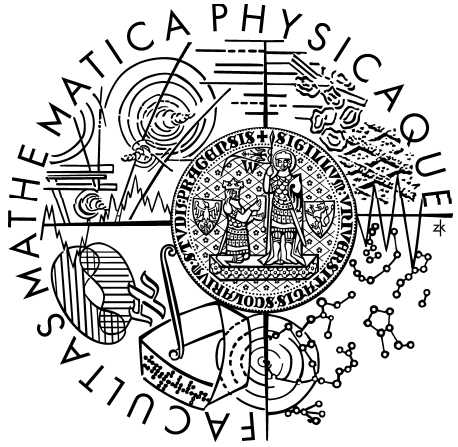
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Michal Rychnovský

Postupná výstavba modelů ohodnocení kreditního rizika

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Pavel Charamza, CSc.
Studijní program: Matematika, Obecná matematika

2008

Chtěl bych poděkovat vedoucímu RNDr. Pavlu Charamzovi, CSc. za poskytnuté materiály a celkovou pomoc při tvorbě práce.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 29. května 2008 Michal Rychnovský

Obsah

Název práce: Postupná výstavba modelů ohodnocení kreditního rizika
Autor: Michal Rychnovský
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: RNDr. Pavel Charamza, CSc.
E-mail vedoucího: pavel.charamza@mediaresearch.cz
Abstrakt: Cílem této práce je přiblížit podstatu výstavby skóringových modelů. Popisujeme zde metodu logistické regrese, odhadování jejích parametrů a testování jejich významnosti. Na základě proměnných odds ratio potom zavádíme independence model jako odhad podmíněné šance splacení klienta. Tento model dále zobecňujeme přidáváním vah jednotlivým skupinám a kategoriím charakteristik klienta. Takto přicházíme k WOE modelu a plnému logistickému modelu. Věnujeme se také měření diverzifikační schopnosti modelů pomocí Lorenzovy křivky a Somerovy d statistiky jako odhadu Giniho koeficientu. Nakonec aplikujeme popsané metody na praktickou výstavbu skóringových modelů a na reálných datech porovnáme vhodnost a diverzifikační schopnost představovaných modelů. Součástí práce je také výstup na internetovou encyklopedii Wikipedia.
Klíčová slova: kreditní riziko, skóringové modely, logistická regrese.

Title: Step by step credit risk model construction
Author: Michal Rychnovský
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Pavel Charamza, CSc.
Supervisor's e-mail address: pavel.charamza@mediaresearch.cz
Abstract: The aim of the present work is to outline a principle of scoring models construction. We describe the logistic regression method, its parameters estimation and their significance testing. On the ground of odds ratio variables we define the Independence model as an estimate of the conditional odds of client's ability to pay. We generalize this model by adding individual weights to groups and categories of clients characteristic. Using this way we come to the WOE model and Full logistic model. We also study the way of measuring the diversification power of the models by the Lorenz curve and Somer's d statistics as an estimate of the Gini coefficient. Finally we apply the described methods to the practical scoring model construction. On a real data we compare suitability and diversification power of the introduced models. Part of this work is also an output for the internet encyclopedia Wikipedia.
Keywords: credit risk, scoring models, logistic regression.

Kapitola 1

Úvod

V dnešní době existuje mnoho bankovních i nebankovních institucí, které poskytují úvěry klientům. Poskytnutím úvěru se taková instituce vystavuje *kreditnímu riziku*, tj. riziku že dotyčný klient úvěr za daných podmínek nesplatí a způsobí tím poskytovateli ztrátu. Proto při každé žádosti o úvěr potřebuje taková instituce toto riziko co nejlépe kvantifikovat a na základě dostupných informací o žadateli rozhodnout, jestli a za jakých podmínek úvěr poskytne. K tomuto v praxi slouží *skóringové modely*.

Skóringový model, jako model ohodnocení kreditního rizika, bývá založen na databázi existujících klientů, kterým kdy byl poskytnut úvěr, společně s informací, kterým z nich se podařilo úvěr splatit. Potom je každému dalšímu žadateli o úvěr na základě tohoto modelu přiděleno *skóre*, které reprezentuje jeho očekávanou schopnost splácet. Podle tohoto skóre se potom instituce rozhoduje, za jakých podmínek úvěr poskytne.

Cílem této práce je přiblížit postup výstavby některých často používaných skóringových modelů na základě databáze existujících klientů. Pro jednoduchost nazvěme *dobrým* takového klienta, který úvěr *splatil* včas a za smluvených podmínek, a *špatným* takového klienta, který některému ze svých závazků nedostál. Toto nazvěme *defaultem*.

Základním kamenem výstavby skóringových modelů je metoda *logistické regrese* popsaná v Kapitole 2. Tato část vychází zejména z publikace [5] Hosmer D. W., Lemeshow S., *Applied Logistic Regression*, str. 1–56 a částečně také z [1] Agresti A. *Categorical Data Analysis*, str. 79–129. Výstupem této

metody je odhad podmíněné pravděpodobnosti splacení klienta s danými charakteristikami.

Jednou z nejpodstatnějších zkoumaných vlastností skóringového modelu je *schopnost diverzifikace*, tedy míra rozlišení dobrých klientů od špatných. Tuto vlastnost v praxi znázorňujeme *Lorenzovou křivkou* a kvantifikujeme *Giniho koeficientem*. Tyto metody jsou popsány v Kapitole 3.

Stěžejní částí práce je potom přirozená výstavba trojice používaných skóringových modelů popsaná v Kapitole 4. Vycházíme zde z odhadu podmíněné *šance* splacení klienta, odkud se postupným zobecňováním dostáváme od nejjednoduššího *independence modelu*, založeného pouze na pozorovaných veličinách, ke komplexnějšímu *WOE modelu* a nakonec k *plnému logistickému modelu*, jejichž parametry odhadujeme metodou logistické regrese popsanou v Kapitole 2.

Poslední kapitola je věnována praktické aplikaci uvedených metod a modelů. Na reálných datech zde vytváříme popsané skóringové modely a porovnáváme jejich vhodnost použití a diverzifikační schopnost.

Součástí práce je též zavedení hesel *kreditní riziko*, *skóringový model*, *Lorenzova křivka* a *Giniho koeficient* do internetové encyklopedie Wikipedia.[1]

[1] http://wikipedia.cz, hesla byla založena 14.5.2008.

Kapitola 2

Logistická regrese

2.1 Model logistické regrese

Naším cílem je najít vyhovující model pro odhadování podmíněné pravděpodobnosti *splacení* klienta v závislosti na hodnotách vysvětlujících proměnných, tzv. *regresorů*. Obecně uvažujme pro i-tého klienta z databáze vektor různých vysvětlujících proměnných $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ik})$ a binární vysvětlovanou proměnnou $Y_{\boldsymbol{x}_i}$, kde $Y_{\boldsymbol{x}_i} = 1$ v případě splacení a $Y_{\boldsymbol{x}_i} = 0$ v případě nesplacení.

Střední hodnotu $Y_{\boldsymbol{x}}$ můžeme spočítat jako

$$\mathrm{E}(Y_{\boldsymbol{x}}) = 1 \cdot \mathrm{P}(Y_{\boldsymbol{x}} = 1) + 0 \cdot \mathrm{P}(Y_{\boldsymbol{x}} = 0) = \mathrm{P}(Y_{\boldsymbol{x}} = 1).$$

Označíme-li dále $\pi(\boldsymbol{x}) = \mathrm{P}(Y_{\boldsymbol{x}} = 1)$ podmíněnou pravděpodobnost splacení klienta s vektorem vysvětlujících proměnných $\boldsymbol{x}$, dostáváme

$$\mathrm{E}(Y_{\boldsymbol{x}}) = \pi(\boldsymbol{x}).$$

Chceme tedy podchytit závislost $\pi(\boldsymbol{x})$ na hodnotách vektoru $\boldsymbol{x}$. Prvním možným modelem, který by nás mohl napadnout, je *lineární regrese* s vektorem parametrů $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k)$,

$$\pi(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}.$$

Tento model však vyhovující není.[1] $Y_{\boldsymbol{x}}$ je binární proměnná, která nabývá

[1] Nazývá se lineární pravděpodobnostní model a v praxi se pro svou jednoduchost přesto někdy používá. Je s ním však spjato mnoho nevýhod (viz např. [6], str 169–171).

jen hodnot 0 a 1, a $\pi(\boldsymbol{x})$ je hodnota pravděpodobnosti z intervalu $[0, 1]$. Jenže proměnná vycházející z lineární regrese může obecně nabývat všech reálných hodnot. Proto definujeme funkci *odds*, nebo také *šance*, jako

$$\text{odds}(\boldsymbol{x}) = \frac{\text{P}(Y_{\boldsymbol{x}} = 1)}{\text{P}(Y_{\boldsymbol{x}} = 0)} = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}. \tag{2.1}$$

Tato funkce již nabývá hodnot v intervalu $[0, \infty)$. Abychom dostali hodnoty z celého $\mathbb{R}$, použijeme logaritmickou transformaci. Takto vytvořená funkce se nazývá *logit* a je definována

$$\text{logit}(\boldsymbol{x}) = \ln\big(\text{odds}(\boldsymbol{x})\big) = \ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right). \tag{2.2}$$

Položíme-li konečně $\text{logit}(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$, dostáváme tak specifický vztah pro logistickou regresi[2] ve tvaru

$$\pi(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}}}. \tag{2.3}$$

2.2 Odhad parametrů

Předpokládejme, že máme n nezávislých pozorování reprezentovaných vektory $(y_i, \boldsymbol{x}_i')$, $i = 1 \dots n$. Naším cílem je najít co nejlepší odhad parametrů modelu, tedy vektoru $\boldsymbol{\beta}$.

Pro odhad parametrů lineární regrese se běžně používá *metoda nejmenších čtverců* založená na minimalizaci součtu druhých mocnin odchylek odhadnutých hodnot od pozorovaných hodnot. Pro odhady modelu logistické regrese se používá obecnější *metoda maximální věrohodnosti* (viz např. [2] str. 146–162).

Metoda maximální věrohodnosti spočívá v konstrukci takzvané *věrohodnostní funkce*. Ta udává pravděpodobnost, s jakou při daném odhadovaném modelu nastanou právě všechny pozorované události (data). Vyhovuje ten model, pro který je tato pravděpodobnost maximální.

[2] Jiný způsob odvození tvaru logistické regrese použitím latentních proměnných uvádí například [6], str. 171–187.

Zkonstruujme tedy věrohodnostní funkci. Vyjdeme-li z označení pravděpodobnosti $\pi(\boldsymbol{x})$, můžeme hledanou podmíněnou pravděpodobnost vyjádřit jako

$$\mathrm{P}(Y_{\boldsymbol{x}_i} = y_i) = \pi(\boldsymbol{x}_i)^{y_i}\big(1 - \pi(\boldsymbol{x}_i)\big)^{1-y_i},$$

tedy pro $y_i = 1$ je to pravděpodobnost $\pi(\boldsymbol{x}_i)$ a pro $y_i = 0$ pravděpodobnost $1 - \pi(\boldsymbol{x}_i)$.

Protože pozorované hodnoty jsou podle předpokladu nezávislé, můžeme definovat věrohodnostní funkci $\mathrm{l}(\boldsymbol{\beta})$ jako součin podmíněných pravděpodobností pro jednotlivá pozorování

$$\mathrm{l}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}\big(1 - \pi(\boldsymbol{x}_i)\big)^{1-y_i}. \tag{2.4}$$

Abychom našli maximum této funkce, provedeme nejprve logaritmickou transformaci. Logaritmus polohu extrému neovlivní, ale výsledná funkce bude vhodnější pro derivaci. Takto dostáváme

$$\mathrm{L}(\boldsymbol{\beta}) = \ln\big(\mathrm{l}(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \Big(y_i \ln\big(\pi(\boldsymbol{x}_i)\big) + (1 - y_i)\ln\big(1 - \pi(\boldsymbol{x}_i)\big)\Big). \tag{2.5}$$

Abychom dostali hledané maximum vzhledem k vektoru parametrů $\boldsymbol{\beta}$, pohlížejme na funkci π jako na funkci proměnných $\boldsymbol{\beta}$ a $\boldsymbol{x}$ a položme jednotlivé parciální derivace funkce $\mathrm{L}(\boldsymbol{\beta})$ podle parametrů $\beta_0, \beta_1, \ldots, \beta_k$ rovny nule. Takto dostaneme soustavu takzvaných *věrohodnostních rovnic* tvaru

$$\sum_{i=1}^{n} \big(y_i - \pi(\boldsymbol{x}_i)\big) = 0 \tag{2.6}$$

a

$$\sum_{i=1}^{n} x_{ij}\big(y_i - \pi(\boldsymbol{x}_i)\big) = 0, \tag{2.7}$$

pro $j = 1, 2, \ldots, k$, kde x_{ij} je j-tá složka vektoru $\boldsymbol{x}_i$.

Tato nelineární soustava rovnic se zpravidla řeší numericky za pomocí specializovaného statistického software (např. SAS, SPSS, EViews a dalších). Řešením dostaneme vektor $\widehat{\boldsymbol{\beta}}$, *maximálně věrohodný odhad* vektoru

parametrů $\boldsymbol{\beta}$.

Z asymptotických vlastností maximálně věrohodných odhadů (viz např. [2] str. 146–162) odhadneme také směrodatné odchylky $\widehat{\sigma}(\widehat{\beta}_j)$ odhadnutých parametrů $\widehat{\beta}_j$. Tento odhad vychází z matice $\mathrm{I}(\boldsymbol{\beta})$, jejíž prvky tvoří hodnoty druhých parciálních derivací $L(\boldsymbol{\beta})$ podle $\boldsymbol{\beta}$ s opačným znaménkem

$$\mathrm{i}(\boldsymbol{\beta})_{jj} = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} = \sum_{i=1}^{n} x_{ij}^2 \pi(\boldsymbol{x}_i)\big(1 - \pi(\boldsymbol{x}_i)\big)$$

a

$$\mathrm{i}(\boldsymbol{\beta})_{jl} = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = \sum_{i=1}^{n} x_{ij} x_{il} \pi(\boldsymbol{x}_i)\big(1 - \pi(\boldsymbol{x}_i)\big).$$

Variační matici $\mathrm{var}(\boldsymbol{\beta})$ potom dostaneme jako inverzi matice $\mathrm{I}(\boldsymbol{\beta})$, tedy $\mathrm{var}(\boldsymbol{\beta}) = \mathrm{I}^{-1}(\boldsymbol{\beta})$. Odtud rozptyl $\mathrm{var}(\beta_j)$ j-té složky je j-tý diagonální prvek matice $\mathrm{var}(\boldsymbol{\beta})$. Nakonec dosazením $\widehat{\boldsymbol{\beta}}$ získáme asymptotický odhad rozptylu $\widehat{\mathrm{var}}(\widehat{\beta}_j)$ a tedy také směrodatné odchylky j-tého parametru

$$\widehat{\sigma}(\widehat{\beta}_j) = \sqrt{\widehat{\mathrm{var}}(\widehat{\beta}_j)}. \tag{2.8}$$

2.3 Testování hypotéz o parametrech modelu

Poté, co jsme získali maximálně věrohodný odhad vektoru parametrů $\widehat{\boldsymbol{\beta}}$, se zaměříme na statistickou významnost jednotlivých koeficientů i modelu jako celku. Nebudeme se nyní zabývat tím, jak dobře model vystihuje data (v absolutním smyslu), ale pouze relativně poměřovat, zda jednotlivé koeficienty statisticky významně přispívají k vypovídající schopnosti modelu či nikoliv.

První způsob testování statistické významnosti jednotlivých parametrů vychází z asymptotické normality odhadu $\widehat{\beta}_i$ (viz [2] str. 146–162), tedy

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}(\widehat{\beta}_i)} \sim N(0,1).$$

Odtud pro testování $\beta_i = 0$, používáme tzv. *Waldův test*, využívající poměru maximálně věrohodného odhadu $\widehat{\beta}_i$ a odhadu jeho směrodatné odchylky,

$$W = \frac{\widehat{\beta}_i}{\widehat{\sigma}(\widehat{\beta}_i)}. \tag{2.9}$$

Nulovou hypotézu, že $\beta_i = 0$ na hladině významnosti α, testujeme porovnáním hodnoty $|W|$ s kvantilem normálního rozdělení $z_{1-\frac{\alpha}{2}}$.

Na základě Waldova testu je možné zkonstruovat také *intervaly spolehlivosti* jednotlivých parametrů pro dané α

$$\beta_i \in \big(\widehat{\beta}_i - z_{1-\frac{\alpha}{2}}\widehat{\sigma}(\widehat{\beta}_i), \widehat{\beta}_i + z_{1-\frac{\alpha}{2}}\widehat{\sigma}(\widehat{\beta}_i)\big). \tag{2.10}$$

Nyní se podívejme na testování statistické významnosti více parametrů nebo kvality modelu jako celku. U modelu lineární regrese používáme pro porovnání vypovídací schopnosti modelu takzvaný *reziduální součet čtverců*

$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

Vlastní test významnosti skupiny koeficientů potom provedeme tak, že porovnáme hodnoty RSS původního modelu s omezeným modelem, ve kterém dané parametry vypustíme. Podobné kritérium zvolíme i u modelu logistické regrese.

U modelu logistické regrese je toto kritérium založeno na logaritmické věrohodnostní funkci. Obecně definujme takzvaný *saturovaný model*, jako model s takovým počtem parametrů, že s pravděpodobností jedna vystihuje pozorovaná data, a $\mathrm{l}_S(\boldsymbol{\beta})$ příslušnou věrohodnostní funkci. Dále definujeme D (z angl. *deviance*) jako

$$D = -2\big(\,\mathrm{L}(\widehat{\boldsymbol{\beta}}) - \mathrm{L}_S(\boldsymbol{\beta}_S)\big) = -2\ln\left(\frac{\mathrm{l}(\widehat{\boldsymbol{\beta}})}{\mathrm{l}_S(\boldsymbol{\beta}_S)}\right). \tag{2.11}$$

Poměru v poslední závorce říkáme *věrohodnostní poměr*. Transformace $-2\ln$ jej upravuje, aby měl známé rozdělení použitelné pro testování hypotéz, které popíšeme dále. Jelikož v našem případě je věrohodnostní funkce saturovaného modelu rovna jedné, případ se nám zjednoduší na tvar, který můžeme dále upravit takto

$$D = -2\ln\big(\mathrm{l}(\widehat{\boldsymbol{\beta}})\big) = -2\ln\Big(\prod_{i=1}^{n}\widehat{\pi}(\boldsymbol{x}_i)^{y_i}\big(1-\widehat{\pi}(\boldsymbol{x}_i)\big)^{1-y_i}\Big),$$

tedy

$$D = -2\sum_{i=1}^{n}\Big(y_i\ln\big(\widehat{\pi}(\boldsymbol{x}_i)\big) + (1-y_i)\ln\big(1-\widehat{\pi}(\boldsymbol{x}_i)\big)\Big). \tag{2.12}$$

Chceme-li nyní zjistit statistickou významnost některých l proměnných modelu, porovnáme hodnoty D původního neomezeného (angl. unrestricted) a omezeného (angl. restricted) modelu, tj. modelu, ve kterém položíme daných l parametrů rovných nule. Definujeme charakteristiku G jako

$$G = D_R - D_U = -2\ln\left(\frac{\mathrm{l}_R(\widehat{\boldsymbol{\beta}}_R)}{\mathrm{l}_U(\widehat{\boldsymbol{\beta}}_U)}\right). \tag{2.13}$$

Tato charakteristika má v logistické regresi podobný význam jako F v regresi lineární. Za platnosti nulové hypotézy, že daných l parametrů modelu se statisticky významně neliší od nuly, se charakteristika G řídí rozdělením χ^2 o l stupních volnosti. Proto, je-li G větší než kvantil $\chi^2_{1-\alpha}(l)$, nulovou hypotézu zamítáme a jeden nebo vice z testovaných parametrů je statisticky významný.

2.4 Vysvětlující proměnné

Abychom mohli vystavět kvalitní model logistické regrese, potřebujeme k tomu příslušnou datovou sadu takzvaných *vysvětlujících proměnných*. Tyto proměnné mohou být buď *kvantitativního* nebo *kvalitativního* charakteru. Vysvětleme si tyto pojmy a uveďme několik příkladů na možných charakteristikách klientů.

Kvantitativní proměnné jsou číselné proměnné vyjadřující počet, množství, velikost míru atp. Dle charakteru je dále dělíme na *diskrétní* a *spojité*. Příkladem diskrétní veličiny by mohl být počet dětí klienta, naopak za spojitou proměnnou bychom mohli považovat například měsíční příjem.

Kvalitativní proměnné označují většinou kategorii, ve které se subjekt nachází. Tyto dále dělíme na *ordinální*, u kterých můžeme kategorie logicky

kvalitativně uspořádat, a *nominální*, které uspořádatelné nejsou. Takovou ordinální proměnnou by mohlo být třeba nejvyšší dosažené vzdělání, které lze uspořádat. Naopak nominální veličinou by mohl být rodinný stav (svobodný, ženatý, rozvedený, vdovec), který kvalitativně uspořádat nelze.

Přestože kvantitativní proměnné mohou do modelu logistické regrese vstupovat přímo svojí hodnotou, většinou se v praxi přikláníme k jejich rozřazení do kategorií, např plat 5000–9999, 10000–14999 atd. Podobně postupujeme také u kvalitativních proměnných. Každá proměnná (používáme rovněž termín *skupina*) je tedy charakterizována sadou *znaků* (nebo též *kategorií*), kterých může nabývat. Ke každému znaku dané skupiny potom přiřadíme tzv. *dummy* proměnnou, tj. binární proměnnou, která má hodnotu 1, pokud prvek daného znaku nabývá, a hodnotu 0 v opačném případě. Pro lepší orientaci budeme skupiny indexovat horním indexem a příslušné kategorie dolním indexem. Tak například pro ženatého muže bychom potom měli $x_1^i = 0$ (není svobodný), $x_2^i = 1$ (je ženatý), $x_3^i = 0$ (není rozvedený) a $x_4^i = 0$ (není vdovec) jako čtveřici kategorií i-té skupiny.

Z takto získaných proměnných potom vytvoříme sloupcový vektor $\boldsymbol{x}' = (x_0, x_1, \ldots, x_k)$, kde položme $x_0 = 1$, abychom do modelu $\boldsymbol{\beta}'\boldsymbol{x}$ přirozeně dostali také úrovňovou konstantu.

K popisu znaků obecně nepotřebujeme všechny dummy proměnné. Pokud je totiž například $x_2^i = 0$ (není ženatý), $x_3^i = 0$ (není rozvedený) a $x_4^i = 0$ (není vdovec), je potom zřejmé, že je tento muž svobodný. Pro každou charakteristiku proto při odhadu vynecháváme jednu dummy proměnnou, aby nedocházelo k multikolinearitě. Testujeme-li potom statistickou významnost proměnných, testujeme obvykle nulovost sady příslušných dummy proměnných jako jednoho celku.

Příklad popisu proměnných nalezneme dále v Kapitole 4, konkrétní aplikaci na reálnou databázi potom v Kapitole 5.

Kapitola 3

Diverzifikační schopnost modelu

3.1 Diverzifikační schopnost modelu

Jednou z nejdůležitějších zkoumaných vlastností skóringového modelu je jeho *schopnost diverzifikace*, tedy míra oddělení dobrých klientů od špatných. V ideálním případě bychom totiž chtěli nalézt takový model, kde by existovala taková hodnota skóre s_0 (skóringová hranice), pro kterou by všichni špatní klienti v databázi byli ohodnoceni skóre nižším než s_0 a naopak všichni dobří klienti skóre větším než s_0. V takovém modelu bychom potom mohli podle dosaženého skóre poměrně dobře rozhodnout o tom, zda se klient zdá dobrý či nikoliv.

V praxi však zpravidla nenajdeme takovou skóringovou funkci, která by neomylně vystihovala kvalitu všech klientů v databázi. Budou se zde jistě vyskytovat takoví klienti, kteří mají sice nízké skóre, ale přesto se jim podařilo splatit, a naopak takoví, kteří přes své vysoké skóre nezaplatili. Skóringová funkce nám potom tedy dobré a špatné klienty rozdělí jen přibližně.

Pro názornost si představme, že jsou všichni klienti seřazeni vzestupně podle přiděleného skóre. V ideálním modelu bychom měli řadu samých špatných klientů a po překročení hranice s_0 řadu samých dobrých klientů. Oproti tomu v reálném modelu dostáváme řadu klientů, kde by sice na začátku byli častěji špatní klienti, ale mezi nimi by se vyskytovali i nějací dobří. Dobrých klientů by postupně přibývalo, až ke konci bychom měli řadu dobrých kli-

entů, mezi kterými by bylo i několik špatných.

A tedy podle toho, jak dobře uspořádání klientů podle skóre odděluje dobré klienty od špatných, posuzujeme kvalitu modelu z hlediska diverzifikační schopnosti.

3.2 Lorenzova křivka

Jedním z nejpoužívanějších způsobů grafického znázornění diverzifikace je *Lorenzova křivka*[1]. Konstrukce Lorenzovy křivky je založena na definici tzv. *distribučních funkcí skóre dobrých a špatných klientů*.

Označme $S = \{s(\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{X}\}$ obor hodnot skóringové funkce $s(\boldsymbol{x})$. Potom pro každou hodnotu skóre $s \in S$ definujme *distribuční funkci skóre dobrých klientů* $\mathrm{F}^G(s)$ jako pravděpodobnost, že náhodně vybraný dobrý klient bude mít skóre menší než s, a *distribuční funkci skóre špatných klientů* $\mathrm{F}^B(s)$ jako pravděpodobnost, že náhodně vybraný špatný klient bude mít skóre menší než s.

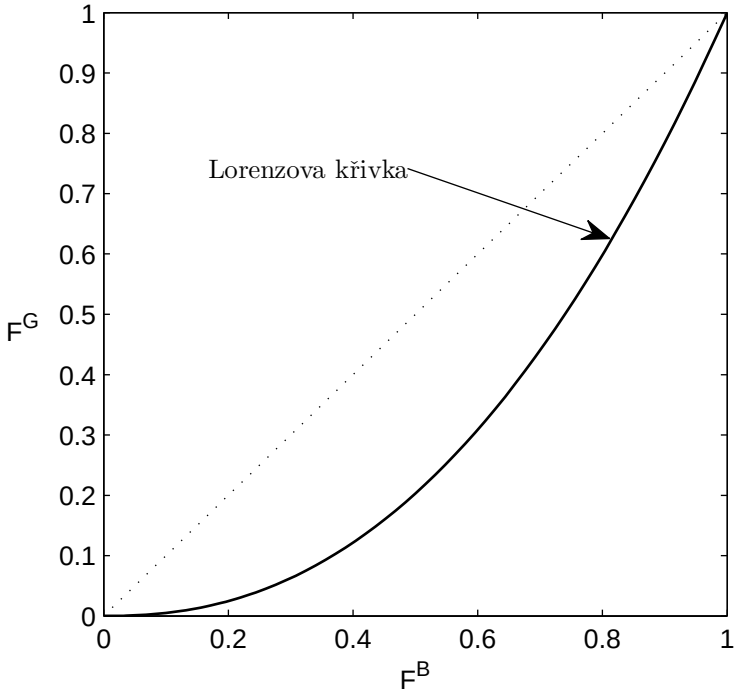
Explicitní distribuční funkce $\mathrm{F}^G(s)$ a $\mathrm{F}^B(s)$ v praxi zpravidla neznáme, proto je nejčastěji nahrazujeme konzistentními odhady. Funkci $\mathrm{F}^G(s)$ odhadujeme jako poměr počtu dobrých klientů se skóre menším než s ku počtu všech dobrých klientů a funkci $\mathrm{F}^B(s)$ jako poměr počtu špatných klientů se skóre menším než s ku počtu všech špatných klientů.

Nakonec definujeme Lorenzovu křivku jako množinu bodů

$$L = \Big\{ \big[\mathrm{F}^B(s), \mathrm{F}^G(s) \big] \in \mathbb{R}^2 : s \in S \Big\}, \tag{3.1}$$

kde $s \in S$ nabývá všech hodnot skóre použité skóringové funkce. Takto zkonstruovaná Lorenzova křivka potom leží uvnitř jednotkového čtverce a spojuje protilehlé vrcholy (Obrázek 3.1). Čím větší má náš model diverzifikační schopnost, tím více se Lorenzova křivka přibližuje stranám čtverce.

[1] V ekonomii se s Lorenzovou křivkou setkáváme především při znázorňování nerovnoměrnosti rozdělení důchodů či bohatství v populaci nějakého celku. My se však budeme zabývat jejím použitím pro hodnocení modelů kreditního rizika.



Obrázek 3.1: Lorenzova křivka

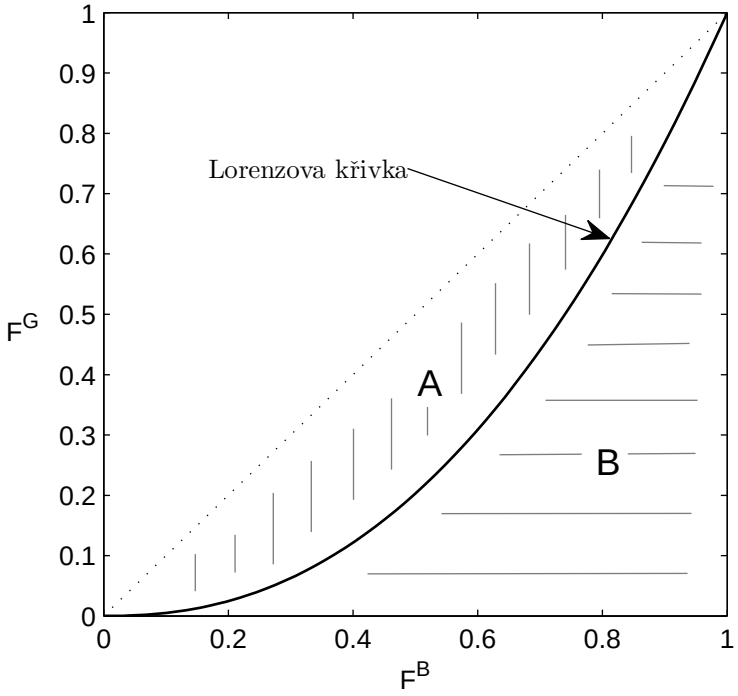
3.3 Giniho koeficient

Jako číselná charakteristika diverzifikační schopnosti modelu se nejčastěji používá tzv. *Giniho koeficient*[2]. Giniho koeficient většinou definujeme jako poměr orientované plochy mezi Lorenzovou křivkou a diagonálou jednotkového čtverce (A) ku celkové ploše pod diagonálou ($A + B$), tedy $GC = \frac{A}{A+B}$ (Obrázek 3.2).

Protože obsah plochy pod diagonálou je polovina jednotkového čtverce, můžeme definici přepsat jako $GC = 2A$ nebo také $GC = 1 - 2B$. Odtud použitím posledního jmenovaného výrazu dostáváme matematický vztah

$$GC = 1 - 2 \int_S \mathrm{F}^G(s)\, \mathrm{dF}^B(s), \tag{3.2}$$

[2]Giniho koeficient má opět veliké uplatnění v ekonomii, kde se jím poměřuje ekvivalence rozložení bohatství a důchodů v jednotlivých územních celcích, nejčastěji státech.



Obrázek 3.2: Giniho koeficient

Jiné vyjádření získáme, vyjdeme-li ze vztahu $GC = 2A$. Potom

$$GC = 2 \int_S \left(\mathrm{F}^B(s) - \mathrm{F}^G(s) \right) \mathrm{dF}^B(s). \tag{3.3}$$

Giniho koeficient je tedy dvojnásobek orientované plochy mezi Lorenzovou křivkou a diagonálou jednotkového čtverce, neboli ekvivalentně poměr této plochy a celkové plochy pod diagonálou. Hodnota Giniho koeficientu proto leží v intervalu $[-1, 1]$, kde hodnota 1 značí perfektní (ideální) diverzifikační schopnost, hodnota 0 značí nulovou diverzifikační schopnost a záporné hodnoty (křivka prohnutá nahoru) značí opačnou klasifikaci skóringové funkce. Naším cílem je tedy hledat skóringovou funkci s co největší hodnotou Giniho koeficientu.

3.4 Odhad Giniho koeficientu

Pro odhad Giniho koeficientu lze v praxi použít více postupů. Jedním z často používaných je odhad pomocí tzv. *Somerovy d statistiky.*

Označíme-li s_j skóre j-tého klienta, můžeme definovat charakteristiky a, b a c následovně:

- a je počet všech dvojic klientů (i, j) , $i > j$ takových, že rozdíly $s_i - s_j$ a $y_i - y_j$ jsou nenulové a mají stejné znaménko (tedy takových dvojic, kde dobrý klient byl ohodnocen větším skóre než špatný klient);
- b je počet všech dvojic klientů (i, j) , $i > j$ takových, že rozdíly $s_i - s_j$ a $y_i - y_j$ jsou nenulové a mají opačné znaménko (tedy takových dvojic, kde dobrý klient byl ohodnocen menším skóre než špatný klient);
- c je počet všech dvojic klientů (i, j) , $i > j$ takových, že $s_i = s_j$ a $y_i \neq y_j$ (tedy takových dvojic, kde dobrý klient byl ohodnocen stejným skóre jako špatný klient).

Potom Somerovu d statistiku definujeme jako

$$d = \frac{a - b}{a + b + c}. \tag{3.4}$$

Takto definovaná hodnota d je potom odhadem Giniho koeficientu ve smyslu popsaných odhadů distribučních funkcí $\mathrm{F}^G(s)$, $\mathrm{F}^B(s)$. Toto ukážeme za předpokladu, že žádní dva klienti nemají stejné skóre, a tedy $c = 0$. Nechť

$$G = \big\{j : j \in \{1, \ldots, n\}, y_j = 1\big\}$$

je množina indexů dobrých klientů a

$$B = \big\{j : j \in \{1, \ldots, n\}, y_j = 0\big\}$$

množina indexů špatných klientů. Dále můžeme tedy psát

$$\mathrm{F}^G(s) = \mathrm{P}(s_j < s | j \in G) = \frac{|\{i : i \in G, s_i < s\}|}{|G|},$$

kde $|\cdot|$ značí mohutnost množiny, a analogicky

$$\mathrm{F}^B(s) = \mathrm{P}(s_j < s | j \in B) = \frac{|\{i : i \in B, s_i < s\}|}{|B|}.$$

Potom integrál z výrazu (3.2) můžeme vyjádřit sumou

$$\int_S \mathrm{F}^G(s)\,\mathrm{dF}^B(s) = \sum_{j=1}^{n} \mathrm{F}^G(s_j)\,\mathrm{P}(s_i = s_j | i \in B).$$

Protože předpokládáme, že žádní dva klienti nemají stejné skóre, je $\mathrm{P}(s_i = s_j | i \in B) = 0$ pro každé $j \in G$. Pro každé $j \in B$ potom pravděpodobnost $\mathrm{P}(s_i = s_j | i \in B)$ odhadujeme jako $\mathrm{P}(s_i = s_j | i \in B) = \frac{1}{|B|}$. Takto výraz dále upravujeme

$$\int_S \mathrm{F}^G(s)\,\mathrm{dF}^B(s) = \sum_{j=1}^{n} \mathrm{F}^G(s)\,\mathrm{P}(s_i = s_j | i \in B) =$$

$$= \frac{1}{|B|} \sum_{j \in B} \mathrm{F}^G(s_j) = \frac{1}{|B|} \sum_{j \in B} \frac{|\{i : i \in G, s_i < s_j\}|}{|G|}.$$

Tedy

$$\int_S \mathrm{F}^G(s)\,\mathrm{dF}^B(s) = \frac{1}{|B| \cdot |G|} \sum_{j \in B} |\{i : i \in G, s_i < s_j\}|.$$

Nyní si zbývá uvědomit, že $|B| \cdot |G|$ značí počet všech dvojic dobrých a špatných klientů, tedy $|B| \cdot |G| = a + b$ (podle předpokladu $c = 0$), a $\sum_{j \in B} |\{i : i \in G, s_i < s_j\}|$ je počet těch dvojic, kde dobrý klient byl ohodnocen menším skóre než špatný klient, tedy $\sum_{j \in B} |\{i : i \in G, s_i < s_j\}| = b$. Takto dostáváme

$$\int_S \mathrm{F}^G(s)\,\mathrm{dF}^B(s) = \frac{b}{a+b}. \tag{3.5}$$

Dosazením do vztahu (3.2) dostáváme

$$GC = 1 - 2\frac{b}{a+b} = \frac{a-b}{a+b}, \tag{3.6}$$

což je právě Somerova statistika d z definice (3.4) pro $c = 0$.

Pokud vynecháme předpoklad, že žádní dva klienti nemají stejné skóre, důkaz je možno vést obdobně. Je však technicky náročnější.

Kapitola 4

Skóringové modely

V této části se vycházíme z práce [4] Benešová P., Charamza P., *Rozlišovací schopnosti různých skóringových funkcí*, ve které jsou uvedeny aplikace tří základních příkladů skóringových modelů. Tyto modely se pokusíme formalizovat a za použití předcházející teorie objasnit podstatu jejich výstavby a odhadu parametrů.

4.1 Definice proměnných typu odds

Nejprve definujme několik proměnných používaných ve všech třech následujících modelech. Předpokládejme, že vysvětlující proměnné tvoří s skupin, kde i-tá skupina obsahuje s_i kategorií (znaků). Označme

$$Z = \big\{(i,j) : i \in \{1 \dots s\}, j \in \{1 \dots s_i\}\big\} \tag{4.1}$$

množinu všech uspořádaných dvojic (i,j), kde i značí skupinu a j její kategorii.

Dále předpokládejme, že každý klient spadá v každé skupině do právě jedné kategorie. Potom tedy pro každého klienta k máme sloupcový vektor $\boldsymbol{x}_k$, jehož prvky tvoří

$$\boldsymbol{x}_k = \big((x_j^i)_k : (i,j) \in Z\big), \tag{4.2}$$

kde v souladu s předchozím vysvětlením dummy proměnných (Kapitola 2) předpokládejme, že pokud nabývá k-tý klient v i-té skupině j-tý znak, pak

$(x_j^i)_k = 1$ a $(x_l^i)_k = 0$ pro všechny ostatní znaky l dané skupiny. Dále označme

$$G_j^i = \left\{k : k \in \{1, \ldots, n\}, y_k = 1, (x_j^i)_k = 1\right\}$$

množinu indexů všech dobrých klientů v j-té kategorii i-té skupiny a

$$B_j^i = \left\{k : k \in \{1, \ldots, n\}, y_k = 0, (x_j^i)_k = 1\right\}$$

množinu indexů všech špatných klientů v j-té kategorii i-té skupiny.

Nyní můžeme definovat proměnnou *odds*, tzv. *šanci celku*, jako poměr počtu dobrých klientů ku počtu špatných klientů

$$odds = \frac{|G|}{|B|} \tag{4.3}$$

a pro jednotlivé znaky j jednotlivých skupin i proměnné $odds_j^i$, tzv. *šance znaku*, jako poměry příslušných počtů dobrých a špatných klientů v dané kategorii

$$odds_j^i = \frac{|G_j^i|}{|B_j^i|}. \tag{4.4}$$

Nakonec zaveďme proměnnou *odds ratio*. Označme OR_j^i podíl $odds_j^i$ příslušné kategorie a *odds* celku, tedy

$$OR_j^i = \frac{odds_j^i}{odds}. \tag{4.5}$$

4.2 Podstata modelů

Na úvod poznamenejme, že zavedené označení proměnné *odds* souvisí s již dříve definovanou funkcí $\mathrm{odds}(\boldsymbol{x})$, viz (2.1), protože vzhledem k přirozeným odhadům pravděpodobnosti můžeme vyjádřit

$$odds = \frac{|G|}{|B|} = \frac{\frac{|G|}{|G \cup B|}}{\frac{|B|}{|G \cup B|}} \approx \frac{\mathrm{P}(Y = 1)}{\mathrm{P}(Y = 0)}.$$

Nyní se pokusme odhadnout teoretickou funkci $\mathrm{odds}(\boldsymbol{x})$ z definice (2.1) v závislosti na empirických hodnotách zavedených proměnných

$$\mathrm{odds}(\boldsymbol{x}) = \frac{\mathrm{P}(Y_{\boldsymbol{x}} = 1)}{\mathrm{P}(Y_{\boldsymbol{x}} = 0)} \approx \frac{\frac{|G_{\boldsymbol{x}}|}{|G_{\boldsymbol{x}} \cup B_{\boldsymbol{x}}|}}{\frac{|B_{\boldsymbol{x}}|}{|G_{\boldsymbol{x}} \cup B_{\boldsymbol{x}}|}} = \frac{|G_{\boldsymbol{x}}|}{|B_{\boldsymbol{x}}|}, \tag{4.6}$$

kde značíme

$$G_{\boldsymbol{x}} = \left\{k : k \in \{1, \ldots, n\}, y_k = 1, \boldsymbol{x}_k = \boldsymbol{x}\right\}$$

množinu indexů všech dobrých klientů s charakteristikou $\boldsymbol{x}$ a

$$B_{\boldsymbol{x}} = \left\{k : k \in \{1, \ldots, n\}, y_k = 0, \boldsymbol{x}_k = \boldsymbol{x}\right\}$$

množinu indexů všech špatných klientů s charakteristikou $\boldsymbol{x}$.

Protože hodnoty $|G_{\boldsymbol{x}}|$ a $|B_{\boldsymbol{x}}|$ závisí na konkrétní kombinaci hodnot vektoru $\boldsymbol{x}$ a těchto kombinací je obecně velmi mnoho (konkrétně $\prod_{i=1}^{s} s_i$), není v praxi vhodné funkci $\text{odds}(\boldsymbol{x})$ odhadovat vztahem (4.6). Proto se tento vztah pokusme dále upravit

$$\text{odds}(\boldsymbol{x}) = \frac{|G_{\boldsymbol{x}}|}{|B_{\boldsymbol{x}}|} = \frac{|G|}{|B|} \frac{\frac{|G_{\boldsymbol{x}}|}{|G|}}{\frac{|B_{\boldsymbol{x}}|}{|B|}}. \tag{4.7}$$

Protože $\frac{|G_{\boldsymbol{x}}|}{|G|}$ je možno interpretovat jako empirický odhad pravděpodobnosti, že dobrý klient bude mít charakteristiku $\boldsymbol{x}$, můžeme za předpokladu nezávislosti regresorů[1] tuto pravděpodobnost přepsat ve tvaru součinu

$$\frac{|G_{\boldsymbol{x}}|}{|G|} = \prod_{(i,j) \in Z} \left(\frac{|G_j^i|}{|G|}\right)^{x_j^i},$$

tedy jako součin činitelů těch kategorií, pro které $x_j^i = 1$. Obdobně rozepíšeme $\frac{|B_{\boldsymbol{x}}|}{|B|}$ a dosadíme do vztahu (4.7). Tento potom dále upravujeme

$$\text{odds}(\boldsymbol{x}) = \frac{|G|}{|B|} \frac{\prod_{(i,j) \in Z} \left(\frac{|G_j^i|}{|G|}\right)^{x_j^i}}{\prod_{(i,j) \in Z} \left(\frac{|B_j^i|}{|B|}\right)^{x_j^i}} = \frac{|G|}{|B|} \prod_{(i,j) \in Z} \left(\frac{\frac{|G_j^i|}{|B_j^i|}}{\frac{|G|}{|B|}}\right)^{x_j^i}. \tag{4.8}$$

Nyní s použitím zavedeného značení můžeme vztah dále přepsat do tvaru

$$\text{odds}(\boldsymbol{x}) = odds \prod_{(i,j) \in Z} \left(\frac{odds_j^i}{odds}\right)^{x_j^i} = odds \prod_{(i,j) \in Z} (OR_j^i)^{x_j^i}. \tag{4.9}$$

Tento vztah je spolu s předpokladem nezávislosti regresorů základem tzv. *Independence modelu*.

[1] Předpoklad nezávislosti regresorů je v praxi většinou těžko dosažitelný, proto často přecházíme ke komplikovanějším modelům.

4.3 Independence model

Independence model je nejjednodušším z trojice představovaných modelů ohodnocení kreditního rizika. Skóringová funkce vychází pouze z vypočítaných hodnot proměnných *odds* a OR_j^i. Na základě předchozích úvah definujme skóringovou funkci $\mathrm{S}^{IM}(\boldsymbol{x})$ jako

$$\mathrm{S}^{IM}(\boldsymbol{x}) = odds \prod_{(i,j)\in Z} (OR_j^i)^{x_j^i}, \tag{4.10}$$

kde $\boldsymbol{x} = \left(x_j^i : (i,j) \in Z\right)$ je sada nezávisle proměnných, která charakterizuje hodnoceného klienta.

Odtud vidíme, že skóringová funkce $\mathrm{S}^{IM}(\boldsymbol{x})$ je tvořena součinem *odds* a OR_j^i právě těch kategorií, ve kterých se příslušný klient nachází. Tento přístup modelování skóringové funkce se často používá právě pro svou jednoduchost. Jeho podstatnou nevýhodou však je předpoklad nezávislosti regresorů a fakt, že model přikládá všem hodnotám OR_j^i stejnou váhu a tím snižuje svou vypovídací schopnost.

V praxi se někdy jako skóre používá logaritmus uvedeného vztahu.

$$\ln\left(\mathrm{S}^{IM}(\boldsymbol{x})\right) = \ln(odds) + \sum_{(i,j)\in Z} x_j^i \ln(OR_j^i). \tag{4.11}$$

4.4 WOE model

Dalším možným přístupem k modelování kreditního rizika pomocí skóringové funkce je *WOE model*. WOE je zkratka z anglického *weight of evidence* a značí, že v modelu přiřadíme každé skupině jinou váhu podle toho, jaký je její statistický vliv na hodnotu vysvětlované proměnné $Y_{\boldsymbol{x}}$. Takový model potom můžeme vyjádřit ve tvaru

$$\mathrm{S}^{WOE}(\boldsymbol{x}, \boldsymbol{\lambda}) = odds \prod_{(i,j)\in Z} (OR_j^i)^{\lambda^i x_j^i}, \tag{4.12}$$

kde $\boldsymbol{x} = \left(x_j^i : (i,j) \in Z\right)$ je opět sada nezávisle proměnných a $\boldsymbol{\lambda} = \left(\lambda^i : i \in \{1 \dots s\}\right)$ je vektor vah jednotlivých skupin.

Takto vytvořená skóringová funkce je opět odhadem funkce odds($\boldsymbol{x}$), proto její logaritmus je funkce logit($\boldsymbol{x}$) a vektor parametrů $\boldsymbol{\lambda} = \left(\lambda^i : i \in \{1 \dots s\}\right)$ je možno odhadovat metodou logistické regrese popsanou v Kapitole 2 pomocí vztahu

$$\text{logit}(\boldsymbol{x}) = \ln\left(\text{S}^{WOE}(\boldsymbol{x}, \boldsymbol{\lambda})\right) = \ln(odds) + \sum_{(i,j)\in Z} \lambda^i x_j^i \ln(OR_j^i) = \boldsymbol{\beta}'\boldsymbol{z}. \quad (4.13)$$

Tento model je výpočetně náročnější, avšak zvláště pro větší databáze poskytuje větší přesnost a částečně tak řeší nedostatky independence modelu. Podle [3] str. 19 je tento model vhodný pro databáze s více než 150 defaulty (případy nesplacení).

4.5 Plný logistický model

Plný logistický model přiřazuje specifickou váhu každému jednotlivému znaku. Takto získáváme pro skóringovou funkci definiční vztah

$$\text{S}^{PLM}(\boldsymbol{x}, \boldsymbol{\lambda}) = odds \prod_{(i,j)\in Z} (OR_j^i)^{\lambda_j^i x_j^i}, \quad (4.14)$$

kde $\boldsymbol{x} = \left(x_j^i : (i,j) \in Z\right)$ je sada nezávisle proměnných a $\boldsymbol{\lambda} = \left(\lambda_j^i : (i,j) \in Z\right)$ vektor vah jednotlivých znaků.

Podobně jako u *WOE modelu* odhadneme vektor parametrů $\boldsymbol{\lambda} = \left(\lambda_j^i : (i,j) \in Z\right)$ metodou logistické regrese

$$\text{logit}(\boldsymbol{x}) = \ln\left(\text{S}^{PLM}(\boldsymbol{x}, \boldsymbol{\lambda})\right) = \ln(odds) + \sum_{(i,j)\in Z} \lambda_j^i x_j^i \ln(OR_j^i) = \boldsymbol{\beta}'\boldsymbol{z}. \quad (4.15)$$

Tento model je nejpřesnější z uvedené trojice modelů, ale také výpočetně nejnáročnější. V praxi se většinou používá pro velmi rozsáhlé databáze, podle [3] str. 19 pro databáze s více jak 1200 defaulty.

Kapitola 5

Skóringové modely na reálných datech

5.1 Data

Data jsme získali z internetových stránek *Institut für Statistik der Ludwig-Maximilians-Universität München* (http://www.stat.uni-muenchen.de), viz také [7] str. 14–22. Databáze obsahuje historické údaje o 1000 klientech jedné německé banky. Pro každého klienta máme k dispozici sadu 20 vysvětlujících proměnných a informaci o tom, zda klient úvěr splatil či nikoliv. Jde zde o 700 případů splacení a 300 defaultů. Všechny proměnné v databázi jsou již rozděleny do kategorií. Popis všech proměnných a příslušných kategorií nalezneme v Příloze A.

V následujících odstavcích se pokusíme na reálných datech porovnat vhodnost a diverzifikační schopnost independence modelu, WOE modelu a plného logistického modelu. Budeme zkoumat závislost proměnné SPLACENO na hodnotách vysvětlujících proměnných ÚČET, SPLATNOST, MORÁLKA, ÚČEL a ÚSPORY.[1] K tomuto účelu tyto proměnné transformujeme na soustavu dummy proměnných podle jednotlivých kategorií.

Pro proměnné ÚČET, SPLATNOST, MORÁLKA, ÚČEL a ÚSPORY znázorněme poměrné zastoupení jednotlivých kategorií na celkovém vzorku

[1]Tuto pětici proměnných jsme získali metodou *stepwise selection* (viz např. [5] str. 116–128) při hodnotě hladiny vstupu $\alpha_1 = 0,005$ a hladiny výstupu $\alpha_2 = 0,005$ v SAS Learning Edition 2.0.

a zastoupení špatných klientů v jednotlivých kategoriích (Příloha B).

5.2 Independence model – zpracování dat

Pro sestrojení *independence modelu* nejprve spočítáme hodnoty $odds_j^i$ a OR_j^i pro jednotlivé kategorie všech proměnných (viz (5.1) a Tabulka 5.1). Pro práci s databází a příslušné výpočty používáme MS Excel 2003. Proměnné ÚČEL 7 nenabývá žádný klient z databáze, proto ji z modelu vypustíme.

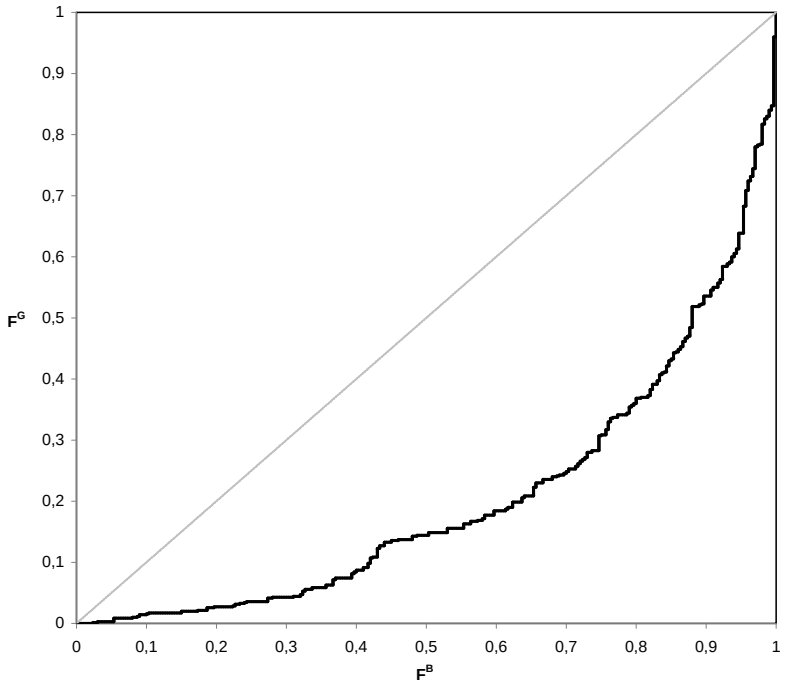
$$odds = \frac{|G|}{|B|} = \frac{700}{300} = 2,333. \tag{5.1}$$

ÚČET	1	2	3	4	
OR_j^i	0,441	0,669	1,500	3,242	
SPLATNOST	1	2	3	4	5
OR_j^i	0,429	0,429	0,295	1,029	0,541
SPLATNOST	6	7	8	9	10
OR_j^i	0,857	1,026	1,003	1,343	3,476
MORÁLKA	0	1	2	3	4
OR_j^i	0,257	0,321	0,915	0,918	2,083
ÚČEL	0	1	2	3	4
OR_j^i	0,698	2,168	0,909	1,507	0,857
ÚČEL	5	6	8	9	10
OR_j^i	0,750	0,545	3,429	0,794	0,600
ÚSPORY	1	2	3	4	5
OR_j^i	0,762	0,870	2,026	3,000	2,022

Tabulka 5.1: Hodnoty OR

Hodnotu skóre pro každého nového klienta potom získáme jako součin celkového *odds* a příslušných OR_j^i těch kategorií, ve kterých se klient nachází. Když takto dopočítáme hodnoty skóre pro všechny klienty v databázi, umožní nám to vykreslit *Lorenzovu křivku* (Obrázek 5.1) a pomocí *Somerovy d statistiky* odhadnout *Giniho koeficient*.

$$\widehat{GC^{IM}} = 0,592.$$



Obrázek 5.1: Independence model – Lorenzova křivka

5.3 WOE model – zpracování dat

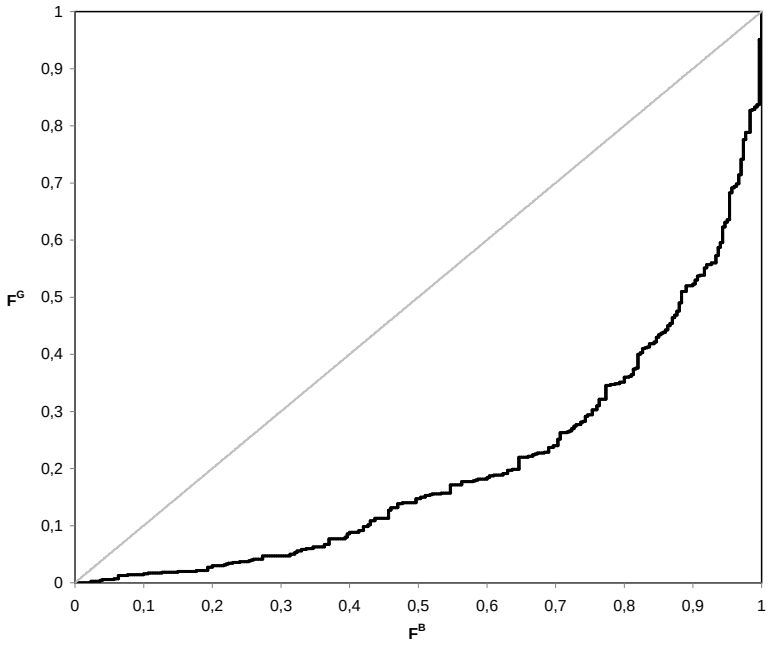
Nyní do modelu přidáme parametrický váhový vektor $\boldsymbol{\lambda}$ s rozdílnou hodnotou pro každou skupinu. Tyto hodnoty odhadneme metodou logistické regrese pomocí statistického software SAS Learning Edition 2.0. K tomuto účelu nejdříve transformujeme databázi, aby pro každého klienta každá skupinová proměnná obsahovala hodnotu $\ln(OR_j^i)$ té kategorie, ve které se klient nachází. K tomuto použijeme opět MS Excel. Odhadnuté hodnoty parametrů uveďme v Tabulce 5.2.

Pro odhad *Giniho koeficientu* a vykreslení *Lorenzovy křivky* (Obrázek 5.2) opět dopočítáme odhadnuté hodnoty skóre.

$$\widehat{GC^{WOE}} = 0,595.$$

Proměnná	$\widehat{\lambda^i}$
INTERCEPT	0,842
ÚČET	0,826
SPLATNOST	0,992
MORÁLKA	0,786
ÚČEL	0,975
ÚSPORY	0,739

Tabulka 5.2: WOE – odhady parametrů



Obrázek 5.2: WOE model – Lorenzova křivka

5.4 Plný logistický model – zpracování dat

V tomto modelu uvažujeme parametrický váhový vektor $\boldsymbol{\lambda}$ s rozdílnými hodnotami pro každou kategorii zvlášť. Tyto parametry odhadneme opět metodou logistické regrese v SAS a shrneme do Tabulky 5.3. Dále spočítáme odhad *Giniho koeficientu* a vykreslíme *Lorenzovu křivku* (Obrázek 5.3).

INTERCEPT	5,039				
ÚČET	1	2	3	4	
$\widehat{\lambda}_j^i$	2,059	3,198	-1,608	0	
SPLATNOST	1	2	3	4	5
$\widehat{\lambda}_j^i$	2,651	2,742	2,149	-46,174	3,455
SPLATNOST	6	7	8	9	10
$\widehat{\lambda}_j^i$	10,916	-58,267	-527,600	-3,398	0
MORÁLKA	0	1	2	3	4
$\widehat{\lambda}_j^i$	1,199	1,498	8,082	7,613	0
ÚČEL	0	1	2	3	4
$\widehat{\lambda}_j^i$	3,003	0,451	5,498	-0,639	4,561
ÚČEL	5	6	8	9	10
$\widehat{\lambda}_j^i$	3,055	2,427	0,638	1,654	0
ÚSPORY	1	2	3	4	5
$\widehat{\lambda}_j^i$	3,161	4,602	-0,556	0,164	0

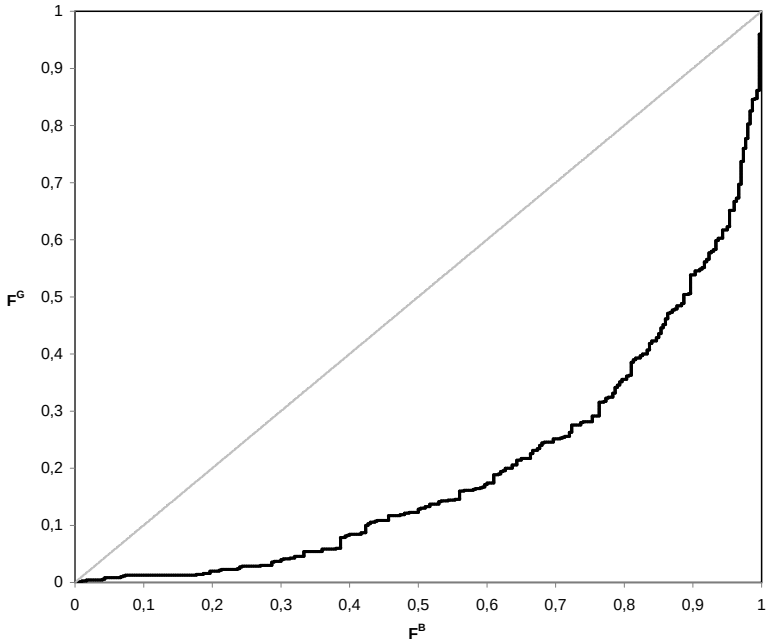
Tabulka 5.3: Plný logistický model – odhady parametrů

$$\widehat{GC^{PLM}} = 0,605.$$

5.5 Porovnání modelů

Porovnáme-li uvedenou trojici modelů z hlediska diverzifikační schopnosti, vidíme, že odhadovaná hodnota Giniho koeficientu je největší u plného logistického modelu. Přesto se však od independence modelu liší jen o 0,013, tedy o 1,3 procentního bodu. Proto pro danou datovou sadu a vybranou pětici vysvětlujících proměnných není onen rozdíl až tolik významný a lze tedy usuzovat, že jednotlivé proměnné a kategorie ovlivňují diverzifikační schopnost modelu podobnou měrou.

Z hlediska výpočetní náročnosti a interpretační jednoduchosti bychom nejspíše vyzdvihli independence model, který nevyužívá odhadu parametrů a vychází pouze z pozorovaných hodnot. Na tomto místě je však vhodné zmínit, že plný logistický model je v nějaké své modifikaci implementován



Obrázek 5.3: Plný logistický model – Lorenzova křivka

ve většině statistických software (včetně SAS) a tím se jeho použití stává výrazně pohodlnějším než vlastní počítání příslušných hodnot *odds* a OR_j^i.

Porovnání vypočítaných parametrů všech tří modelů je nejlépe názorné v Tabulce 5.4. V prvních sloupcích jsou pro jednotlivé proměnné spočítány hodnoty $odds_j^i$, OR_j^i a $\ln(OR_j^i)$, které přímo vystupují v independence modelu a jsou výchozími hodnotami pro odhad parametrů dalších dvou modelů. V dalších sloupcích jsou odhadnuté hodnoty parametrů λ^i a λ_j^i jako vah příslušných skupin a kategorií. Poslední tři sloupce tabulky tvoří tzv. *WOE* jednotlivých modelů, které získáme jako součin $\ln(OR_j^i)$ a příslušného parametru zvětšený o poměrnou část interceptu. Pro independence model (sloupec IM) jsou to hodnoty $\ln(OR_j^i) + \frac{1}{5}\ln(odds)$, pro WOE model (sloupec WOE) je to součin $\lambda^i \ln(OR_j^i) + \frac{1}{5}\lambda^0$ a pro plný logistický model součin $\lambda_j^i \ln(OR_j^i) + \frac{1}{5}\lambda_0^0$. Sečtením příslušných hodnot WOE takto pro každého klienta získáme skóre odpovídající logaritmické skóringové funkci použitého modelu.

Rozsáhlejší databáze bývají v praxi často rozděleny na *vývojovou část*, pomocí které se sestrojí vhodný model a odhadnou jeho parametry, a *ověřovací část*, na které se potom model testuje. Takové rozdělení sice sníží datový vzorek pro odhad parametrů, ale poskytne informace o stabilitě modelu.

5.6 Kompletní model logistické regrese

Pro úplnost ještě uveďme výsledky použití *plného logistického modelu* na úplné databázi s použitím všech proměnných. Tento výpočet proveďme za pomocí implementovaného postupu v SAS.

Zvolíme-li metodu *stepwise selection* (viz např. [5] str. 116–128) s hladinou vstupu do modelu i výstupu z modelu rovnou $\alpha = 0,05$, dostaneme model s deseti vysvětlujícími proměnnými: ÚSPORY, ÚČET, SPLATNOST, DOBAZAM, VÝŠE, MORÁLKA, RUČENÍ, ÚČEL, BYDLENÍ a CIZINEC. Pro tento model máme spočítánu také hodnotu D (deviance) (5.2) a odhad Giniho koeficientu (5.3) pomocí Somerovy d statistiky. Můžeme také vykreslit Lorenzovu křivku (Obrázek 5.4).

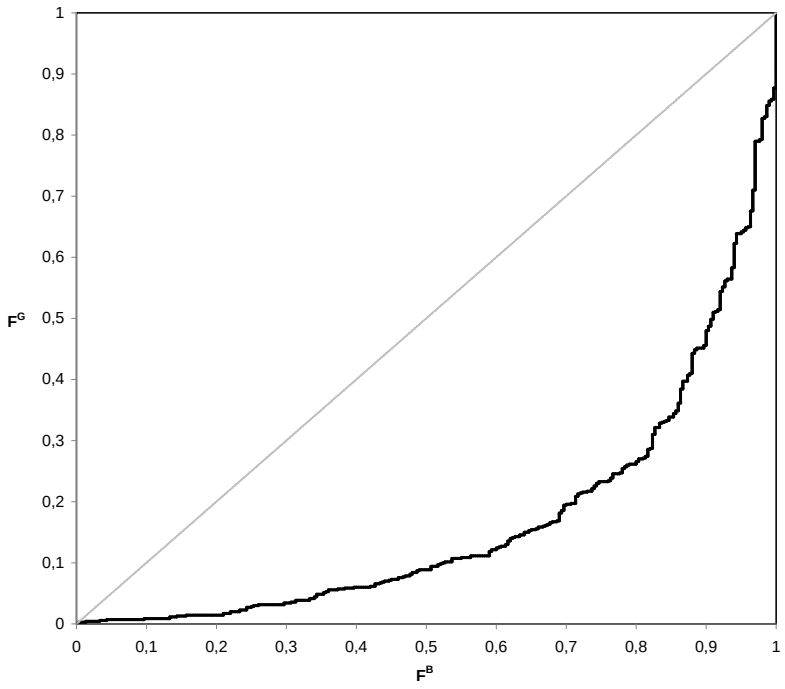
$$D = 905,955. \tag{5.2}$$

$$\widehat{GC} = 0,667. \tag{5.3}$$

Všechny proměnné tohoto modelu jsou sice na hladině $\alpha = 0,05$ statisticky významné, ale na příslušný rozsah databáze (1000) je počet parametrů modelu (48) až příliš velký a s ohledem na stabilitu modelu bychom se v tomto případě nejspíš rozhodli pro WOE model, který má parametrů výrazně méně. To také odpovídá doporučení z [3] str. 19 pro 300 defaultů.

Proměnná		*odds*	OR	$\ln(OR)$	λ^i	λ^i_j	IM	WOE	PLM
ÚČET	1	1,030	0,441	-0,818	0,826	2,059	-0,649	-0,533	-0,830
	2	1,562	0,669	-0,401	0,826	3,198	-0,232	-0,189	-0,430
	3	3,500	1,500	0,405	0,826	-1,068	0,575	0,478	0,421
	4	7,565	3,242	1,176	0,826	0,000	1,346	1,115	0,854
SPLATNOST	1	1,000	0,429	-0,847	0,992	2,651	-0,678	-0,698	-1,392
	2	1,000	0,429	-0,847	0,992	2,742	-0,678	-0,698	-1,469
	3	0,688	0,295	-1,222	0,992	2,149	-1,053	-1,069	-1,773
	4	2,400	1,029	0,028	0,992	-46,174	0,198	0,171	-0,447
	5	1,263	0,541	-0,614	0,992	3,455	-0,444	-0,466	-1,267
	6	2,000	0,857	-0,154	0,992	10,916	0,015	-0,010	-0,829
	7	2,394	1,026	0,026	0,992	-58,267	0,195	0,168	-0,640
	8	2,339	1,003	0,003	0,992	-527,600	0,172	0,145	-0,490
	9	3,134	1,343	0,295	0,992	-3,398	0,465	0,435	-0,149
	10	8,111	3,476	1,246	0,992	0,000	1,415	1,378	0,854
MORÁLKA	0	0,600	0,257	-1,358	0,786	1,199	-1,189	-0,924	-0,775
	1	0,750	0,321	-1,135	0,786	1,498	-0,966	-0,749	-0,846
	2	2,136	0,915	-0,088	0,786	8,082	0,081	0,073	0,140
	3	2,143	0,918	-0,085	0,786	7,613	0,084	0,076	0,206
	4	4,860	2,083	0,734	0,786	0,000	0,903	0,719	0,854
ÚČEL	0	1,629	0,698	-0,359	0,975	3,003	-0,190	-0,208	-0,225
	1	5,059	2,168	0,774	0,975	0,451	0,943	0,897	1,203
	2	2,121	0,909	-0,096	0,975	5,498	0,074	0,049	0,329
	3	3,516	1,507	0,410	0,975	-0,639	0,580	0,543	0,592
	4	2,000	0,857	-0,154	0,975	4,561	0,015	-0,008	0,151
	5	1,750	0,750	-0,288	0,975	3,055	-0,118	-0,138	-0,025
	6	1,273	0,545	-0,606	0,975	2,427	-0,437	-0,448	-0,617
	8	8,000	3,429	1,232	0,975	0,638	1,402	1,344	1,640
	9	1,853	0,794	-0,231	0,975	1,654	-0,061	-0,082	0,473
	10	1,400	0,600	-0,511	0,975	0,000	-0,341	-0,356	0,854
ÚSPORY	1	1,779	0,762	-0,271	0,739	3,161	-0,102	-0,058	-0,004
	2	2,029	0,870	-0,140	0,739	4,602	0,030	0,040	0,212
	3	4,727	2,026	0,706	0,739	-0,556	0,876	0,664	0,462
	4	7,000	3,000	1,099	0,739	0,164	1,268	0,954	1,034
	5	4,719	2,022	0,704	0,739	0,000	0,874	0,663	0,854

Tabulka 5.4: Porovnání modelů



Obrázek 5.4: Kompletní model logistické regrese – Lorenzova křivka

Kapitola 6

Závěr

V této práci jsme se nejdříve věnovali popisu metody logistické regrese, kterou jsme dále použili na odhad parametrů vybraných skóringových modelů. Zabývali jsme se podstatou výstavby těchto modelů a charakteristikami míry jejich diverzifikační schopnosti. Výstavbu modelů jsme nakonec demonstrovali na reálném datovém vzorku.

Cílem práce bylo uceleným způsobem popsat matematický základ vybraných skóringových modelů a vysvětlit jejich odvození. Cílem praktické části práce bylo potom na konkrétním příkladě porovnat vhodnost použití jednotlivých modelů, jejich výpočetní náročnost a diverzifikační schopnost.

V našem případě se osvědčil již nejjednodušší independence model, jehož diverzifikační schopnost byla srovnatelná s oběma složitějšími modely. Zvláště pro databáze menšího rozsahu je tento model vhodný. Použití plného logistického modelu obecné přináší větší diverzifikační schopnost, ale pro zachování stability modelu je zapotřebí velmi rozsáhlá databáze s vývojovou a ověřovací částí. Ve všech případech je zapotřebí následný monitoring WOE charakteristik po implementaci modelu.

Literatura

[1] Agresti A.: *Categorical Data Analysis*, John Wiley & Sons, Inc., 1990.

[2] Anděl J.: *Základy matematické statistiky*, MATFYZPRESS, 2007.

[3] Aspey J., Hinder J., Lucas A.: *Rhino Risk Mission Statement*, http://www.crc.man.ed.ac.uk/conference/archive/2003/presentations/lucas2.pdf.

[4] Benešová P., Charamza P.: *Rozlišovací schopnosti různých skóringových funkcí*, FSV UK 2008.

[5] Hosmer D. W., Lemeshow S.: *Applied Logistic Regression*, John Wiley & Sons, Inc., 2000.

[6] Hušek R.: *Ekonometrická analýza*, Vysoká škola ekonomická v Praze, Nakladatelství Oeconomica, 2007.

[7] Kračmerová L.: *Metody zpracování kategoriálních finančních dat*, MFF UK 2007.

Příloha A

Popis proměnných

Název proměnné	Význam	Kategorie	Skóre
SPLACENO	splacení úvěru	úvěr splacen	1
		úvěr nesplacen	0
ÚČET	množství peněz na účtu (DM)	žádné nebo debet	2
		méně než 200	3
		více jak 200	4
		nemá účet	1
SPLATNOST	doba do splatnosti (měsíce)	méně než 6	10
		6 – 12	9
		12 – 18	8
		18 – 24	7
		24 – 30	6
		30 – 36	5
		36 – 42	4
		42 – 48	3
		48 – 54	2
		více než 54	1
MORÁLKA	splácení předchozích úvěrů	žádné předchozí úvěry	2
		splacené úvěry	4
		současné úvěry spláceny	3
		váhavé splácení	0
		úvěry u jiných bank	1

Tabulka A.1: Popis proměnných

Název proměnné	Význam	Kategorie	Skóre
ÚČEL	účel úvěru	nový automobil	1
		ojetý automobil	2
		nábytek	3
		rádio nebo televize	4
		zařízení bytu	5
		opravy	6
		vzdělání	7
		dovolená	8
		rekvalifikace	9
		obchod	10
		jiný	0
VÝŠE	výše úvěru (DM)	méně než 500	10
		500 – 1000	9
		1000 – 1500	8
		1500 – 2500	7
		2500 – 5000	6
		5000 – 7500	5
		7500 – 10000	4
		10000 – 15000	3
		15000 – 20000	2
		více než 20000	1
ÚSPORY	výše úspor a cenných papírů (DM)	méně než 100	2
		100 – 500	3
		500 – 1000	4
		více než 1000	5
		žádné nebo nezjištěno	1
DOBAZAM	doba současného zaměstnání (roky)	nezaměstnaný	1
		méně než 1	2
		1 – 4	3
		4 – 7	4
		více než 7	5

Tabulka A.2: Popis proměnných

Název proměnné	Význam	Kategorie	Skóre
POMĚR	poměr výše splátky ku příjmu (%)	více než 35	1
		25 – 35	2
		20 – 25	3
		méně než 20	4
STAV	pohlaví a rodinný stav	M: rozvedený	1
		Ž: vdaná, rozvedená, M: svobodný	2
		M: ženatý, vdovec	3
		Ž: svobodná	4
RUČENÍ	způsob ručení	žádný	1
		spolužadatel	2
		ručitel	3
DOBABYD	v současné domácnosti (roky)	méně než 1	1
		1 – 4	2
		4 – 7	3
		více než 7	4
AKTIVA	cenná aktiva	vlastník nemovitosti	4
		stavební spoření, životní pojištění	3
		automobil nebo jiná	2
		žádná dostupná	1
VĚK	věk (roky)	méně než 25	1
		26 – 39	2
		40 – 59	3
		60 – 64	5
		více než 65	4
ÚVĚRY	další úvěry	v jiných bankách	1
		v obchodech	2
		žádné	3
BYDLENÍ	typ bydlení	zdarma	1
		byt v nájmu	2
		vlastní byt	3

Tabulka A.3: Popis proměnných

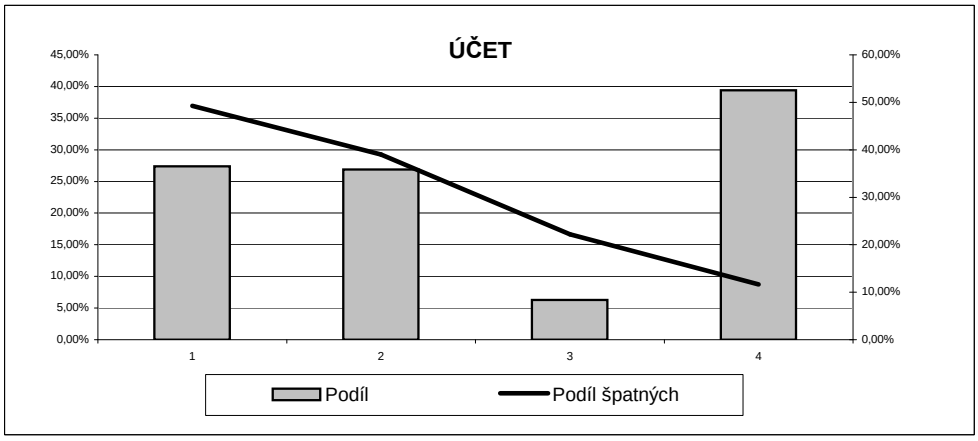
Název proměnné	Význam	Kategorie	Skóre
POČETÚVĚRŮ	počet úvěrů v bance	1	1
		2 – 3	2
		4 – 6	3
		více než 6	4
ZAMĚSTNÁNÍ	zaměstnání	nezaměstnaný	1
		nevyučený	2
		kvalifikovaný	3
		vedoucí pracovník	4
VYŽIVOVANÍ	počet vyživovaných	méně než 3	2
		3 a více	1
TELEFON	telefon	ne	1
		ano	2
CIZINEC	pracující cizinec	ano	1
		ne	2

Tabulka A.4: Popis proměnných

Příloha B

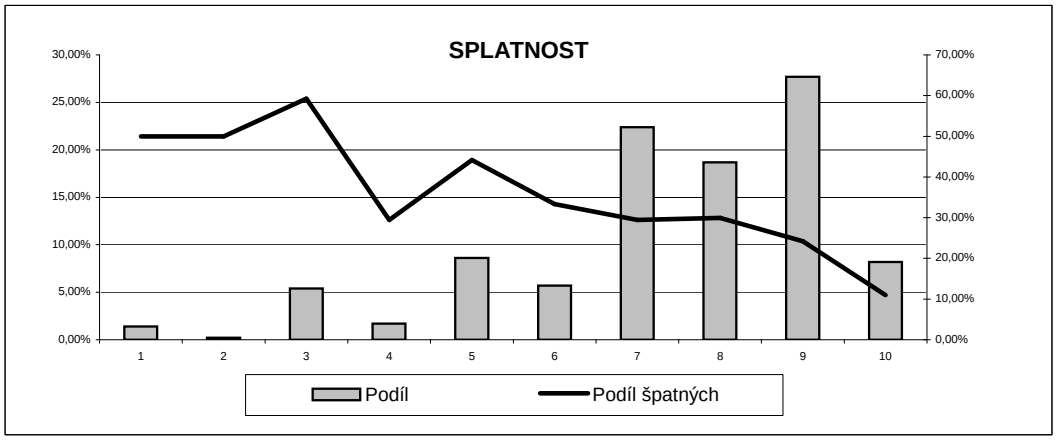
Tabulky splacení

ÚČET	Počet	Podíl	Počet špatných	Podíl špatných
ÚČET 1	274	27,40%	135	49,27%
ÚČET 2	269	26,90%	105	39,03%
ÚČET 3	63	6,30%	14	22,22%
ÚČET 4	394	39,40%	46	11,68%
Celkem	1000	100,00%	300	30,00%



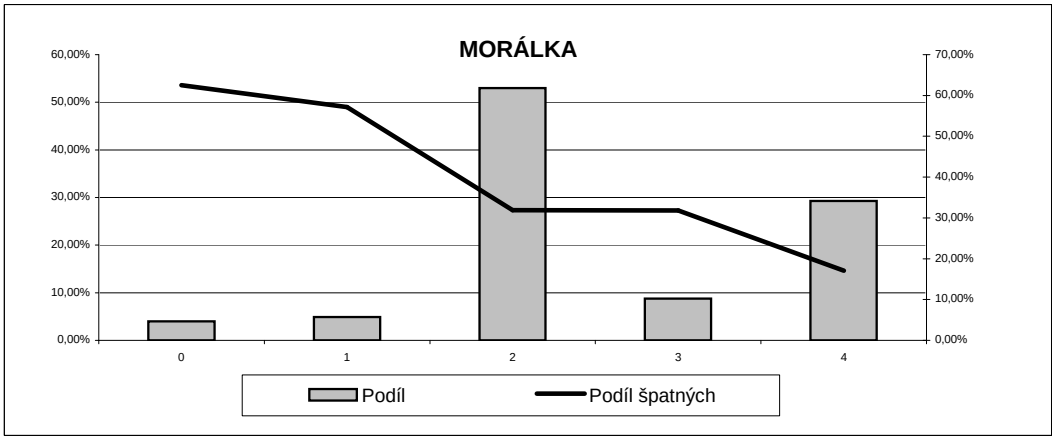
Tabulka B.1: ÚČET – množství peněz na účtu

SPLATNOST	Počet	Podíl	Počet špatných	Podíl špatných
SPLATNOST 1	14	1,40%	7	50,00%
SPLATNOST 2	2	0,20%	1	50,00%
SPLATNOST 3	54	5,40%	32	59,26%
SPLATNOST 4	17	1,70%	5	29,41%
SPLATNOST 5	86	8,60%	38	44,19%
SPLATNOST 6	57	5,70%	19	33,33%
SPLATNOST 7	224	22,40%	66	29,46%
SPLATNOST 8	187	18,70%	56	29,95%
SPLATNOST 9	277	27,70%	67	24,19%
SPLATNOST 10	82	8,20%	9	10,98%
Celkem	1000	100,00%	300	30,00%



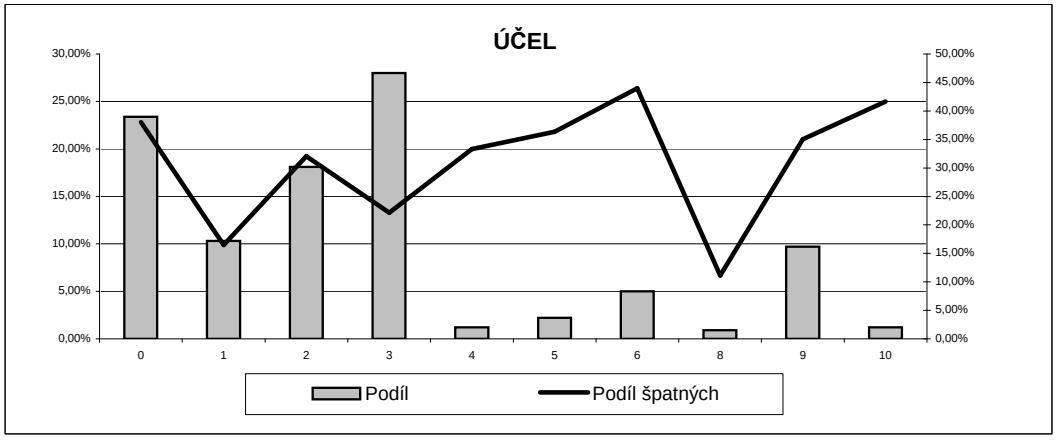
Tabulka B.2: SPLATNOST – doba do splatnosti

MORÁLKA		Počet	Podíl	Počet špatných	Podíl špatných
MORÁLKA	0	40	4,00%	25	62,50%
MORÁLKA	1	49	4,90%	28	57,14%
MORÁLKA	2	530	53,00%	169	31,89%
MORÁLKA	3	88	8,80%	28	31,82%
MORÁLKA	4	293	29,30%	50	17,06%
Celkem		1000	100,00%	300	30,00%



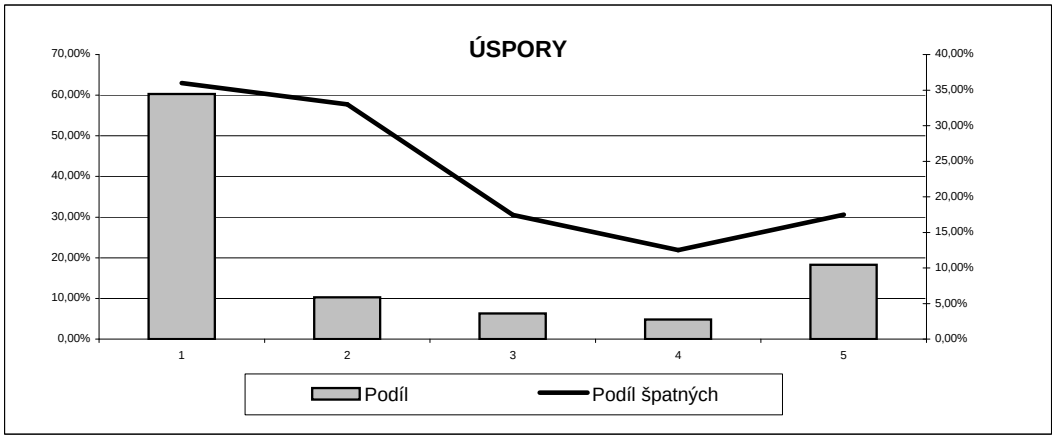
Tabulka B.3: MORÁLKA – splácení předchozích úvěrů

ÚČEL		Počet	Podíl	Počet špatných	Podíl špatných
ÚČEL	0	234	23,40%	89	38,03%
ÚČEL	1	103	10,30%	17	16,50%
ÚČEL	2	181	18,10%	58	32,04%
ÚČEL	3	280	28,00%	62	22,14%
ÚČEL	4	12	1,20%	4	33,33%
ÚČEL	5	22	2,20%	8	36,36%
ÚČEL	6	50	5,00%	22	44,00%
ÚČEL	8	9	0,90%	1	11,11%
ÚČEL	9	97	9,70%	34	35,05%
ÚČEL	10	12	1,20%	5	41,67%
Celkem		1000	100,00%	300	30,00%



Tabulka B.4: ÚČEL – účel úvěru

ÚSPORY		Počet	Podíl	Počet špatných	Podíl špatných
ÚSPORY	1	603	60,30%	217	35,99%
ÚSPORY	2	103	10,30%	34	33,01%
ÚSPORY	3	63	6,30%	11	17,46%
ÚSPORY	4	48	4,80%	6	12,50%
ÚSPORY	5	183	18,30%	32	17,49%
Celkem		1000	100,00%	300	30,00%



Tabulka B.5: ÚSPORY – výše úspor a cenných papírů