jednotlivých rocích.

rok	# defaultů
6	5
7	74
8	411
9	258
10	487
11	262
12	127
13	126
14	69

Tabulka 32- počet defaultů v jednotlivých letech

DisbursementDate

Je datum, kdy přišly z banky peníze na účet klientům. Jsem si jist, že tato proměnná nijak nebude ovlivňovat šance na splacení úvěru, proto ji vyřadím.

Disbursement Gross

Je částká, která přišla klientovi na účet. Tato částka je v datech zapsána v dolarech. Program R s tímto neumí pracovat, proto musíme číslo $10,000.00 převést na 10000.

```
data$DisbursementGross <- as.numeric(gsub('[$,]', '', data$DisbursementGross))
```

Nadále s proměnnou můžeme pracovat jako s číselným vektorem.

```
   Min. 1st Qu.  Median   Mean   3rd Qu.      Max.
    195   25000   66373  177508  168900     2688052
```

Proměnná RevLine může tuto proměnnou značně navýšit. Tohoto si lze všimnout, když tuto proměnnou porovnáme s proměnnou GrAppr (částka schválená bankou). Zjistil jsem, že u 3114 klientů (58%) se tato částka rovná s částkou schválenou bankou.

Zároveň by tato částka nikdy neměla být nižší než částka schválená bankou. Nicméně v datech se tato situace objevuje u 200 případů. To mi přijde jako moc velká část celých dat, proto pokud se stane, že Disbursement bude menší než schválený úvěr, tuto částku přepíšu na výši schváleného úvěru. Myšlenka je taková, že to, co klientovi schválila banka, je minimum, co na účet může dostat.

Balance Gross

Je částka, kterou klient stále dluží bance – v mém případě by měla být na každém řádku rovna nule – klienti, kteří mají jiný stav než konečný stav (default/paid off) nás nezajímají. Proto tuto proměnou odstraním.

MIS_Status

Jedná se o jedinou vysvětlovanou proměnnou. Má dvě úrovně 0, která značí splacení úvěru a 1, která značí nesplacení úvěru a splňuje všechny definice *defaultu.*

ChgOffPrinGr

Proměnná, která značí částku, kterou dlužil klient bance v době defaultu. Tato proměnná nebude nijak ovlivňovat pravděpodobnost defaultu.

Nicméně proměnná je důležitá pro kontrolu předchozí proměnné, kterou je označení, zda k defaultu došlo či nikoliv. Je jasné, že pokud klientovi byla odepsána nějaká částka, musel klient dosáhnout defaultu.

Pokud ovšem spočítám, kolik nenulových položek obsahuje tato proměnná, dostanu se na číslo 1819. Toto číslo je zcela stejné jako číslo nenulových položek u proměnné ChgOffDate. Pokud spočítám, kolik nenulových položek obsahuje MIS_Status, dostanu číslo 1749. Je jasné, že je něco špatně. Tato proměnná mi vykazuje o 70 více defaultů než vysvětlovaná proměnná.

Proto je jasné, že proměnná MIS_Status obsahuje chyby. Všude, kde je nenulová částka ChgOffPrinGr, musí být nenulová úroveň defaultu.

```
for (i in c(1:length(a$a))) {

  if(a$a[i] >0 & a$b[i] == 0 )  a$b[i] =1

}
```

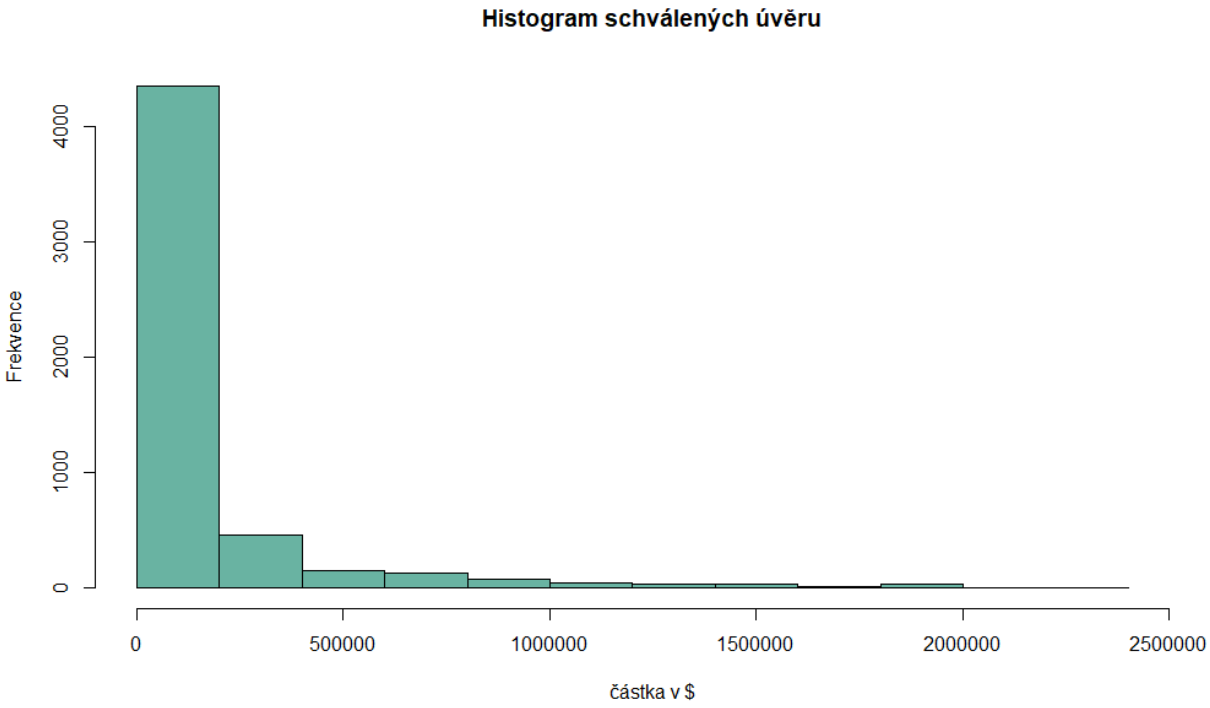
Počet defaultů se mi zvýšil o 70 na celkové číslo 1819.

GrAppr

Vyjadřuje částku, která byla schválena bankou. Ve většině případů by se tato částka měla zároveň rovnat částce, kterou banka připsala na účet klienta. Proměnná RevLine ale umožňuje, aby byla celková částka připsaná na účet klienta větší než částka, která byla schválená.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5000   25000   40000  156104  140000 2369000
```

Nejčastěji schvalovaná částka je 40 000$ s průměrem 156 104 dollarů. Nejmenší možný úvěr je evidentně 5000 $. Ještě se podívám na histogram.



Obrázek 26 - rozdělení výše schválených úvěrů

SBA_Appv

Poslední proměnná v původních datech vyjadřuje absolutní výši garance SBA. Pokud klient zdefaultuje, tak vládní organizace garantuje. Tato částka je vždy menší nebo stejná než předchozí proměnná GrAppr.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2500   12500   22750  118662  100000 2369000
```

A1

Toto bude nová proměnná, kterou si vytvořím z proměnné SBA_Appv a GrAppr. Bude mě zajímat jaká část schváleného dluhu je pokryta garancí. Hodnoty v nové proměnné mají celkem 34 úrovní. Nejobsáhlejší je 50% krytí dluhu, které obsahuje 3176 úvěrů. Druhé nejobsáhlejší je 85% krytí, které obsahuje 1069 úvěrů. Nicméně tato proměnná je mírně problematická kvůli tomu, že obsahuje hodně hodnot, které nezapadají nikam do větších úrovní. Zobrazeno na spodním výřezu obrazovky.

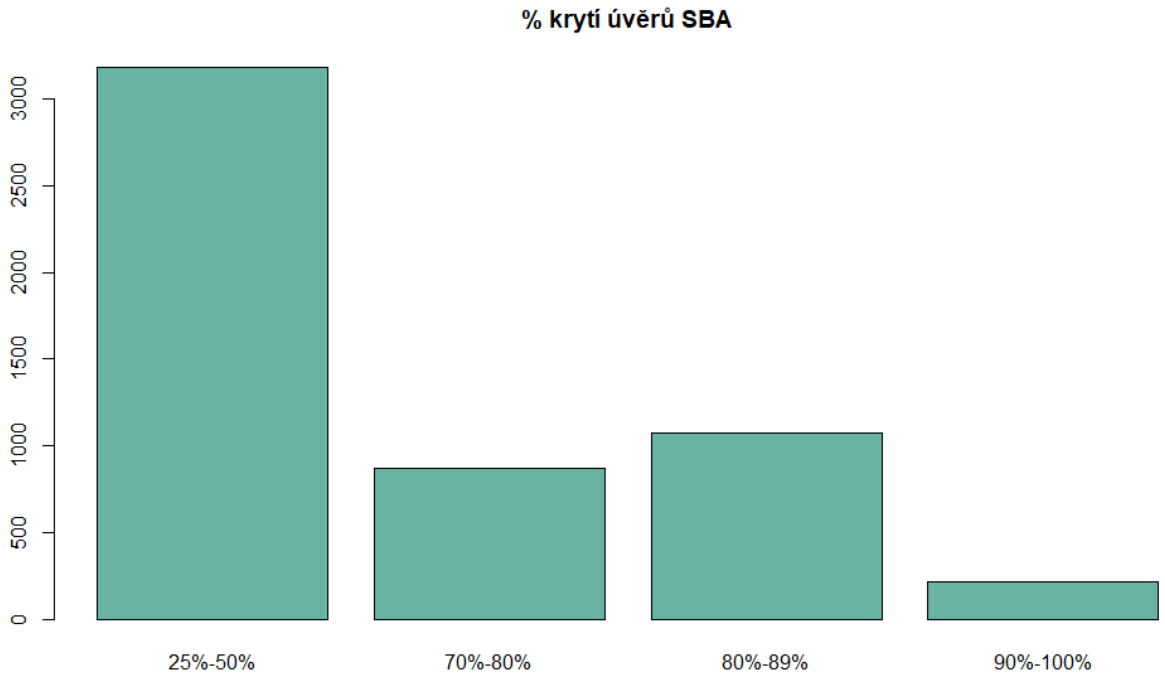
0.8500000	1069
0.8500002	1
0.8500034	1
0.8500045	1
0.8500058	1
0.8500067	1
0.8999998	1
0.9000000	7
1.0000000	212

Obrázek 27 - chyby v A1

Zde je vidět, že mezi úrovní s 85% krytí a 100% krytí je velké množství „vyjímek“. S těmito hodnotami se budu muset vypořádat. Jedna možnost je tyto hodnoty vymazat, druhá možnost je tyto hodnoty sloučit k nějaké nejbližší možné. Hodnotu 0.8500002 bych „zaokrouhlil“ k hodnotě 0.85. Toto jsem zajistil následující podmínkou.

```
data$A1<-
as.factor(ifelse(data$A1<=0.5,'0.5',ifelse(data$A1<=0.76,'0.75',ifelse(data$A1<=0.86,'0.85',ifelse(data
$A1<=0.91,'0.9','1'))))))
```

Horní řádka mi z původních 34 kategorií udělá pouze čtyři.



Obrázek 28 - % krytí pomocí SBA garance

Zároveň tato nově vytvořená proměnná mi nahradí ty proměnnou GrAppr a SBA_Appv.

```
data =data[,-10]
```

```
data =data[,-10]
```

A2

Další mnou vytvořená proměnná bude ta, že pokud jsou úvěry vypsány na více než 20 let, tak je nutné tento úvěr krýt nějakou nemovitostí. Proto proměnnou Term nahradím proměnnou, která značí, zda je úvěr zajištěn nemovitostí. 1 značí zajištění, 0 pak značí nezajištěný úvěr.

```
data$A2= replicate(5345, 1)

for (i in c(1:length(data$A2))) {

  if(data$Term[i] < 240 )  data$A2[i] = 0

}
```

Z mého vzorku dat je celkem 4940 úvěrů nezajištěných nemovitostí. Zbytek 405 úvěrů je zajištěných.

Jako závěrečnou věc přeházím proměnné v modelu, tak, aby první byla vysvětlovaná proměnná, následovat budou všechny spojité proměnné a po nich budou všechny diskrétní proměnné.

```
data=data[,c(9, 1, 2,3, 8, 4, 5, 6 ,7 ,10, 11)]
```

Příloha 2 – vytvořené funkce + kód použitý v kapitole Univariate Analysis

V této příloze naleznete nejdůležitější části kódu a krátký komentář.

Tato funkce vytvoří tabulku s WoE, IV, počtem špatných a počtem dobrých úvěrů. Je v ní i statistika zvaná efficiency a grp_score, které jsem v mé práci nepoužil.

```
gbpct <- function(x, y=data$MIS_Status){

  mt <- as.matrix(table(as.factor(x), as.factor(y))) # x -> independent variable(vector), y->dependent
variable(vector)

  Total <- mt[,1] + mt[,2]

  Total_Pct <- round(Total/sum(mt)*100, 2)

  Bad_pct <- round((mt[,1]/sum(mt[,1]))*100, 2)

  Good_pct <- round((mt[,2]/sum(mt[,2]))*100, 2)

  Bad_Rate <- round((mt[,1]/(mt[,1]+mt[,2]))*100, 2)

  grp_score <- round((Good_pct/(Good_pct + Bad_pct))*10, 2)

  WOE <- round(log(Good_pct/Bad_pct)*10, 2)

  g_b_comp <- ifelse(mt[,1] == mt[,2], 0, 1)

  IV <- ifelse(g_b_comp == 0, 0, (Good_pct - Bad_pct)*(WOE/10))

  Efficiency <- abs(Good_pct - Bad_pct)/2

  otb<-as.data.frame(cbind(mt, Good_pct,  Bad_pct,  Total,

                           Total_Pct,  Bad_Rate,

                           WOE, IV ))

  otb$Names <- rownames(otb)

  rownames(otb) <- NULL

  otb[,c(10,1,2,4,3,5:9)]

}
```

Pro faktorizaci spojitých proměnných jsem použil jednoduchou ifelse podmínku:

```
data$Term <-as.factor(ifelse(data$Term<=63,'1-63',

                           ifelse(data$Term<=84,'64-84', '85-312')))
```

následně jsem si zobrazil tabulku pomocí mnou vytvořené funkce:

```
A9 <- gbpct(data$Term)
```

Pro zobrazení IV grafu a distribuci good-bad úvěrů v jednotlivých proměnných jsem použil:

```
plot(as.factor(data$Term), as.factor(data$MIS_Status),

    main="Term ~ Good-Bad",
```

```
    xlab="kategorie",
    ylab="")

barplot(A9$IV, col="brown", names.arg=c(A9$Names),
       main="IV v Term",
       xlab="kategorie",
       ylab="IV")
```

A následně jsem tuto proměnnou dal do modelu:

```
m1_4 <- glm(MIS_Status~ Term,data=data,family=binomial())
summary(m1_4)
```

Pro ROC jsem použil:

```
plot(m3_perf, lwd=2, colorize=TRUE, main="AUC m3")
lines(x=c(0, 1), y=c(0, 1), col="red", lwd=1, lty=3);
```

Pro graf sensitivity a specificity:

```
m3_perf_precision <- performance(m3_pred, measure = "spec")
m3_perf_acc <- performance(m3_pred, measure = "sens")

plot(m3_perf_precision, col= 'blue', main="m3:specificity vs sensitivity ", ylab =
'Specificity/sensitivity')
plot(m3_perf_acc, add= TRUE, col= 'green')
```

Pro výpočet optimálního bodu cut-off:

```
opt.cut = function(m3_perf, m3_pred){
  cut.ind = mapply(FUN=function(x, y, p){
    d = (x - 0)^2 + (y-1)^2
    ind = which(d == min(d))
    c(sensitivity = y[[ind]], specificity = 1-x[[ind]],
      cutoff = p[[ind]])
  }, m3_perf@x.values, m3_perf@y.values, m3_pred@cutoffs)
}
print(opt.cut(m3_perf, m3_pred))
```

A Pro výpočet KS, Gini a AUC:

```
m3_KS <- round(max(attr(m3_perf,'y.values')[[1]]-attr(m3_perf,'x.values')[[1]])*100, 2)
m3_AUROC <- round(performance(m3_pred, measure = "auc")@y.values[[1]]*100, 2)
m3_Gini <- (2*m3_AUROC - 100)
```

Příloha 3 – Státy USA a defaultní míry z úvěrů SBA z roku 2006

ME	79	408	0,193627	G1
MT	48	328	0,146341	G1
ND	35	267	0,131086	G1
NE	74	371	0,199461	G1
NM	58	305	0,190164	G1
VT	45	296	0,152027	G1
WY	9	95	0,094737	G1

AK	20	98	0,204082	G2
CT	281	974	0,288501	G2
HI	67	252	0,265873	G2
IA	135	510	0,264706	G2
ID	170	688	0,247093	G2
KS	150	573	0,26178	G2
KY	190	643	0,29549	G2
LA	147	571	0,257443	G2
MA	593	2034	0,291544	G2
MN	411	1704	0,241197	G2
MO	388	1299	0,298691	G2
NH	182	751	0,242344	G2
OH	919	3104	0,29607	G2
OK	173	605	0,28595	G2
PA	819	2759	0,296847	G2
SD	39	156	0,25	G2
UT	428	1559	0,274535	G2
WI	323	1453	0,222299	G2

MS	165	462	0,357143	G3
NC	522	1306	0,399694	G3
NY	2112	6123	0,344929	G3
OR	272	867	0,313725	G3
RI	185	592	0,3125	G3
TX	1744	5064	0,344392	G3
VA	481	1227	0,392013	G3
WA	572	1826	0,313253	G3
WV	60	187	0,320856	G3
MD	446	1205	0,370124	G3
IN	500	1632	0,306373	G3
DE	78	224	0,348214	G3
AR	119	373	0,319035	G3
CO	622	1762	0,353008	G3

FL	2188	4366	0,501145	G4
AL	242	523	0,462715	G4
AZ	672	1490	0,451007	G4
CA	4206	9873	0,42601	G4
GA	1013	2026	0,5	G4
IL	1339	3185	0,420408	G4
MI	1188	2669	0,445111	G4
NJ	921	2241	0,410977	G4
NV	348	734	0,474114	G4
SC	194	438	0,442922	G4
TN	396	919	0,430903	G4

Seznam obrázků a grafů

Obrázek 1 - Lineární regrese 9
Obrázek 2 - Binární proměnná 10
Obrázek 3- Binární proměnná a lineární vztah 11
Obátek 4 - Logistická křivka 13
Obrázek 5 - optimální cut-off bod, zdroj: (Hosmer, 2000) 24
Obrázek 6 - ROC curve (Witzany, 2017) 25
Obrázek 7 - cumulative accuracy profile, zdroj: (Witzany, 2017) 26
Obrázek 8 - heat map - počet defaultů 29
Obrázek 9- Heat map - míra defaultu v USA 30
Obrázek 10 - Univariate analysis proměnné NAICS 34
Obrázek 11- Univariate analysis proměnné Term 36
Obrázek 12- Univariate analysis proměnné NoEmp 37
Obrázek 13- Univariate analysis proměnné DisbursementGross 39
Obrázek 14- Univariate analysis proměnné NewExist 41
Obrázek 15- Univariate analysis proměnné FranchiseCode 42
Obrázek 16- Univariate analysis proměnné UrbanRural 43
Obrázek 17- Univariate analysis proměnné RevLineCR 44
Obrázek 18- Univariate analysis proměnné A1 45
Obrázek 19- Univariate analysis proměnné A2 47
Obrázek 20- AUC závěrečného modelu fitovaného na populaci Texasu 50
Obrázek 21- optimal cut-off point finálního modelu fitovaného na populaci Texasu 51
Obrázek 22- rozložení good-bad úvěrů v upravené proměnné State 56
Obrázek 23- AUC finálního modelu fitovaného na celou populaci USA 57
Obrázek 24- Cut-off point finálního modelu fitovaného na celou populaci USA 58
Obrázek 25 - distribuce schálených úvěrů v jednotlivých měsících 63
Obrázek 26 - distribuce defaultů v jednotlivých měsících 64
Obrázek 27 - domácí banka vs mimostátní banka 68
Obrázek 28 - frekvence defaultů v jednotlivých letech 73
Obrázek 29 - rozdělení výše schválených úvěrů 75
Obrázek 30 - chyby v A1 76
Obrázek 31 - % krytí pomocí SBA garance 76

Knižní zdroje

HOSMER, David W. a Stanley LEMESHOW. Applied logistic regression. 2nd ed. New York: John Wiley, 2000. Wiley series in probability and statistics. ISBN 0-471-35632-8.

WITZANY, Jiří. Credit risk management and modeling. Ed. 1st. Praha: Oeconomica, 2010. Odborná kniha s vědeckou redakcí. ISBN 978-80-245-1682-0.

Hull, J. (2011). Options, Futures, and Other Derivatives (8th Edition). Upper Saddle River, N.J: Pearson/Prentice Hall. ISBN-13: 978-0133456318

Jonh C. Hull (2018). Risk Management and Financial Institutions (5th Edition). ISBN: 978-1-119-44811-2

ANDERSON, Raymond. The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. Oxford: Oxford University Press, 2007. ISBN 0199226407.

CRISTIANINI, Nello a John SHAWE-TAYLOR. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press, 2000. ISBN 0-521-78019-5.

SIDDIQI, Naeem. Intelligent credit scoring: building and implementing better credit risk scorecards. Second edition. Hoboken, New Jersey: Wiley, [2017].

Basel Committee on Banking Supervision International Convergence of Capital Measurements and Capital Standards A revised Framework June 2004

Logistic regression and its application in credit scoring. Christine Bolton. Cape Town, South Africa, 2009.
DOBSON, A.J. & Barnett, A.G., c2008. An introduction to generalized linear models 3rd ed., Boca Raton: CRC Press.

Internetové zdroje

Cartwheel technologies. (2017, December 28). *Credit scoring*. dostupné z https://rpubs.com/Cartwheel/creditscoring

Akul Mahajan. (2018, March 3). *Logistic regression and CART in R*. dostupné z http://rstudio-pubs-static.s3.amazonaws.com/366219_fd424698e4e0460e8254d5ae055e8055.html

Ali Sultan. (2016, April 25). *Accuracy vs area under the roc curve*. dostupné z https://stats.stackexchange.com/questions/225210/accuracy-vs-area-under-the-roc-curve

R. Berwick. (2018, October 26). *An idiot's guide to SVM*. dostupné z http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf

Devin Soni. (2018, June 8). *An intorduction to bayesian networks*. dostupné z https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e

Ayhan Dis. (2016, January 5). *Introduction to creditR*. dostupné z https://www.analyticsvidhya.com/blog/2019/03/introduction-creditr-r-package-enhance-credit-risk-scoring-validation-r-codes/

Dr. Marcel Dettling. (2016, November 17). Statistical Analysis of Financial Data. dostupné z https://www.ethz.ch/content/dam/ethz/special-

interest/math/statistics/sfs/Education/Advanced%20Studies%20in%20Applied%20Statistics/course-material/finance/00_Notes_v00.pdf

Petr Gurný, Martin Gurný. (2013, September 14). Comparison of credit scoring models on PD estimation for US banks. dostupné z https://ideas.repec.org/a/prg/jnlpep/v2013y2013i2id446p163-181.html

ZLÁMAL, Filip. Logistická regrese v R. Brno, 2013. Bakaláøská práce. Masarykova univerzita.

Matoušková, Zuzana. Model logistické regrese s fixními a smíšenými efekty v hodnocení kreditního rizika v Brno, 2016. Diplomová práce. Masarykova univerzita.

David Nikodem. Možnosti odhadu rizikových parametrů kreditního rizika v Praze, 2018. VŠE.