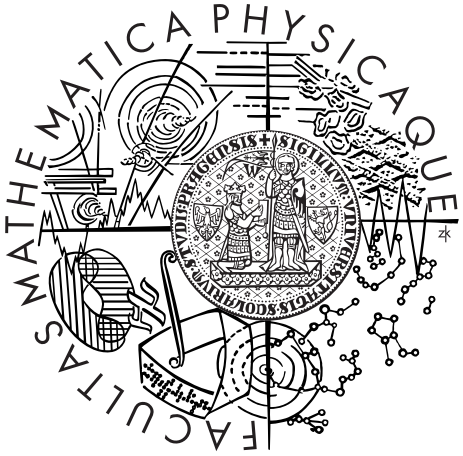
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Michal Polak

Míry diskriminace v kreditním riziku

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Michal Pešta, Ph.D.
Studijní program: Matematika
Studijní obor: Finanční matematika

Praha 2013

Děkuji RNDr. Michal Peštovi, Ph.D. za cenné připomínky a rady, které byly přínosné pro vypracování této bakalářské práce.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 2.8.2013 Michal Polak

Název práce: Míry diskriminace v kreditním riziku

Autor: Michal Polak

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Skóringové modely jsou základním nástrojem pro moderní řízení kreditního rizika. Přispívá tomu hlavně značný vývoj informačních technologií. Využívají se nejenom při poskytování úvěru, ale i ve strategiích týkajících se budoucího řízení kreditního rizika nebo ve strategiích spojených s vymáháním pohledávek. V přiložené práci se věnujeme mírám diskriminace používané pro validaci diverzifikační schopnosti logistických skóringových modelů. Prvně se zabýváme pojmem rizika. Poté uvedeme základní dělení skóringových modelů. Dále popisujeme metodu logistické regrese, odhadování a význam parametrů a testování jejich významnosti. Pro změření a znázornění diverzifikační schopnosti modelu jsme uvedli nejběžněji používané metody jako Lorenzovu a ROC křivku, Giniho koeficient, c-statistiku a taky Kolmogorov-Smirnovův test. Závěrem aplikujeme teoretické poznatky na reálných datech. Zkonstruujeme skóringový model a posléze porovnáme míry diskriminace v uvedeném modelu. Klíčová slova: kreditní riziko, skóringové modely, logistická regrese, míry diskriminace

Title: Discrimination Measures in Credit Risk

Author: Michal Polak

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Scoring models represent a fundamental tool for the modern management of credit risk. This is mainly due to a significant development in the field of information technology. Such models are used not only when providing credit, but also in strategies relating to the future management of credit risk, or in strategies connected with enforcing receivables. In my thesis I deal with discrimination measures used in the validation of diversification potential of logistic scoring models. At the beginning, I focus on the term 'risk'. Then, I introduce a basic division of scoring models. Next, I describe the method of scoring logistic regression, I concentrate on estimating parameters, their significance and on testing their relevance. For the measurement and illustration of diversification potential of the model I mention the most commonly used methods such as the Lorenz and ROC curve, the Gini coeficient, the c-statistic as well as the Kolmogorov-Smirnov test. Finally, I apply the theoretical knowledge to real data. I design a scoring model and subsequently compare the discrimination measures which it contains.

Keywords: credit risk, scoring models, logistic regression, discrimination measures

Obsah

Úvod **2**

1 Definice rizika **3**
1.1 Kreditní riziko . . . 3

2 Credit scoring **5**
2.1 Základní dělení . . . 5
2.2 Podstata skóringových modelů . . . 5

3 Logistická regrese **6**
3.1 Základní popis . . . 6
3.2 Význam parametrů . . . 6
3.3 Odhad parametrů . . . 7
3.4 Test statistické významnosti parametrů . . . 9

4 Diskriminační míry **11**
4.1 Diskriminační míry . . . 11
4.1.1 Lorenzova křivka . . . 11
4.1.2 Giniho koeficient . . . 11
4.1.3 Kolmogorova-Smirnovova statistika . . . 14
4.1.4 ROC křivka . . . 14

5 Aplikace na data **16**
5.1 Popis dat . . . 16
5.2 Tvorba modelu . . . 16
5.2.1 Příklad . . . 18
5.3 Aplikace diskriminačních měr . . . 19
5.3.1 Lorenzova křivka . . . 20
5.3.2 C-statistika . . . 20
5.3.3 ROC křivka . . . 21
5.3.4 Kolmogorov-Smirnovova statistika . . . 21
5.4 Shrnutí modelu . . . 22

Závěr **23**

Literatura **24**

Tabulky **25**

Úvod

Nedílnou součásti bankovní činnosti je poskytování úvěrů a s nimi spojené řízení rizik, zvláště pak řízení kreditního rizika.

Problému měření kreditního rizika se věnuje hodně analytiků. Význam kreditního rizika se v posledních desetiletích značně změnil. Z pasivního procesu se stal strategickým nástrojem v bankovní činnosti. Banky vycítily, že bankovní operace mají vliv na ekonomické prostředí, ale taky, že samy na něm závisí. Ekonomické prostředí se zpravidla pojí se zvýšeným rizikem, především kreditním rizikem. Na druhou stranu vytváří rovněž prostor pro zisk.

Credit scoring je základním nástrojem pro moderní řízení kreditního rizika. Napomáhá tomu hlavně značný vývoj informačních technologií. Scoring se využívá nejenom při poskytování úvěru, ale i ve strategiích týkajících se budoucího řízení kreditního rizika nebo ve strategiích spojených s vymáháním pohledávek.

Po začlenění skóringového modelu v bance začne proces validace, tj. ověření, zda model správně rozhoduje o poskytnutí úvěru. Špatná rozhodnutí mohou být dvojího druhu: odmítnutím žádosti o úvěr dobrému klientovi banka přichází o zisk a poskytnutím úvěru špatnému klientovi naopak způsobí bance ztrátu. Správné ohodnocení klienta umožní bance určit i vhodnou úrokovou sazbu.

V první kapitole této práce objasníme pojem rizika a jeho dělení. Dále uvedeme definici defaultu a přiblížíme význam kreditního rizika ve finančním sektoru. V druhé kapitole popíšeme základní dělení credit scoringu a vysvětlíme podstatu skóringových modelů.

Ve třetí kapitole popíšeme model logistické regrese, objasníme význam jednotlivých parametrů a následně se budeme zabývat jejich odhadem a statistickou významností.

Čtvrtá kapitola se zaměřuje na ohodnocení kvality skóringového modelu. Popíšeme nejběžnější diskriminační míry a také si ukážeme, čím se řídit při volbě tzv. cut-off bodu, který definujeme ve 4. kapitole, a podle kterého budeme rozdělovat klienty do dvou skupin: na dobré a na špatné klienty.

V páté kapitole teoretické poznatky aplikujeme na reálná data a znázorníme schopnost diverzifikace našeho modelu Lorenzovou a ROC křivkou.

1. Definice rizika

Pro účely této musíme práce nejdříve objasnit pojem rizika. Slova riziko se údajně poprvé použilo v 17. století v námořnictví, kde označovalo úskalí, kterému se museli plavci vyhnout. Od té doby se význam rizika značně rozšířil a znamená např.:

1.pravděpodobnost či možnost vzniku ztráty, obecně nezdaru.
2.variabilitu možných výsledků nebo nejistotu jejich dosažení.
3.odchýlení skutečných od očekávaných výsledků.

Jelikož neexistuje všeobecné vymezení definice rizika, budeme dále riziko považovat za míru nebo ohodnocení ohrožení či nebezpečí vycházejícího z výskytu možných událostí na nás nezávislých anebo pramenícího z následku nějakého rozhodnutí.

Všeobecně se rizika klasifikují podle zdroje, odkud nejistota pochází. V podnikatelské sféře můžeme i nefinanční ztrátu vyčíslit v penězích, takže ve finále je zpravidla každé riziko nositelem finančního efektu a podle kontextu úvahy může být tedy pojmenováno jako finanční riziko. Podle knihy Mejstříka a kol. Teplý Petr and Magda [2008] můžeme finanční riziko rozdělit na rizika:

- kreditní - riziko ztráty vyplývající ze selhání smluvní strany tím, že nedostojí svým závazkům podle podmínek smlouvy, na základě které se banka/firma stala věřitelem smluvní strany. Např. u bank se podílí na přibližně 50-70% všech bankovních rizik, a proto tvoří nejvýznamnější složku v řízení finančního rizika. Podrobněji toto riziko rozebereme v následující podkapitole.
- tržní - riziko ztráty nebo nejistota budoucích zisků vyplývající ze změn cen, kurzů a sazeb na finančních trzích.
- tikvidní - riziko, že banka/firma ztratí schopnost dostát svým finančním závazkům v době, kdy se stanou splatnými nebo nebude schopna financovat svá aktiva.

1.1 Kreditní riziko

V bankovním sektoru představuje úvěr finanční aktivum, které vyplývá z doručení peněz nebo jiného aktiva od věřitele k dlužníkovi za určitých podmínek splácení dluhu, ke kterým patří výška splátek, úrok a doba splatnosti. Úvěrové riziko tak v bance znamená default klienta v úvěrovém vztahu s bankou a vyjadřuje nejistotu spojenou s očekávanými výnosy. Zpravidla se za default považuje situace, kdy klient nesplácí úvěr alespoň 90 dní po vypršení doby splatnosti. Vyšší podstupované riziko banky je zpravidla kompenzováno vyšším výnosem z poskytnutého úvěru.

Kreditní riziko je nejběžnějším rizikem, se kterými se finanční instituce musí vypořádat. I z tohoto důvodu byly sepsány normy, podle nichž se musí řídit fi-

nanční instituce podléhající licenčnímu řízení nebo obchodované na specifických finančních trzích. Pro banky se stal základním kamenem těchto norem Basilejská dohoda BASEL II.(viz Anna [2012])
Hlavním cílem měření kreditního rizika je vyčíslení potencionálních ztrát vzniklých při úvěrových obchodech banky nebo finanční instituce. Koncept Basel II stanovuje 3 metody měření kreditního rizika(rizikové váhy):

- standardizovaný přístup(STA-Standardised approach)
- základní IRB přístup(FIRB-Foundation Internal Ratings-Based Approach)
- pokročilý IRB přístup(AIRB-Advanced Internal Ratings-Based Approach)

Při standardním přístupu zjištění rizikové váhy se používají výhradně ratingy uznávaných ratingových agentur.
Přístupy založené na interním ratingu dovolují bankám, které jsou schopny statisticky změřit příslušné riziko určitého financování, aby svou kapitálovou vybavenost upravovaly adekvátně dle svého individuálního rizika. Použití vlastních interních hodnocení musí být schváleno národním regulátorem. Modely používané při IRB přístupu jsou založeny na rovnici

$$EL = PD * LGD * EAD.$$

EL značí očekávané ztráty, LGD ztrátu v případě selhání, EAD vystavenou hodnotu aktiv vůči selhání a PD znamená pravděpodobnost selhání v průběhu roku od uzavření smlouvy. PD je hlavní složka této rovnice, LGD je dáno národním regulátorem a EAD určíme z rozvahy banky.
V této práci se zaměříme na modely zabývajícími se vypočtením PD.

2. Credit scoring

2.1 Základní dělení

Credit scoring je souhrn metod sloužících k ohodnocení žadatele o úvěr. V credit skóringu se pokoušíme rozdělit klienty do dvou tříd: na "dobré"klienty, tedy spolehlivé, kteří nejspíš splatí úvěr, a na ty "špatné", u kterých je vysoká pravděpodobnost defaultu. Skóringové modely můžeme rozdělit do tří skupin:

- **expertní**
  Tento model vzniká jen na základě rozhodnutí experta. Jeho hlavní výhodou je, že nejsou ke vzniku potřebná data. Podle určité soustavy otázek budou klienti rozděleni do různých tříd.

- **etatistický**
  Vzniká na základě statistického zpracování dat o minulém chování klientů.

- **hybridní**
  Jsou to statistické skóringové modely upravené ratingovým analytikem.

Dále můžeme skóringové modely rozdělit na aplikační a behaviorální. U aplikačního skóringu vycházíme z dat, která získáme od klienta při žádosti o úvěr, kdežto u behaviorálního modelu předpokládáme, že klient již má u instituce úvěrovou historii, a můžeme využít i charakteristiky popisující průběh splacení.

2.2 Podstata skóringových modelů

Usilujeme o to, aby vyšší hodnoty odpovídaly spolehlivým a nižší nespolehlivým klientům. Ke konstrukci skóringové funkce se používají např. klasifikační stromy, neuronové sítě nebo logistická regrese.
V této práci se zaměříme pouze na *logistický model*. Ten je jedním z nejběžnějších modelů používaných pro modelování binární proměnné a je založen na předpokladu, že vysvětlující proměnné násobené příslušnými koeficienty mají lineární vztah vzhledem k přirozenému logaritmu četnosti případů splacení úvěru a defaultu(viz Teplý Petr [2007]).

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{n} \boldsymbol{\beta}_i \boldsymbol{x}_i,$$

kde p je pravděpodobnost splacení úvěru, β jsou příslušné koeficienty a $\boldsymbol{x}$ nezávislé proměnné (více ve 3. kapitole).
Z této rovnice pak snadno odvodíme pravděpodobnost splacení úvěru p:

$$p = \frac{\exp\{\beta_0 + \sum_{i=1}^{n} \boldsymbol{\beta}_i \boldsymbol{x}_i\}}{1 + \exp\{\beta_0 + \sum_{i=1}^{n} \boldsymbol{\beta}_i \boldsymbol{x}_i\}}.$$

3. Logistická regrese

3.1 Základní popis

Logistická regrese je v praxi nejpoužívanější metoda pro konstrukci skóringové funkce. Je to statistická metoda, se kterou se snažíme určit závislost binární náhodné veličiny (závislá proměnná) na jiných skutečnostech, u kterých si myslíme, že mají vliv na zkoumanou veličinu. Tyto skutečnosti (nezávislé proměnné, regresory) mohou být prezentovány pomoci binárních, kategoriálních nebo spojitých proměnných. Zkoumanou veličinu budeme nazývat závislou proměnnou (odezva, vysvětlovaná proměnná), pod kterou si můžeme představit např. pravděpodobnost výskytu nějakého jevu.

Matematicky zápis Označme Y jako závislou binární proměnnou, $\boldsymbol{\beta}$ jako vektor parametrů, které chceme odhadnout a $\boldsymbol{x}_i$ je vektor regresorů. Pak lze jednorozměrný model logistické regrese zapsat následujícím způsobem:

$$E(Y \mid x) = \pi(x), \text{ kde}$$

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

znamená pravděpodobnost Y nabytí hodnoty 1 při daných nezávislých proměnných. Lineární regrese je založená na modelu $y = E(Y \mid x) + \epsilon$, kde ϵ je chyba a vyjadřuje odchylku od podmíněné střední hodnoty. Obvykle se předpokládá, že ϵ má normální rozdělení se střední hodnotou rovnající se nule a nějakým rozptylem, a tedy závislá proměnná má normální rozdělení se střední hodnotou nula a konstantním rozptylem. V našem případě, kde odezva podmíněna x se rovná $y = \pi(x) + \epsilon$ nabývá dvě různé hodnoty. Pokud se $y = 1$, potom $\epsilon = 1 - \pi(x)$ s pravděpodobností $\pi(x)$, pokud $y = 0$, pak $\epsilon = -\pi(x)$ s pravděpodobností $1 - \pi(x)$. Z výše uvedeného vyplývá, že náš model má binomické rozdělení s parametrem $\pi(x)$.

3.2 Význam parametrů

Pro lepší pochopení modelu potřebujeme určit relaci mezi vysvětlujícími a vysvětlovanou proměnnou a následně definovat jednotku přírůstku v nezávislých proměnných.

Transformace $g(x)$ nazýváme logit a využijeme ji v popisech v další kapitole, ve které si pro názornost přiblížíme, jak se dájí získat parametry pro jednorozměrný model.

V lineární regresi odhadnuté parametry znamenaly, že pokud zvýšíme regresor xb_i o jedna při fixních ostatních regresorech, pak se změní hodnota Y právě o β_i. Proto nejdříve sestavíme funkci vysvětlované proměnné jako lineární funkci vysvětlujících proměnných. Tato funkce se obecně v anglické literatuře nazývá *link function* a speciálně pro náš model se jmenuje logit a je tvaru:

$$g(\pi(\boldsymbol{x})) = \ln\left[\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right] = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + ... + \boldsymbol{\beta}_n x_n.$$

Z toho dostaneme β_i rozdílem

$$g(\pi(x_1,...,x_i+1,...,x_n) - g(\pi(x_1,...,x_i,...,x_n) = \beta_i.$$

Tento rozdíl můžeme poté přepsat do tvaru:

$$g(\pi(x_1,...,x_i+1,...,x_n) - g(\pi(x_1,...,x_i,...,x_n) = \ln\left(\frac{\frac{\pi(x_1,...,x_i+1,...,x_n)}{1-\pi(x_1,...,x_i+1,...,x_n)}}{\frac{\pi(x_1,...,x_i,...,x_n)}{1-\pi(x_1,...,x_i,...,x_n)}}\right),$$

a tedy

$$\frac{\frac{\pi(x_1,...,x_i+1,...,x_n)}{1-\pi(x_1,...,x_i+1,...,x_n)}}{\frac{\pi(x_1,...,x_i,...,x_n)}{1-\pi(x_1,...,x_i,...,x_n)}} = \exp(\beta_i).$$

Výše uvedený podíl v literatuře najdeme pod pojmem *odds ratio*. Je to podíl šancí a udává nám o kolik je pravděpodobnější výskyt jevu, jestliže zvýšíme hodnotu regresoru x_ij o jedna. Atribut x_ij můžeme porovnávat v různých rozmezích, podle potřeby výzkumu. Např. pokud se nezávislá proměnná pohybuje v rozmezí od 0 po 1, pak bude realističtější brát v úvahu menší přírůstky než 1.

3.3 Odhad parametrů

Hodnoty $\boldsymbol{x}$ jsou známy, a tedy musíme pouze najít co nejlepší odhad parametru $\boldsymbol{\beta}$. Odhad parametru $\boldsymbol{\beta}$ v modelu logistické regrese se provádí metodou maximální věrohodnosti. Princip této metody spočívá v sestavení a následně nalezení maxima věrohodnostní funkce. Tato funkce je sdružená distribuční funkce vysvětlujících proměnných ($f\,(Y_1,Y_2,...,Y_n \mid \mathbb{X})$), jinými slovy nám udává pravděpodobnost,s jakou nastanou námi pozorované jevy při daných regresorech. V této sekci si rozebereme odhad parametru $\boldsymbol{\beta}$ pouze pro jednorozměrný případ. S předpokladem nezávislosti pozorovaných dat, kdy jev Y_i nastává s pravděpodobností $\pi(\boldsymbol{x})$, je věrohodnostní funkce ve tvaru:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\,(1-\pi(x_i))^{1-y_i}\,. \tag{3.1}$$

Pro hledání extrémů položíme parciální derivace podle β_0 a β_1 rovny nule:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0. \tag{3.2}$$

Pro jednodušší derivování zlogaritmujeme $\log(L(\boldsymbol{\beta})) = \ell(\boldsymbol{\beta})$. Touto transformací se ale nemění bod, ve kterém $L(\boldsymbol{\beta})$ nabývá svého maxima. Tímto dostaneme

$$\ell(\pi) = \log\left(\prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}\right) = \sum_{i=1}^{n} y_i \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) + \sum_{i=1}^{n} \log\left(1-\pi(x_i)\right)$$

$$= \sum_{i=1}^{n} y_i g(x_i) + \sum_{i=1}^{n} \log\left(1-\pi(x_i)\right). \tag{3.3}$$

Platí následující vztahy:

$$\frac{\partial}{\partial g(x_i)} \log\left(1-\pi(x_i)\right) = -\frac{\partial}{\partial g(x_i)} \log\left(1+\exp^{g(x_i)}\right) = -\frac{\exp^{g(x_i)}}{1+\exp^{g(x_i)}} = -\pi(x_i)$$

$$\frac{\partial g(x_i)}{\partial \beta_0} = 1, \frac{\partial g(x_i)}{\partial \beta_1} = \boldsymbol{x}_i.$$

Pak po zderivování rovnice 3.3 podle β_0 máme

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left(y_i - \pi(x_i)\right),$$

a podle β_1

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left(y_i - \pi(x_i)\right) x_i.$$

Zbývá nám dokázat, že vyřešením 3.2 rovnic získáme maximum v $\boldsymbol{\beta}$.
Druhou derivací podle β_0 dostáváme

$$\frac{\partial \ell}{\partial \beta_0^2} = -\sum_{i=1}^{n} \left(\pi(x_i)(1-\pi(x_i))\right)$$

a pro β_1

$$\frac{\partial \ell}{\partial \beta_1^2} = -\sum_{i=1}^{n} \left(\pi(x_i)(1-\pi(x_i))\right) x_i^2,$$

a dále pro

$$\frac{\partial \ell}{\partial \beta_1 \beta_0} = \frac{\partial \ell}{\partial \beta_0 \beta_1} = -\sum_{i=1}^{n} \left(\pi(x_i)(1-\pi(x_i))\right) x_i.$$

Z toho vidíme, že Hessián je negativně definitní nebo negativně semidefinitní, a tedy věrohodnostní funkce je konkávní. Rovnici 3.2 pak vyřeší příslušný statistický software. Podobným způsobem získáme odhad parametrů i u vícerozměrného modelu. Uvažujeme rovnici

$$\ln\left[\ell(\beta)\right] = \sum_{i=1}^{n} y_i \ln\left[\pi(x_i)\right] + \left(1-y_i\right) \ln\left[1-\pi(x_i)\right].$$

U takového modelu ale nemusí mít všechny nezávislé proměnné vliv na zkoumaný jev, a proto se provádí *test statistické významnosti parametrů*.

3.4 Test statistické významnosti parametrů

S vypočtenými parametry modelu nás bude zajímat, zda všechny parametry obsažené v modelu mají vliv na odezvu modelu. V této podkapitole se nebudeme zabývat tím, zda se zvolenými nezávislými proměnnými výsledný model přesněji vystihuje data než s jinou množinou regresorů, ale zda se bude výrazně lišit odezva.(viz Lemeshow a kol.)
Jeden ze způsobů jak změřit významnost daného parametru je podobný jako v lineární regresi s využitím *residuálního součtů čtverců* neboli *SSE*. Tento součet se dá vyjádřit vztahem

$$SSE = \sum_{i=1}^{n} (y_i - y_i^2).$$

Provedeme porovnání SSE původního modelu s modelem, ve kterém vynecháme některé nezávislé proměnné. Pokud bude rozdíl velký, potom jsou dané nezávislé proměnné statisticky významné.

V logistické regresi budeme místo SSE porovnávat logaritmickou věrohodnostní funkci saturovaného modelu a modelu s vynecháním některých proměnných. *Saturovaný model* je takový model, ve kterém jsou obsaženy všechny proměnné z námi zkoumaných dat. Porovnání saturovaného modelu a modelu s vynecháním některých proměnných se dá zapsat výrazem, jenž označíme D (z angl. deviance):

$$D = -2ln \left[\frac{\ell(\widehat{\boldsymbol{\beta}})}{(\ell_S(\boldsymbol{\beta}_S))} \right].$$

Poměr $\dfrac{\ell(\widehat{\boldsymbol{\beta}})}{(\ell_S(\boldsymbol{\beta}_S))}$ se nazývá *věrohodnostní poměr*. Přidáním mínus dvojky jsme získali výraz, pro nějž známe distribuční funkci, a tedy jej můžeme použít pro testování hypotéz. Jelikož známe tvary jejich věrohodnostních funkcí, můžeme D zapsat jako:

$$D = -2 \sum_{i=1}^{n} \left[y_i ln \left(\frac{\widehat{\pi}_i(x_i)}{y_i} \right) + (1 - y_i) ln \left(\frac{1 - \widehat{\pi}_i(x_i)}{1 - y_i} \right) \right].$$

Výše uvedenou rovnici můžeme dále upravit do tvaru $D = -2ln\ell(\beta)$, protože saturovaný model vystihuje s pravděpodobností 1 pozorovaná data, tedy $\ell_S = \prod_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{(1-y_i)} = 1$.
Pro zjištění statistické významnosti l námi vybraných nezávislých proměnných provedeme porovnání D původního modelu se všemi regresory s D bez daných l nezávislých proměnných. S využitím předchozího máme:

$$\begin{aligned} G &= D(\text{model se všemi proměnnými}) - D(\text{model s vynecháním l proměnných}) \\ &= -2ln \left[\frac{\ell(\widehat{\boldsymbol{\beta}}_l)}{\ell(\widehat{\boldsymbol{\beta}})} \right], \end{aligned} \tag{3.4}$$

kde G se v literatuře najdeme pod pojmem standardní odchylka.

Budeme testovat hypotézu H_0, zda se koeficienty u námi l vybraných nezávislých proměnných rovnají nule oproti alternativě H_1. Pokud platí hypotéza H_0, pak G má Chí-kvadrat rozdělení o l stupních volnosti. Test na hladině α hypotézy H_0 zamítáme, pokud $G > \chi^2_l(1-\alpha)$.

Dalším způsobem testování statistické významnosti parametrů je Waldův test, který je založen na asymptotické normalitě odhadnutých parametrů:

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

Uvažujeme hypotézu H_0, že $\beta_i = 0$ oproti alternativě H_1, že $\beta_i \neq 0$ a sestrojíme Waldovu statistiku:

$$W = \frac{\beta_i}{\widehat{\sigma}(\widehat{\beta}_i)} \sim N(0,1) \text{ za platnosti H0.}$$

Pokud je $|W| > (1-\alpha/2)$ kvantil normálního rozdělení, potom zamítáme hypotézu H_0.

4. Diskriminační míry

4.1 Diskriminační míry

Za míru diskriminace v kreditním riziku považujeme schopnost modelu oddělit "špatné"klienty od těch "dobrých". Snahou je vytvořit takový model, v němž bychom byli schopni zvolit skóre s_0 tak, že hodnoty nižší než s_0 budou mít všichni špatní klienti a hodnoty vyšší než s_0 naopak klienti dobří. V praxi toho samozřejmě nelze dosáhnout. Skóre s_0 se v literatuře popisuje jako *cut-off bod* neboli prahový bod.
Míry diskriminace lze použít jenom na modely, které přidělují skóre všem klientům, a tudíž ji nemůžeme použít např. na modelech používající klasifikační stromy nebo neuronové sítě.

4.1.1 Lorenzova křivka

Lorenzova křivka je nejpopulárnějším nástrojem pro zobrazení diverzifikace. Pro použití Lorenzovy křivky na skóringový model musíme znát distribuční funkce dobrých a špatných klientů.
Mějme skóringovou funkci s:$\boldsymbol{X} \rightarrow$ [0,1]. Pravděpodobnost, že náhodně vybraný klient bude mít nižší skóre než zvolené $a \in$ [0,1] nazveme distribuční funkci náhodné veličiny s($\boldsymbol{X}$).
Pro konstrukci Lorenzovy křivky dále rozdělíme data do dvou skupin a k nim označíme podmíněné distribuční funkce F_G-splatili úvěr a F_B-nesplatili úvěr, pak pro tyto dvě funkce definujeme množinu $LC(s) = \{[F_B(a),F_G(a)] \in [0,1]^2, a \in [0,1]\}$ jako Lorenzovu křivku. V bezchybné skóringonvé funkci by LC vedla z bodu [0,0] k [1,0], poté k [1,1], pro každou jinou rozumnou s($\boldsymbol{X}$) je $F_B(a) \geqslant F_G(a)$. Lc ležící na diagonále by pak znamenala, že naše skóringová funkce nemá žádnou rozlišovací schopnost.
V praxi se pro získání F_B a F_G používá konzistentní odhad pro distribuční funkce, který vypadá následovně:

$$\widehat{F}_B(a) = \sum_{i=1}^{n_B} \frac{I[s(\boldsymbol{X}_B) \leqslant a]}{n_B}$$

$$\widehat{F}_G(a) = \sum_{i=1}^{n_G} \frac{I[s(\boldsymbol{X}_G) \leqslant a]}{n_G},$$

kde $\boldsymbol{X}_B$ jsou data špatných klientů a $\boldsymbol{X}_G$ těch dobrých, n_B, n_G jejich počet.

4.1.2 Giniho koeficient

Lorenzova křivka nám graficky znázorňuje diverzifikaci skóringové funkce, ale abychom s ní mohli pracovat, budeme ji muset kvantifikovat. K tomu slouží Giniho koeficient. Pro distribuční funkce F_B a F_G jej definujeme jako:

$$GC = 2 \int_0^1 |F_G(a) - F_B(a)| dF_G(a).$$

Pokud známe plochu A mezi Lorenzovou křivkou a diagonálou a plochu B pod Lorenzovou křivkou, pak můžeme definici GC přepsat do tvaru

$$GC = \frac{A}{A+B}.$$

Slovy, jedná se o poměr plochy mezi Lorenzovou křivkou a diagonálou a celkovou plochou pod diagonálou. Jelikož Lorenzova křivka je vykreslena v jednotkovém čtverci, můžeme odvodit následující vztahy:

$$A + B = \frac{1}{2} \Rightarrow GC = 2A = 1 - 2B. \tag{4.1}$$

Z výše uvedeného vyplývá, že Giniho koeficient udává hodnoty mezi [−1,1], kde 1 znamená, že skóringová funkce rozdělila klienty do dvou skupin, kdežto nulová hodnota poukazuje na totožnost distribučních funkcí F_G a F_B, a tedy náš model postrádá rozlišovací schopnost. Záporná hodnota značí opačnou rozlišovací schopnost.

Jeden z možných způsobů(viz Říha Samuel [2012]) odhadu Giniho koeficientu je odhadnout plochu pod Lorenzovou křivkou pomocí obdélníků. Každý z těchto obdélníků bude mít šířku $\frac{1}{n_1}$ a výšku $\frac{C_i}{n_0}$, kde C_i je počet klientů, kteří nesplatili úvěr počínaje i-tým klientem. Seřadíme je vzestupně podle jejich skóre a budeme předpokládat, že žádnému z nich nebylo přiřazeno stejné skóre. Pak

$$B = \sum_i^n Y_i \frac{1}{n_1} \frac{C_i}{n_0}, \tag{4.2}$$

kde

$$C_i = \sum_j^n (1 - Y_j).$$

Po dosazení a úpravě rovnice 4.2 dostáváme

$$
\begin{aligned}
B &= \sum_{i}^{n} Y_i \frac{1}{n_1} \frac{\sum_{i}^{n}(1-Y_i)}{n_0} \\
&= \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{j=i}^{n} Y_i (1-Y_j) \\
&= \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{j=i}^{n} Y_i - \sum_{i=1}^{n} \sum_{j=i}^{n} Y_i Y_j \\
&= \frac{1}{n_0 n_1} \sum_{i=1}^{n} (n+1-i) Y_i - \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} \sum_{j>i}^{n} Y_i Y_j \\
&= \frac{1}{n_0 n_1} \left[n_1(n+1) - \sum_{i=1}^{n} i Y_i - n_1 - \binom{n_1}{2} \right] \\
&= \frac{1}{n_0 n_1} \left[n_1(n+1) - \sum_{i=1}^{n} i Y_i - n_1 - \frac{n_1(n_1-1)}{2} \right]
\end{aligned}
$$

Dosazením do rovnice 4.1 dostaneme

$$
\begin{aligned}
GC &= 1 - 2\frac{1}{n_0 n_1} \left[n_1(n+1) - \sum_{i=1}^{n} i Y_i - n_1 - \frac{n_1(n_1-1)}{2} \right] \\
&= \frac{1}{n_0 n_1} \left[-n_1 n - n_1 - 2 \sum_{i=1}^{n} i Y_i \right] \\
&= \frac{2}{n_0 n_1} \sum_{i=1}^{n} i Y_i - \frac{n+1}{n_0}
\end{aligned}
$$

Pokud bychom neměli seřazené klienty podle jejich skóre, potom bychom dostali GC jako

$$
GC = \frac{2}{n_0 n_1} \sum_{i=1}^{n} R_i Y_i - \frac{n+1}{n_0}, \tag{4.3}
$$

kde R_i značí pořadí skóre klientů. Z toho vyplývá, že Giniho koeficient je závislý pouze na pořadí jejich skóre.

C-statistika

Dalším ukazatelem míry diverzifikace je c-statistika(viz Karel [2004]), která je úzce spjata z Giniho koeficientem jak později ukážeme.
C-statistiku definujeme jako

$$
c(s) = P\left[s(\boldsymbol{X}_B i) \leqslant s(\boldsymbol{X}_G j) \mid Y_i = 0, Y_j = 1 \right] \tag{4.4}
$$

Je to pravděpodobnost, že klient, který splatil úvěr, bude mít vyšší skóre než ten, který úvěr nesplatil. Pro jeho odhad použijeme výpočet

$$
c(s) = \frac{\sum_{s_j \in S_1} \sum_{s_i \in S_0} I_(s_j > s_i)}{n_0 n_1},
$$

kde S_1 je množina všech skóre klientů, kteří splatili úvěr a S_2, kteří nesplatili úvěr. Dalšími úpravami získáme vyjádřeni c-statistiky pomocí Giniho koeficientu.

$$c(s) = \frac{\sum_{s_j \in S_1} \sum_{s_i \in S_0} I_(s_j > s_i)}{n_0 n_1} = \frac{\sum_{s_j \in S_1} (n_0 - C_j)}{n_0 n_1} = \frac{\sum_{i=1}^{n} Y_i (n_0 - C_i)}{n_0 n_1}$$
$$= 1 - \frac{\sum_{i=1}^{n} Y_i C_i}{n_0 n_1} = 1 - \sum_{i=1}^{n} Y_i \frac{1}{n_1} \frac{C_i}{n_0} = 1 - B = \frac{1}{2}(1 + GC)$$

Pokud by nastala situace, že jednomu klientu přiřadí skóringová funkce stejné skóre, pak identifikátor $I_{(s_j > s_i)}$, pro $s_j \in S_1$, $s_i \in S_0$ vynásobíme $\frac{1}{2}$, pro C_i přičteme $\frac{1}{2}$.

4.1.3 Kolmogorova-Smirnovova statistika

Další z používaných diskriminačních měr je Kolmogorova-Smirnovova statistika nebo taky supremální kritérium. Můžeme ji popsat pomocí výše zmíněných distribučních funkcí.

$$KS = \sup_{s \in R} |F_G(s) - F_B(s)| \tag{4.5}$$

Jedná se o maximální vertikální vzdálenost mezi distribučními funkcemi dobrých a špatných klientů. Pokud bychom vykreslili jejich hustoty, potom bod s, ve kterém KS nabývá svého suprema, bude v bodě, ve kterém se tyto hustoty protínají. V tomto bodě bude pravděpodobně skóringový model správně rozdělovat na dobré a špatné klienty.

4.1.4 ROC křivka

ROC křivka

Jednou z nejpoužívanějších metod k určení predikční schopnosti modelu je ROC křivka (Receiver Operating Characteristic Curve). Ta podobně jako Lorenzova křivka vyjadřuje vztah mezi distribučními funkcemi dobrých a špatných klientů. (viz Zbyněk [2008]) ROC křivku definujeme pomocí pojmů sensitivita a specificita modelu:

- *sensitivita* udává pravděpodobnost, že dobrý klient bude klasifikován jako dobrý (označíme jako S_e).
- *specificita* udává pravděpodobnost, že špatný klient bude klasifikován jako špatný (označíme jako S_p).

Křivku ROC pak tvoří množina všech možných dvojic (S_e,$1 - S_p$) pro různé hodnoty s. 1-specificita vyjadřuje podíl špatných klientů, které náš model označil jako dobré.
V perfektním modelu by ROC křivka měla tvar funkce $y = 1$. Pokud by ROC křivka kopírovala diagonálu, pak to znamená, že náš model má nulovou rozlišovací schopnost. Při porovnání několika modelů pak pro dané skóre vybereme ten, pro který je ROC křivka položena nejvýše.

AUC (Area under the ROC curve)

Podobně jako u Lorenzovy křivky se pro kvantifikaci diskriminace používá Giniho koeficient, u ROC křivky se používá AUC neboli plocha pod ROC křivkou. Přesný vztah mezi AUC a Giniho koeficientem můžeme popsat rovnicí:

$$GI = 2AUC - 1.$$

AUC můžeme interpretovat jako pravděpodobnost, že náhodně vybraný nevhodný klient dopadl v testu hůře než náhodně vybraný dobrý klient, a tedy je rovna c-statistice.

5. Aplikace na data

Našim cílem bude vytvořit model pro odhadování podmíněné pravděpodobnosti splacení úvěru klientem v závislosti na získaných údajích. Tuto pravděpodobnost dále budeme považovat za skóre klienta. Poté na výsledný model aplikujeme diskriminační míry uvedené v teoretické části.

5.1 Popis dat

Data jsme získali z internetových stránek `http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29`. Data obsahují informace o tisíci klientech jedné německé banky. Pro každého klienta máme 20 charakteristik, z toho 7 numerických a 13 kategoriálních. Bližší popis dat nalezneme v tabulkách 5.3.

5.2 Tvorba modelu

K tvorbě modelu použijeme statistický software R. Pro výpočet parametrů metodou maximální věrohodnosti a výběr proměnných využijeme zabudované funkce v R. Pro všechny ostatní úkony provedeme vlastní výpočet.
Nejdříve vytvoříme model se všemi proměnnými, jehož odhadnuté parametry uvádíme v tabulkách 5.4. Pro lepší přehlednost uvedeme původní i finální model v logitovém tvaru. Původní model je ve tvaru:

$$\begin{aligned} g(\pi(\boldsymbol{x})) = \ln\left[\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right] &= \beta_0 + \boldsymbol{\beta}_1 * Ucet + \beta_2 * Splatnost + \boldsymbol{\beta}_3 * Moralka \\ &+ \boldsymbol{\beta}_4 * Ucel + \beta 5 * Objem + \boldsymbol{\beta}_6 * Uspory + \boldsymbol{\beta}_7 * Soucazam \\ &+ \beta_8 * Podil + \boldsymbol{\beta}_9 * Stav + \boldsymbol{\beta}_{10} * Ruceni + \beta_{11} * DBydleni \\ &+ \boldsymbol{\beta}_{12} * StruktMajetku + \beta_{13} * Vek + \boldsymbol{\beta}_{14} * Uvery + \boldsymbol{\beta}_{15} * Bydleni \\ &+ \beta_{16} * PocetUveru + \boldsymbol{\beta}_{17} * Zamestnani + \boldsymbol{\beta}_{18} * Vyzivovani \\ &+ \beta_{19} * Telefon + \beta_{20} * Cizinec. \end{aligned} \tag{5.1}$$

U kategoriálních proměnných vystupuje parametr β_i jako vektor. Pro vysvětlení předvedeme jeden ze způsobů zakódování proměnné Ucet.

Ucet	**D1**	**D2**	**D3**
Ucet[T.A11]	0	0	0
Ucet[T.A12]	1	0	0
Ucet[T.A13]	0	1	0
Ucet[T.A14]	0	0	1

Tabulka 5.1: Kódování proměnné Ucet

Kategorie Ucet[T.A11] byla zvolena jako referenční kategorie. $D1$,$D2$ a $D3$ v literatuře najdeme pod pojmem *dummy proměnné*. Proměnná Ucet poté bude v logitu našeho modelu ve tvaru:

$$\beta_0 + \beta^1_{Ucet} * D1 + \beta^2_{Ucet} * D2 + \beta^3_{Ucet} * D3, \tag{5.2}$$

kde pro náš případ bude β_0 intercept celého modelu a $\beta^i_{promenna}$ značí i-tou složku parametru $\boldsymbol{\beta}_{promenna}$.

V dalším kroku vybereme proměnné metodou *Stepwise selection*, neboli krokový výběr česky. Její podstatou je postupné přidávání nebo odebírání proměnných podle jejich statistické významnosti. Tato metoda je podrobně popsána v knize Lemeshowa a kol.(Jr. and Stanley [2000]).

Touto metodou jsme získali model

$$\begin{aligned} g(\pi(\boldsymbol{x})) = \ln\left[\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right] &= \beta_0 + \beta^1_{Moralka} * Moralka[T.A31] \\ &+ \beta^2_{Moralka} * Moralka[T.A32] + \beta^3_{Moralka} * Moralka[T.A33] \\ &+ \beta^4_{Moralka} * Moralka[T.A34] + \beta^1_{Splatnost} * Splatnost \\ &+ \beta^1_{Ucet} * Ucet[T.A12] + \beta^2_{Ucet} * Ucet[T.A13] \\ &+ \beta^3_{Ucet} * Ucet[T.A14]. \end{aligned} \tag{5.3}$$

Odhady parametrů uvádíme v následující tabulce:

Koeficienty	**odhad**	**Standardní odchylka**	**W hodnota**	(**P**(> |**W**|)
Intercept	-0.219009	0.406640	-0.539	0.59017
Moralka[T.A31]	0.121639	0.474729	0.256	0.79777
Moralka[T.A32]	0.932781	0.371104	2.514	0.01195
Moralka[T.A33]	0.961880	0.433850	2.217	0.02662
Moralka[T.A34]	1.569489	0.393796	3.986	¿0.0001
Splatnost	-0.034479	0.006328	-5.448	¿0.0001
Ucet[T.A12]	0.514770	0.184817	2.785	0.00535
Ucet[T.A13]	1.091141	0.336022	3.247	0.00117
Ucet[T.A14]	1.898803	0.205785	9.227	¡0.0001

Tabulka 5.2: Odhad parametrů i s výsledkem Waldova testu

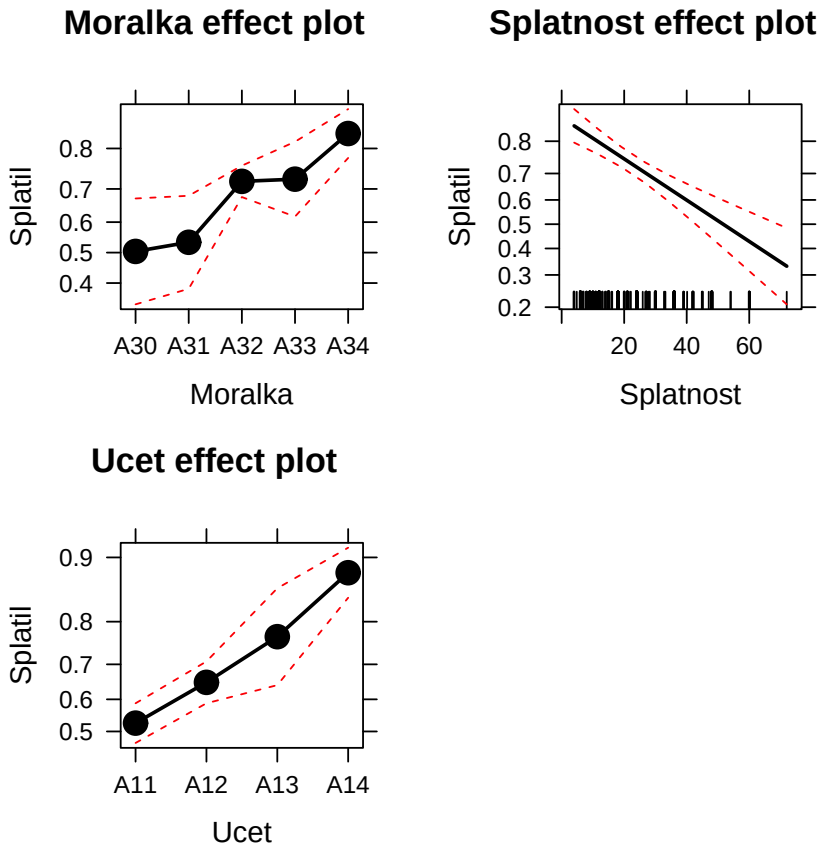
Jako skóre i-tého klienta budeme považovat hodnotu $\hat{\pi}(\boldsymbol{x}_i)$, která představuje odhad podmíněné pravděpodobnosti splacení úvěru.

V grafu 5.1 pak vidíme výstupy modelu pro jednotlivé kategorie proměnných Moralka, Splatnost a Ucet. Na ose Splatil je znázorněn odhad pravděpodobnosti splacení úvěru podle kategorie nebo hodnoty dané proměnné, a tedy můžeme pozorovat změnu výstupu modelu při změně kategorií nebo hodnoty dané proměnné.

První z trojice grafů má jinou tendenci jakou bychom logicky mohli usuzovat. Z grafu vyplývá, že klient, který má potíže se splácením jiného úvěru, má větší

šanci získat úvěr než klient s bezúhonnou platební morálkou.

Z dalšího grafu můžeme vyčíst, že s rostoucí dobou splatnosti úvěru klesají žadateli šance získání úvěru. Poslední graf se vyvíjí dle očekávání, a tedy klient s vyšší sumou na účtu má větší pravděpodobnost splacení úvěru. Nejvyšší šance splacení úvěru má pak osoba, která si u dané banky nevede žádný účet.



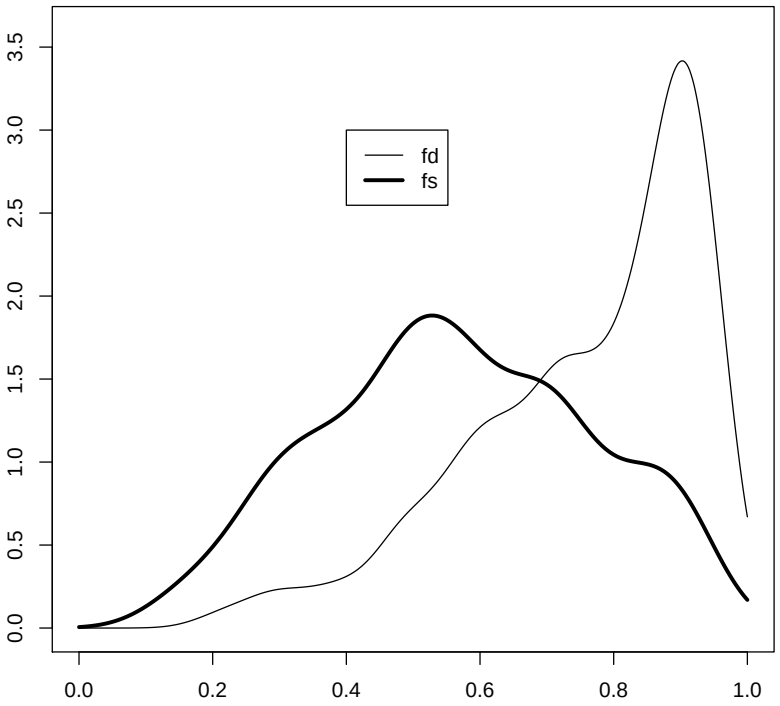
Obrázek 5.1: Grafy pro jednotlivé proměnné při zafixovaných ostatních proměnných

V grafu 5.2 máme porovnání odhadnutých hustot skóre špatných(fs) a dobrých(fd) klientů. Pro jejich odhad jsme použili jádrový odhad. Bod jejich protnutí bude taky výsledkem Kolmogorovy- Smirnovovy statistiky, který spočteme v další podkapitole.

5.2.1 Příklad

Pro lepší pochopení modelu si zde uvedeme jednoduchý příklad. Předpokládejme studenta ve věku 20 let, který si vzal půjčku na školu, kterou splácí bez žádných zpoždění a žádá další půjčku na ojeté auto ve výši $5000DM$ s dobou splatností 72 měsíců. Půjčku bude splácet z výdělku z brigády. Splátky tvoří 25%. Dále předpokládáme, že stále bydlí u rodičů a nemá žádné výdaje spojené s ubytováním. Kromě $100DM$ na účtu a mobilního telefonu nemá žádný jiný majetek.

Výše uvedené charakteristiky vložíme do našeho modelu.



Obrázek 5.2: Hustota skóre špatných a dobrých klientů

$$\begin{aligned} g(\pi(\boldsymbol{x})) = \ln\left[\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right] &= \beta_0 + \beta_1 * Moralka[T.A31] \\ &+ \beta_2 * Moralka[T.A32] + \beta_3 * Moralka[T.A33] \\ &+ \beta_4 * Moralka[T.A34] + \beta_5 * Splatnost \\ &+ \beta_6 * Ucet[T.A12] + \beta_7 * Ucet[T.A13] \\ &+ \beta_8 * Ucet[T.A14] \\ &= -0.219009 + 0.932781 * 1 - (0.034479 * 72) + 0.51477 \end{aligned} \tag{5.4}$$

Převedeme vzorec pro vyjádření odhadu podmíněné pravděpodobností:

$$\hat{\pi}(\boldsymbol{x}) = \frac{\exp\{-0.219009 + 0.932781 * 1 - (0.034479 * 72) + 0.51477\}}{1 + \exp\{-0.219009 + 0.932781 * 1 - (0.034479 * 72) + 0.51477\}} = 0.22. \tag{5.5}$$

S takovým skóre student má velmi malé šance získat úvěr.

5.3 Aplikace diskriminačních měr

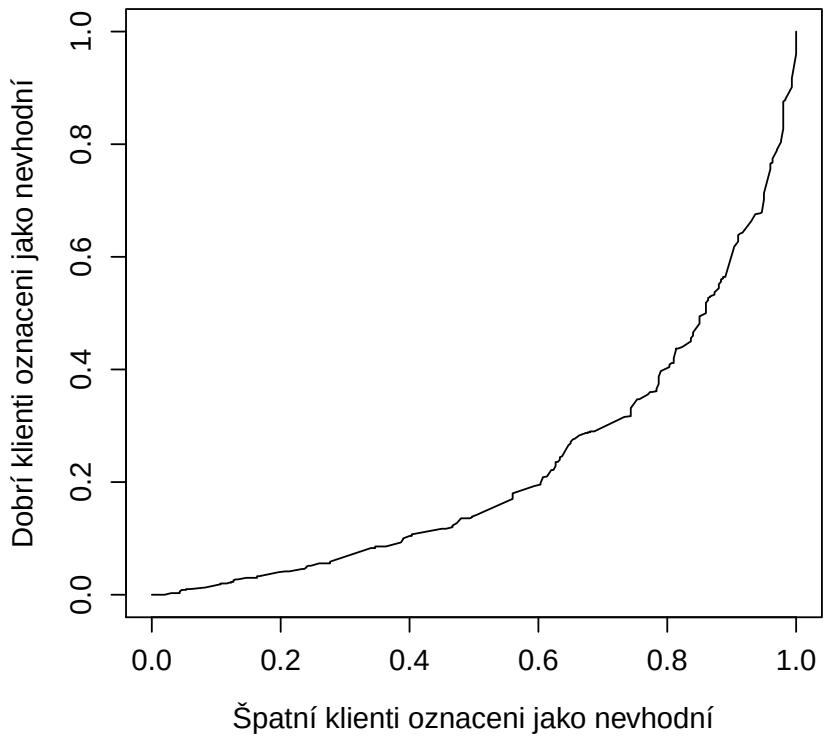
V této podkapitole změříme diverzifikační schopnost modelu.

5.3.1 Lorenzova křivka

Lorenzovu křivku jsme sestrojili pomocí odhadnutých podmíněných distribučních funkcí skóre dobrých a špatných klientů. Jednotlivé hodnoty skóre budeme považovat jako cut-off body. Poté klienty, pro které náš model spočítal nižší skóre než daný cut-off bod, budeme považovat jako nespolehlivé. V grafu 5.3 pak vidíme, že Lorenzova křivka je jenom mírně pod diagonálou. Podle uvedeného postupu 4.3 spočteme *Giniho koeficient*. Tento odhad je ale pouze přibližný, protože náš model nesplňuje předpoklad prostého zobrazení. Z dat vygeneroval pouze 229 různých hodnot skóre na 1000 klientů.

$$GI \approx 0.5416476 \tag{5.6}$$

Laicky můžeme říci, že se náš model alespoň z poloviny blíží modelu ideálnímu.



Obrázek 5.3: Lorenzova křivka

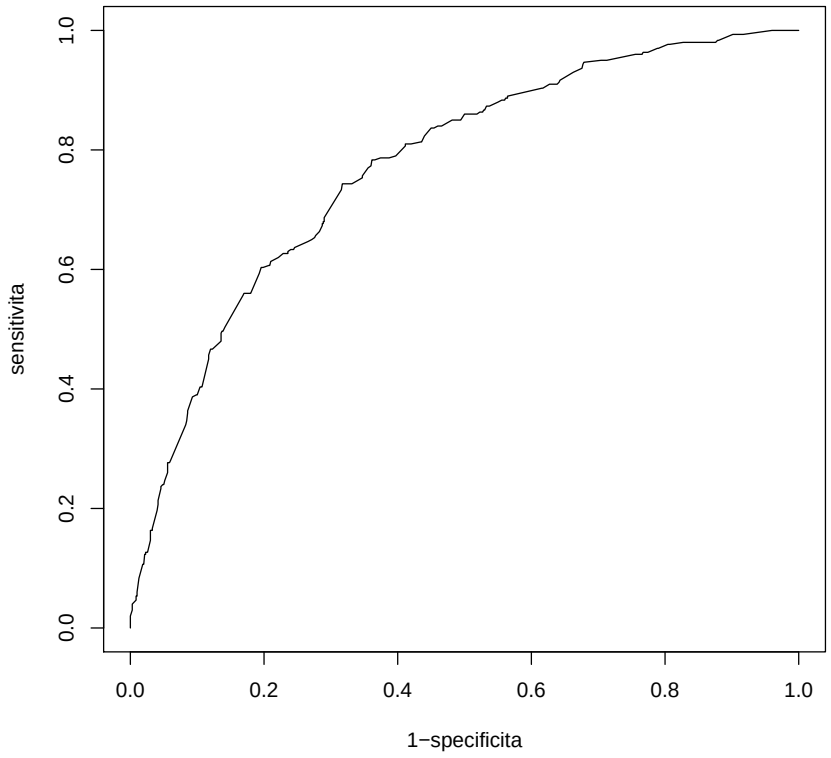
5.3.2 C-statistika

Dalším zmíněným ukazatelem je *c-statistika*, která má jednoznačný vztah k *Giniho koeficientu* a udává nám pravděpodobnost, že dobrý klient bude mít vyšší skóre něž špatný klient. Při výpočtu budeme postupovat podle rovnice 4.4.

$$c(s) = 0.7760714 \tag{5.7}$$

5.3.3 ROC křivka

Stejně jako Lorenzovu křivku, jsme i ROC křivku sestrojili pomocí empirických distribučních funkcí skóre dobrých a špatných klientů. Znázorňuje poměr počtu špatně přiřazených dobrých klientů k počtu správně přiřazených špatných klientů pro jednotlivé hodnoty cut-off bodu. Plochu pod křivkou nám kvantifikuje zmíněná c-statistika. Stejně jako Lorenzova se i ROC křivka dost liší od ideálního tvaru.



Obrázek 5.4: ROC křivka

5.3.4 Kolmogorov-Smirnovova statistika

V posledním kroku určíme Kolmogorov-Smirnovovu statistiku. Výpočtem 4.5 jsme získali hodnotu

$$KS = \sup_{s \in R} |F_G(s) - F_B(s)| = 0.697. \tag{5.8}$$

Pro tuto hodnotu nám test hypotézy H_0, tj. rovnosti distribučních funkcí špatných a dobrých klientů, na hladině 0.05 dává p hodnotu rovnou nule, a tedy hypotézu H_0 zamítáme.

5.4 Shrnutí modelu

Výsledné hodnoty jednotlivých diskriminačních měr svědčí o tom, že náš model není příliš efektivní. Např. odhadnutá hodnota Giniho koeficientu je 0.5416476, kdežto v bankovní praxi se ustálil názor, že hodnota kolem 0.7 se považuje za uspokojivou. Jako cut-off bod bychom mohli považovat výsledek Kolmogorov-Smirnovovy statistiky, která se rovná 0.697. Podle tohoto bodu by náš model správně označil 68.3% dobrých klientů a 74.3% špatných klientů. Zajímavé je, že jsme našim postupem selekce proměnných považovali výši poskytovaného úvěru jako statisticky nevýznamnou charakteristiku.

Závěr

V této práci jsme se nejdříve zabývali kreditním rizikem a credit scoringem. Ukázali jsme postup konstrukce logistického modelu a následně jsme se zabývali metodami pro určení jeho diverzifikační schopnosti. V praktické části jsme se věnovali hlavně jednotlivým mírám diskriminace. Zprvu jsme na reálných datech zkonstruovali logistický model. Poté jsme jeho validační schopnost znázornili Lorenzovou křivkou a kvantifikovali ji Giniho koeficientem. Jako další způsob vizualizace diverzifikační schopnosti jsme sestrojili ROC křivku. K té jsme uvedli c-statistiku jakožto její číselný ukazatel. Na konci jsme uvedli výsledek Kolmogorov-Smirnovova testu. Všechny tyto míry diskriminace poukazují na to, že by se náš model v praxi neosvědčil.

Cílem práce bylo srozumitelně popsat matematický základ potřebný k sestrojení skóringových modelů a následně jej na reálných datech zkonstruovat. Čtenář tak po přečtení práce bude mít konkrétní představu o výstavbě skóringového modelu založeného na logistické regresi.

Literatura

Bluhm Christian; Overbeck Ludger; and Christoph Wagner. *Introduction to Credit Risk Modeling*. Chapman and Hall, 2nd edition edition, 2010.

Bortlíček Zbyněk. *ROC křivky*. Diplomová práce, Masarykova univerzita v Brně, 2008.

Hosmer David W. Jr. and Lemeshow Stanley. *Applied logistic regression*. John Wiley and Sons, Inc., 2000.

Jakubík Petr Teplý Petr. Skóring jako indikátor finanční stability. `http://www.cnb.cz/miranda2/export/sites/www.cnb.cz/cs/financni_stabilita/zpravy_fs/fs_2007/FS_2007_clanek_2.pdf`, 2007. červenec 2013.

Komorád Karel. *Statistické metody klasifikace a jejich využití pro kreditní skórování*. Diplomová práce, Univerzita Karlova v Praze, 2004.

Matuszyk Anna. *Credit scoring*. 1.Wydanie. CeDeWu Sp.z.o.o., Warszawa, 2012.

Mejstřík Michal Teplý Petr and Pečená Magda. *Základní principy bankovnictví*. Karolinum, 2008.

Polak Roman. *Finanční rizika leasingové společnosti*. PhD thesis, VŠB-Technická univerzita Ostrava, 2008.

Rychnovský Michal. *Postupná výstavba modelů kreditního rizika*. Bakalářská práce, Univerzita Karlova v Praze, 2008.

Říha Samuel. *Maximalizace Giniho koeficientu v binární logistické regresi*. Bakalářská práce, Univerzita Karlova v Praze, 2012.

Řezáč Martin and Řezáč František. How to Measure the Quality of Credit Scoring Models. *Czech Journal of Economics and Finance*, 61(5):486–507, 2011.

Tabulky

Proměnná	**Popis**
Ucet	množství peněz na účtu(DM)
A11	žádné nebo debet
A12	méně než 200
A13	více než 200
A14	nemá účet
Splatnost	doba do splatnosti(měsíce)
Moralka	splácení předchozích úvěrů
A30	žádné předchozí úvěry
A31	splacené úvěry
A32	současně splácené úvěry
A33	váhavé splácení
A34	úvěry u jiných bank
Ucel	účel půjčky
A40	nový automobil
A41	ojetý automobil
A42	nábytek
A43	rádio nebo televize
A44	zařízení bytu
A45	opravy
A46	vzdělání
A47	dovolená
A48	rekvalifikace
A49	obchod
A410	jiný
Objem	výše půjčky(DM)
Uspory	výše úspor/cenných papírů(DM)
A61	méně než 100
A62	$100 - 500$
A63	$500 - 1000$
A64	více než 1000
A65	žádné nebo nezjištěno
Souczam	doba současného zaměstnání(roky)
A71	nezaměstnaný
A72	méně než 1
A73	$1 - 4$
A74	$4 - 7$
A75	více než 7
Podil	poměr výše splátek ku příjmu

Proměnná	**Popis**
Stav	pohlaví a rodinný stav
A91	Muž, rozvedený
A92	Žena, vdaná nebo rozvedená
A93	Muž, svobodný
A94	Muž, ženatý nebo vdovec
A95	Žena, svobodná
Ruceni	způsob ručení
A101	žádné
A102	spolužadatel
A103	ručitel
DBydleni	doba bydlení v současné domácnosti
StruktMajetku	ve vlastnictví
A121	nemovitosti
A122	stavební spoření
A123	životní pojištění
A124	automobil nebo jiný
A125	neznámo nebo nemá
Vek	věk(roky)
Uvery	další úvěry
A141	v jiných bankách
A142	v obchodech
A143	žádné
Bydleni	typ bydlení
A151	zdarma
A152	byt v nájmu
A153	vlastní byt
PocetUveru	počet úvěrů
Zamestnani	zaměstnání
A171	nezaměstnaný
A172	nevyučený
A173	kvalifikovaný
A174	vedoucí pracovník
Vyzivovani	počet vyživovaných
Telefon	telefon
A191	ano
A192	ne
Cizinec	pracující cizinec
A201	ano
A202	ne
Splatil	splacení úvěru
splacen	1
nesplacen	0

Tabulka 5.3: Popis proměnných

Koeficienty	odhad	Standardní odchylka	z hodnota	(P(> |z|)
(Intercept)	-4.005e-01	1.084e+00	-0.369	0.711869
Bydleni[T.A152]	4.436e-01	2.347e-01	1.890	0.058715
Bydleni[T.A153]	6.839e-01	4.770e-01	1.434	0.151657
Cizinec[T.A202]	1.392e+00	6.258e-01	2.225	0.026095
DBydleni	-4.776e-03	8.641e-02	-0.055	0.955920
Moralka[T.A31]	-1.434e-01	5.489e-01	-0.261	0.793921
Moralka[T.A32]	5.861e-01	4.305e-01	1.362	0.173348
Moralka[T.A33]	8.532e-01	4.717e-01	1.809	0.070470
Moralka[T.A34]	1.436e+00	4.399e-01	3.264	0.001099
Objem	-1.283e-04	4.444e-05	-2.887	0.003894
PocetUveru	-2.721e-01	1.895e-01	-1.436	0.151109
Podil	-3.301e-01	8.828e-02	-3.739	0.000185
Ruceni[T.A102]	-4.360e-01	4.101e-01	-1.063	0.287700
Ruceni[T.A103]	9.786e-01	4.243e-01	2.307	0.021072
Souczam[T.A72]	6.691e-02	4.270e-01	0.157	0.875475
Souczam[T.A73]	1.828e-01	4.105e-01	0.445	0.656049
Souczam[T.A74]	8.310e-01	4.455e-01	1.866	0.062110
Souczam[T.A75]	2.766e-01	4.134e-01	0.669	0.503410
Splatnost	-2.786e-02	9.296e-03	-2.997	0.002724
Stav[T.A92]	2.755e-01	3.865e-01	0.713	0.476040
Stav[T.A93]	8.161e-01	3.799e-01	2.148	0.031718
Stav[T.A94]	3.671e-01	4.537e-01	0.809	0.418448
StruktMajetku[T.A122]	-2.814e-01	2.534e-01	-1.111	0.266630
StruktMajetku[T.A123]	-1.945e-01	2.360e-01	-0.824	0.409743
StruktMajetku[T.A124]	-7.304e-01	4.245e-01	-1.721	0.085308
Telefon[T.A192]	3.000e-01	2.013e-01	1.491	0.136060
Ucel[T.A41]	1.666e+00	3.743e-01	4.452	¡0.0001
Ucel[T.A410]	1.489e+00	7.764e-01	1.918	0.055163
Ucel[T.A42]	7.916e-01	2.610e-01	3.033	0.002421
Ucel[T.A43]	8.916e-01	2.471e-01	3.609	0.000308
Ucel[T.A44]	5.228e-01	7.623e-01	0.686	0.492831
Ucel[T.A45]	2.164e-01	5.500e-01	0.393	0.694000
Ucel[T.A46]	-3.628e-02	3.965e-01	-0.092	0.927082
Ucel[T.A48]	2.059e+00	1.212e+00	1.699	0.089297
Ucel[T.A49]	7.401e-01	3.339e-01	2.216	0.026668
Ucet[T.A12]	3.749e-01	2.179e-01	1.720	0.085400
Ucet[T.A13]	9.657e-01	3.692e-01	2.616	0.008905
Ucet[T.A14]	1.712e+00	2.322e-01	7.373	¿0.0001
Uspory[T.A62]	3.577e-01	2.861e-01	1.250	0.211130
Uspory[T.A63]	3.761e-01	4.011e-01	0.938	0.348476
Uspory[T.A64]	1.339e+00	5.249e-01	2.551	0.010729
Uspory[T.A65]	9.467e-01	2.625e-01	3.607	0.000310
Uvery[T.A142]	1.232e-01	4.119e-01	0.299	0.764878
Uvery[T.A143]	6.463e-01	2.391e-01	2.703	0.006871
Vek	1.454e-02	9.222e-03	1.576	0.114982
Vyzivovani	-2.647e-01	2.492e-01	-1.062	0.288249
Zamestnani[T.A172]	-5.361e-01	6.796e-01	-0.789	0.430160
Zamestnani[T.A173]	-5.547e-01	6.549e-01	-0.847	0.397015
Zamestnani[T.A174]	-4.795e-01	6.623e-01	-0.724	0.469086