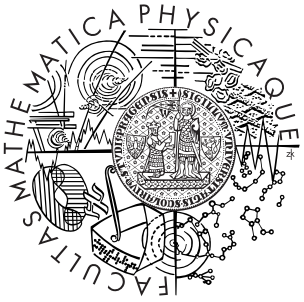
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Aleh Masaila

Vliv zamítání klientů na predikční schopnost skóringových modelů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Jaroslav Ševčík

Studijní program: Matematika, Obecná matematika

2009

Chtěl bych poděkovat vedoucímu Mgr.Jaroslavu Ševčíkovi za zajímavé téma a pomoc při tvorbě této 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 7. srpna 2009 Aleh Masaila

Obsah

Název práce: Vliv zamítání klientů na predikční schopnost skóringových modelů
Autor: Aleh Masaila
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Jaroslav Ševčík
e-mail vedoucího: sevcik@karlin.mff.cuni.cz

Abstrakt: Skóringové modely dělí žadatele o úvěr na přijaté a zamítnuté klienty. Pokud chceme vyvinout nový nebo upravit stávající model, pak informace o splacení či nesplacení úvěru máme jenom pro přijaté klienty. Pokud ale nový model bude stavěn jenom na těchto klientech, pak neznalost úvěruschopnosti zamítnutých klientech může způsobit, že postavený model nebude přesný. *Reject inference* je označení pro proces, kdy se snažíme využít údaje o zamítnutých klientech za účelem vylepšit skóringový model. Cílem této práce je vysvětlit problém reject inference, popsat několik známých způsobů řešení a na reálných datech ověřit vliv zamítnutých klientů na predikční schopnosti skóringových modelů.

Klíčová slova: reject inference, chybějící data, odchylka výběru

Title: An analysis of the prediction capability of scoring models
Author: Aleh Masaila
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Jaroslav Ševčík
Supervisor's e-mail address: sevcik@karlin.mff.cuni.cz

Abstract: Scoring models separate applicants for credit into accepted and rejected clients. If we want to design a new or adjust an existing model, we only have information about repayment or non-repayment of credit on the accepted clients. However, if the new model is only built on these clients, the ignorance of the credit-worthiness of the rejected clients can cause that this model will not be correct. The reject inference is a term for a process during which we try to include the data of the rejected clients to improve a scoring model. The aim of this work is to explain the problem of the reject inference. We will also describe several known ways of solution and verify the influence of the rejected clients on the prediction capability of the scoring models using real data.

Keywords: reject inference, missing data, sample bias

Kapitola 1

Úvod

Slovo *úvěr* se stalo nedílnou součástí našeho slovníku. Co je vlastně úvěr? Běžně se definuje jako půjčka, která slouží k uspokojení finančních potřeb fyzických nebo právnických osob. Obvykle jako věřitel v této situaci vystupují bankovní instituce, ale není to podmínkou. Tím, že banky poskytují úvěr, se vystavují *úvěrovému riziku*, což je riziko, že protistrana nebude schopná splnit své závazky. Ve většině případů to znamená, že klient, který obdržel úvěr, ho nebude schopný splácet.

Přirozenou snahou bank je omezit toto riziko neboli poskytovat úvěr jenom spolehlivým klientům, u kterých nehrozí riziko *defaultu*[1]. Z těchto důvodu byl vyvinut *kreditní skóring*.

Pojem kreditní skóring popisuje statistické metody, které pomáhají věřitelům odhadovat rizikovost žadatelů o úvěr neboli určit pravděpodobnost defaultu. Tato pravděpodobnost je stanovena na základě skóre, od toho také název skóring. Klientovo skóre se dá zjednodušeně popsat jako body, které dostal za údaje, které o sobě věřitelovi poskytl. Tyto údaje se liší podle typu klienta. Pro fyzické osoby to jsou sociodemografické údaje, jako například příjem, vzdělání nebo věk. Pro právnické osoby to jsou finanční ukazatele, jako ukazatel likvidity nebo zadluženosti.

Proces, který se odehrává od žádosti potenciálního dlužníka po splacení, respektive nesplacení úvěru, se může charakterizovat dvěma otázkami.

1. Byl úvěr poskytnut?
2. Byl úvěr splacen?

Odpověď na první otázku získáváme díky znalostí údajů a chování klientů, kterým byl úvěr v minulosti poskytnut. Na základě těchto informací je vybu-

[1] budeme se držet zažitého označení, a proto budeme používat pojem *default* místo českého ekvivalentu *selhání*

dován *skóringový model*. Někdy je skóringový model označován jako skóringová funkce. Tento model je pak aplikován na i-tého zájemce o úvěr. Výsledkem je skóre daného klienta, a plynoucí ze skóre pravděpodobnost defaultu. Samotné rozhodnutí o poskytnutí úvěru obdržíme, když porovnáme dosažené skóre S_i, s námi stanovenou prahovou hodnotou c.

Zadefinujeme $A \in \{0, 1\}$ jako proměnnou, která označuje přijetí nebo zamítnutí žadatele.

$$A = \begin{cases} 1 & \text{pokud } S_i \geq c \text{ (úvěr je poskytnut, neboli klient je } \textit{přijatý}) \\ 0 & S_i < c \text{ (úvěr není poskytnut, neboli klient je } \textit{zamítnutý}) \end{cases}$$

Populaci tak můžeme rozdělit na dvě skupiny, $(A = 0)$ a $(A = 1)$. Pro přijaté klienty zadefinujeme proměnnou $Y \in \{0, 1\}$, předpisem

$$Y = \begin{cases} 1 & \text{takový klient je označován jako } \textit{dobrý} \\ 0 & \text{tento klient je } \textit{špatný} \end{cases}$$

Nebudeme blíže specifikovat, jaký přesně klient je označován za dobrého a jaký za špatného. Toto rozdělení na dobré a špatné se liší dle typu úvěru, toho, co chceme zkoumat, nebo podle vnitřních předpisů banky. Nemusí to ale vždy nutně znamenat, že klient v jistém okamžiku selhal, a dále už úvěr nesplácí. Za špatného klienta se tak například může považovat ten, kdo nezaplatí splátku po dobu tří měsíců. Ale tento detail není pro naší další činnost důležitý, proto se jim nebudeme zabývat.

Samotný skóringový model se dá vytvořit pomocí řady statistických metod, například rozhodovací stromy, diskriminační analýza nebo neutronové sítě. Nejpoužívanější metodou je pak metoda *logistické regrese*. Výhodou logistické regrese je její poměrně snadná interpretace. Pomocí ní budou vybudované skóringové modely, které se objeví v této práci. Blíže logistickou regresi popíšeme ve 2. kapitole.

Pro výstavbu nového modelu nebo úpravu stávajicího modelu vycházíme z údajů o splacení/nesplacení pouze těch klientů, kteří byli přijatí. Toto může představovat problém, neboť klient, který byl zamítnut, není automaticky špatný klient. Stejně jako mezi přijatými žadateli se vyskytují špatní klienti, tak i mezi zamítnutými jsou klienti dobří. A byli bychom rádi, kdyby náš nový model dokázal odhalit i tyto dobré klienty. Na otázku, jak nám mohou s tímto problémem pomoci zamítnutí klienti, se snaží odpovědět proces *reject inference*.

Účelem této práce je popsat reject inference a jak to souvisí s vychýlením modelu. Dále rozebereme několik konkrétních technik, podíváme se na způsob, jak pracují, a na jejich případné nedostatky. To bude obsahem 3. kapitoly.

Na problém reject inference se dá také dívat jako na problém *chybějících dat*. Proto ve 4. kapitole se budeme věnovat vztahu mezi reject inference a chybějících dat. Uvedeme několik druhů analýzy chybějících dat, především ty, které se

dají použít při stanovení skutečné úvěruschopnosti zamítnutých klientů a jejích následném využití.

Praktická část této práce je obsažena v 5. kapitole. V ní navrhneme několik způsobů, jak můžeme prozkoumat vliv zamítnutých klientů na skóringové modely. Následně s využitím statistického softwaru a reálných dat to prakticky otestujeme a rozebereme výsledky.

Naším cílem nebude podrobně prozkoumat nějakou určitou část reject inference, ani udělat rozsáhlou simulační studii. Cílem je obecně přiblížit tento proces a vytvořit jeho základní přehled.

Kapitola 2

Logistická regrese

Tato část práce vychází především z práce Hosmer, Lemeshow (2000), a bakalářské práce Rychnovský(2008). Vice informace o logistické regresi nebo ohledně diverzifikačních schopnostech skóringových modelů lze proto hledat tam.

2.1 Model logistické regrese

Logistická regrese se používá v případech, kdy potřebujeme popsat vztah mezi nezávislými proměnnými a závislou proměnnou. V našem případě roli nezávislých vysvětlujících proměnných hrají individuální charakteristiky klienta, které označíme jako $\boldsymbol{X} = (X_1, \ldots, X_k)$. Jak již bylo řečeno v úvodu, vstupními daty mohou být různé sociodemografické ukazatelé klienta, jako například věk nebo příjem. Výstupní hodnotou $Y_{\boldsymbol{X}}$ je pak informace, zda klient je dobrý nebo špatný.

Zajímá nás pravděpodobnost, zda klient s konkrétními charakteristikami $\boldsymbol{x}$ je dobrý. Označme tuto pravděpodobnost jako $\pi(\boldsymbol{x}) = \mathrm{P}(Y_{\boldsymbol{x}} = 1) = P(Y = 1|\boldsymbol{X})$.

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

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

Pokud bychom chtěli použít pro výpočet této pravděpodobnosti lineární regresi, tak bychom museli vyřešit tuto rovnici:

$$\pi(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = \beta_0 + \boldsymbol{\beta x}$$

Ale výstupní hodnota řešící tuto rovnici by nemusela být mezi 0 a 1, což je nutné, neboť nás zajímá pravděpodobnost $\pi(\boldsymbol{x})$. Proto, abychom se dostali k potřebnému modelu, definujeme nejprve funkci *odds*, neboli *šance*.

$$odds(\boldsymbol{x}) = \frac{\mathrm{P}(Y_{\boldsymbol{x}} = 1)}{\mathrm{P}(Y_{\boldsymbol{x}} = 0)} = \frac{\mathrm{P}(Y_{\boldsymbol{x}} = 1)}{1 - \mathrm{P}(Y_{\boldsymbol{x}} = 1)} = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}$$

Nyní provedeme logistickou transformaci:

$$logit(\boldsymbol{x}) = \ln odds(\boldsymbol{x}) = \ln \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}$$

Hodnoty $logit(\boldsymbol{x})$ již nabývají hodnot z intervalu $(-\infty, \infty)$, a proto tuto hodnotu můžeme vyjádřit pomocí rovnice:

$$logit(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = \beta_0 + \boldsymbol{\beta x}$$

Zpět k *odds* se dostaneme pomocí exponenciální funkce.

$$\exp(\beta_0 + \boldsymbol{\beta x}) = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}$$

Z této rovnice již snadno dostaneme potřebný model:

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

2.2 Odhad parametrů

Uvažujeme, že máme n pozorování, $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$. Chceme nejlépe odhadnout parametry modelu. Tyto parametry $\boldsymbol{\beta}$ se odhadují *metodou maximální věrohodnosti*. Máme:

$$\mathrm{P}(Y_{\boldsymbol{x}_i} = y_i) = \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i} \tag{2.1}$$

kde y_i označuje, zda i-tý klient je dobrý ($y_i = 1$), nebo špatný ($y_i = 0$). Protože předpokládáme, že jednotlivé případy jsou nezávislé, můžeme psát

$$l(\beta) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i} \tag{2.2}$$

Abychom si zjednodušili počítání, tak tuto rovnici zlogaritmujeme.

$$\mathrm{L}(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} y_i \ln[\pi(\boldsymbol{x}_i)] + (1 - y_i) \ln[1 - \pi(\boldsymbol{x}_i)] \tag{2.3}$$

Metoda maximální věrohodnosti spočívá v tom, že pro každé β_i najdeme jeho maximální hodnotu. Toho dosáhneme tak, že zderivujeme dle jednotlivých parametrů a výsledné rovnice položíme rovno nule.

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(\boldsymbol{x}_i)] = 0 \tag{2.4}$$

a

$$\frac{\partial L(\beta)}{\partial \beta_i} = \sum_{i=1}^{n} \boldsymbol{x}_i [y_i - \pi(\boldsymbol{x}_i)] = 0, \text{ pro } i = 1, \ldots, n \tag{2.5}$$

2.3 Diverzifikační schopnosti skóringových modelů

Když mluvíme o *diverzifikačních schopnostech* modelu, pak máme na mysli jeho schopnost oddělit dobré klienty od špatných. Tato schopnost modelu je v podstatě jeho nejdůležitější vlastností. Existuje mnoho způsobů, pomocí kterých můžeme zjistit, zda námi sestavený model má pro nás dostatečné diverzifikační schopnosti. Jsou to například Kolmogorov-Smirnovův test, Lorenzova křivka, Giniho koeficient, ROC křivka, AUC a další.

Stručně popíšeme ROC křivku, z důvodů její poměrně snadné interpretace a všeobecné rozšířenosti.

ROC křivka

ROC křivka *(Reciever Operating Charasteristic)* je nástroj užívaný pro hodnocení a grafické znázornění chování klasifikačních pravidel při klasifikaci objektů do dvou tříd. Uplatnění se nachází v mnoha oborech počínaje medicínou a konče bankovnictvím. Abychom ji přesně popsali, potřebujeme si vysvětlit několik důležitých pojmů.

V úvodu máme definované *rozhodovací pravidlo* neboli *skóringovou funkci*, která přiřadí klientovi s charakteristikami $\boldsymbol{x}$ skóre S. Dále jsme si řekli, co to je *prahová hodnota* c, veličina Y, která určovala, zda se z žadatele stal dobrý nebo špatný klient, a A, která označovala, zda klient obdržel úvěr. Nyní pomocí následujících předpisů zavedeme pojmy *senzitivita*(Se) a *specificita*(Sp):

$$Se(c) = P(S \geq c|Y = 1) = P(A = 1|Y = 1) \tag{2.6}$$

$$Sp(c) = P(S < c|Y = 0) = P(A = 0|Y = 0) \tag{2.7}$$

Senzitivita nám udává, s jakou pravděpodobností je klient označený za dobrého, je ve skutečnosti dobrý, a specificita, s jakou pravděpodobností klient označený za špatného, je skutečně špatný. Empirické odhady těchto veličin se počítají z tzv. *matice záměn*, vyjádřenou kontingenční tabulkou.

	$Y = 1$	$Y = 0$
$A = 1$	TP	FP
$A = 0$	FN	TN

Tabulka 2.1: Matice záměn

TP - klient je správně označen za dobrého (*True Positives*)

FP – klient je nesprávně označen za dobrého, ve skutečnosti je špatný (*False Positives*)

FN - klient je nesprávně označen za špatného, ve skutečnosti je (*False Negatives*)

TN - klient je správně označen za špatného (*True Negatives*)

Rovnice (2.6),(2.7) se vyjádří jako

$$\widehat{Se}(c) = \frac{TP}{TP + FN} \tag{2.8}$$

$$\widehat{Sp}(c) = \frac{TN}{TN + FP} \tag{2.9}$$

V kreditním skóringu se, ale místo pojmu senzitivita a specificita používají *true positive rate*, TRP, a *false positive rate*, FRP. Vztah mezi nimi je následující

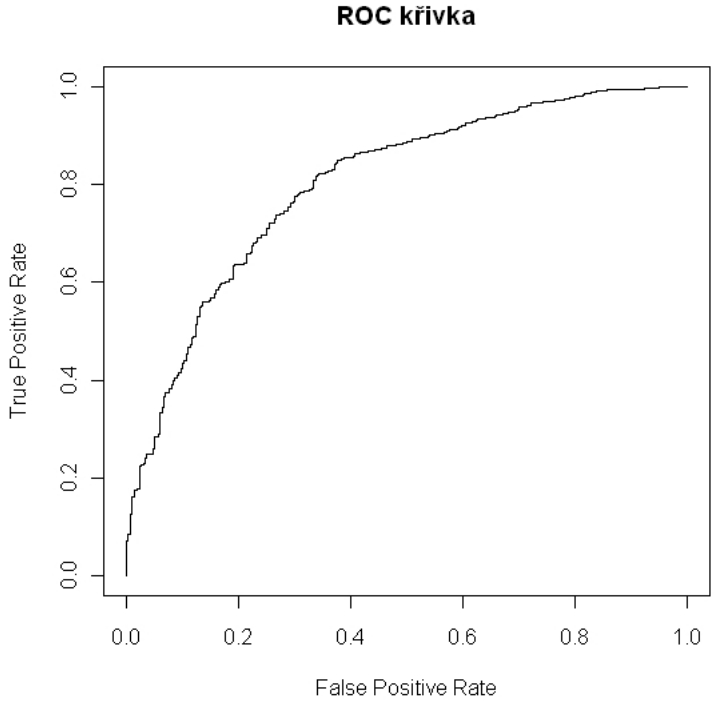
$$\text{TPR} = Se \tag{2.10}$$

$$\text{FPR} = 1 - Sp = P(S \geq c | Y = 0) \tag{2.11}$$

TPR je tedy pravděpodobnost, že klient je označen za dobrého, za podmínky, že je ve skutečnosti dobrý, kdežto FPR je pravděpodobnost, že klient je označen za dobrého, za podmínky, že je ve skutečnosti špatný. Empirické odhady TPR, FPR dostaneme tak, že za Se, resp. Sp dosadíme $\widehat{Se}$, resp $\widehat{Sp}$.

Při grafickém zobrazení ROC křivky a na vodorovnou osu dáváme FPR, a na svislou TPR. Křivka ROC je pak tvořena body

$$\{[\text{FPR}(c), \text{TPR}(c)], c \in \mathbb{R}\},$$



Model s perfektní diverzifikační schopností by měl ROC jdoucí z levého dolního rohu, dále do levého horního rohu a odtud do pravého horního rohu.

Statistika, která vyjádří vlastnosti ROC jedním číslem, je AUC(area under curve). AUC vyjadřuje, jaká je pravděpodobnost, že náhodně zvolený špatný klient bude mít nižší skóre, než náhodně zvolený dobrý klient. Místo AUC je často používán Giniho koeficient, což je taky číselná charakteristika diverzifikační schopností modelu. Zatímco AUC je plocha pod ROC křivkou, Gini je dvojnásobek orientované plochy mezi Lorenzovou křivkou a diagonálou jednotkového čtverce (viz. Rychnovský (2008)). Mezi AUC a Giniho koeficientem platí následující vztah Gini=2 · AUC - 1. AUC nabývá hodnot z intervalu $[0, 1]$, zatímco Gini z $[-1, 1]$. Ale pro praktické účely mají význam hodnoty AUC jenom z $\left[\frac{1}{2}, 1\right]$, resp. Gini z $[0, 1]$, protože AUC$< \frac{1}{2}$, respektive Gini< 0 znamenají, že náš model je horší než náhodné přidělování úvěru klientům. V tom případě bychom měli změnit definici dobrých a špatných klientů - dobré označit za špatné a obráceně. Touto změnou dostaneme model, který už je lepší než náhodný.

Kapitola 3

Reject inference

Jak již bylo naznačeno v úvodu, *reject inference* je označení pro proces, kdy se snažíme nějakým způsobem využít zamítnuté klienty, abychom zlepšili náš skóringový model.

Nastává otázka, zda bychom se měli o tyto klienty zajímat a zda si nevystačíme se znalostí přijatých klientů. Pokud by klienti, na jejichž výsledcích je postaven skóringový model, dostávali úvěr náhodně, pak by výběr z takových klientů věrně reprezentoval populaci. Proto všechny informace, které by byly potřeba k identifikaci profilu špatného klienta, by byly pro nás dostupné. V takovém případě bychom nepotřebovali zamítnuté klienty. Ale taková situace nenastává často, spíše naopak.

Proto máme opodstatněný zájem o chování zamítnutých klientů. Pokud získáme nějaké vědomosti o jejích chování, pak to můžeme využít pro stavbu nových a účinnějších skóringových modelů. Obyčejně totiž pro výstavbu nového modelu máme k dispozici dva druhy údajů: (a) individuální charakteristiky žadatelů o úvěr (tento vektor $\boldsymbol{X}$ je dostupný pro každého, jak pro zamítnuté, tak i pro přijaté), a (b) výstupní údaje těch klientů, kterým byl úvěr poskytnut.

Model postavený jenom na výsledcích přijatých klientů, přináší s sebou několik problémů. Největší z nich je ten, že při další aplikaci na populaci jako celek může dávat zkreslené výsledky. Pod pojmem zkreslené výsledky, nebo taky *vychýlení* modelu, rozumíme situaci, kdy model odhaduje nepřesně pravděpodobnost defaultu klienta. Vychýlení může vzniknout tak, že takový model nebude brát v úvahu některé proměnné, které jsou důležité pro popis chování náhodného žadatele z populace. Jako příklad si můžeme představit, že model, postavený na výsledcích přijatých, vypustí proměnnou, která zjišťovala druh bydlení žadatele. Udělá to proto, že z hlediska vysvětlení schopnosti klienta splatit úvěr, tato proměnná nebyla významná skupině. Nicméně v případě celé populace tato proměnná může být významná.

Všechny tyto skutečnosti vedou k tomu, že se zdá být užitečné zkoumat, zda by se ze zamítnutých stali dobří či špatní klienti. Kombinace reject inference a již známých výsledků přijatých klientů by mohla vést ke zlepšení stávajících skóringových modelů.

Dá se říci, že existují dva pohledy na problém reject inference. Jeden zkoumá tento problém přímo, vyšetřuje rozdělení klientů a techniky, které jsou zde uplatňované, jsou vyvinuté přímo pro potřeby kreditního skóringu a zamítnutých klientů. Ten druhý, který se objevuje později, nahlíží na reject inference obecněji, a to jako na *problém chybějících dat*. Táto kapitola se věnuje prvnímu přístupu. Vycházíme z práce Hand, Henley (1993). Autoři zde definuji problém reject inference, jak vzniká a taky popisují několik způsobů, jakými se tento problém může řešit.

3.1 Extrapolace z přijatých

Nechť $P(Y = 1|\boldsymbol{x})$ je pravděpodobnost, že klient s konkrétními charakteristikami $\boldsymbol{x}$ je dobrý. Pravděpodobnostní funkci označme jako $p(\boldsymbol{x}|Y = 1)$. Obdobné značení zavedeme i pro klienta, který je špatný.

Nejjednodušší přístup pro řešení problému reject inference je přímá extrapolace z prostoru přijatých do prostoru zamítnutých klientů. To znamená,že využíváme znalosti o charakteristikách a údajů o splacení u přijatých klientů a budujeme model na předpovědi jejich pravděpodobnosti splácení, který pak používáme i na zamítnutých klientech.

Tento postup ale nemusí být vždy úspěšný. Předně záleží na počtu zamítnutých klientů. Pokud tento počet bude příliš velký, pak z přijatých nezískáme dostatek informací, abychom vybudovali kvalitní model.

V případě extrapolace musíme rozlišovat dva základní přístupy, pomoci kterých odhadujeme $P(Y = 1|\boldsymbol{X})$. První odhaduje tuto pravděpodobnost přímo. Jedná se například o logistickou regresi, se kterou jsme se již seznámili. Druhý odhaduje hustoty rozdělení dobrých a špatných klientů $p(\boldsymbol{X}|Y = 1)$, $p(\boldsymbol{X}|Y = 0)$, a pak určí $P(Y = 1|\boldsymbol{X})$ pomocí Baysove věty.

$$P(Y = 1|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|Y = 1)P(Y = 1)}{p(\boldsymbol{X}|Y = 1)P(Y = 1) + p(\boldsymbol{X}|Y = 0)P(Y = 0)},$$

kde $P(Y = 1)$ a $P(Y = 0)$ je pravděpodobnost toho, že klient je dobrý, respektive špatný. Druhý způsob odhadu $P(Y = 1|\boldsymbol{X})$ používá např. diskriminační analýza.

Mezi těmito dvěma přístupy je rozdíl, který je pro nás důležitý. Hand, Henley (1993) tvrdí, že pokud pro stavbu nového modelu použijeme výběr z populace, a ne celou populaci, pak metody odhadující $P(Y = 1|\boldsymbol{X})$ pomocí odhadů $p(\boldsymbol{X}|Y = 1)$ a $p(\boldsymbol{X}|Y = 0)$, budou zdeformované. Pokusíme se tento efekt ukázat.

Nechť $T = \{T_0, T_1\}$ je sada dat, na které vyvíjíme model, $\boldsymbol{x}$ je konkrétní hodnota $\boldsymbol{X}$. T_0 označuje podmnožinu špatných klientů, T_1 dobrých. Obě dvě podmnožiny se následně použijí k odhadu podmíněných hustot $\hat{p}(\boldsymbol{x}|Y = i)$, $i = 0, 1$ a pravděpodobností $\hat{P}(Y = 0)$ a $\hat{P}(Y = 1)$. S pomoci Baysovy věty obdržíme $\hat{P}(Y = 1|\,\boldsymbol{x})$.

Označme $T^A = \left\{T_0^A, T_1^A\right\}$ jako klienty, kteří spadají do oblasti přijatých. Protože rozdělení na přijaté a zamítnuté závisí na $\boldsymbol{x}$, pak odhad $\hat{p}(\boldsymbol{x}|T_i^A, Y = i)$ $i = 0, 1$ bude vychýlen. Ukážeme si to na příkladě.

Nechť rozhodnutí o přijetí závisí jenom na jediné charakteristice x. Uvažujeme $p(\boldsymbol{x}|Y = 0) = N(\mu, \sigma^2)$ a hranici c. Potom platí, že

$$E\left[x|x \geq c\right] = \mu + \sigma\lambda(\alpha),$$

kde $\alpha = (c - \mu)/\sigma$ a

$$\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$$

Tento výraz se nazývá *Mills ratio*, $\phi(\cdot)$ je hustota normálního rozdělení a $\Phi(\cdot)$ je kvantilová funkce normálního rozdělení. Pro rozptyl takto "ořezaných" hodnot platí

$$Var\left[x|x \geq c\right] = \sigma^2(1 - \delta(\alpha)),$$

kde $\delta(\alpha) = \lambda(\alpha)(\lambda(\alpha) - \alpha)$

Pro lepší představu zkusíme jeden ilustrativní příklad[1]":

Nechť $p(x|Y = 0) = N(2, 1)$, $p(|Y = 1) = N(6, 1)$, $P(Y = 0) = P(Y = 1) = 1/2$ a stanovíme hranici $c = 3$. Dále označme x_0 jako špatného klienta, x_1 jako dobrého. Pak $E\left[x_0|x_0 \geq 3\right] \approx 3.53$ a $E\left[x_1|x_1 \geq 3\right] \approx 6$. Dále $Var\left[x_0|x_0 \geq 3\right] \approx 0.2$ a $Var\left[x_1|x_1 \geq 3\right] \approx 0.99$. Pozorujeme, že $\hat{p}(x|x \geq 3, Y = 0)$ bylo stanovenou hranici velice zdeformováno. Střední hodnota vzrostla z 2 na 3.53, zatímco rozptyl poklesl z 1 na 0.2. Na druhé straně $\hat{p}(x|x \geq 3, Y = 1)$ zůstalo téměř zachováno, protože jenom relativně malé množství dobrých klientů bylo zamítnuto. Dále $P(Y = 0|x \geq 3) \approx 0.14$ a $P(Y = 1|x \geq 3) \approx 0.86$, což znamená, že množství dobrých klientů v populaci bylo značně nadhodnoceno.

Na rozdíl od diskriminační analýzy logistická regrese odhaduje $P(Y = 1|\boldsymbol{x})$ přímo,

$$P(Y = 1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta x})}{1 + \exp(\beta_0 + \boldsymbol{\beta x})}$$

Protože při odhadu nepoužívá odhady $\hat{P}(Y = 0)$ a $\hat{P}(Y = 1)$, ani $\hat{p}(\boldsymbol{x}|Y = i)$, $i = 0, 1$, které mohou byt vychýlené, nedochází při použití logistické regrese k tak výrazným odchylkám, jako v předchozím případě.

[1]tento příklad je převzat z Feelders (2003)

Jako důsledek si můžeme vyvodit to, že pokud rozhodnutí o poskytnutí úvěru záleží jenom na charakteristikách $\boldsymbol{x}$, pak lze očekávat, že metody typu logistická regrese budou účinnější než klasická diskriminační analýza.

3.2 Využití zamítnutých klientů

Předpokládejme, že určité množství žadatelů s charakteristikami $\boldsymbol{x}$ je přijato na základě svých charakteristik. Pak pravděpodobnost, že klient s charakteristiky $\boldsymbol{x}$ je dobrý za podmínky, že je přijatý, se bude rovnat pravděpodobnosti, že klient se stejnými charakteristikami $\boldsymbol{x}$ je dobrý za podmínky, že je zamítnutý. To ale platí jenom za určitých podmínek. Hand, Henley (1993) tvrdí, že rovnost $P(Y = 1|\boldsymbol{x}, A = 1) = P(Y = 1|\boldsymbol{x}, A = 0)$ bude platí, pokud:

(a) rozhodnutí o přijetí/zamítnutí klienta, záleží jenom na pozorovatelných charakteristikách $\boldsymbol{X}$. Nejsou využívané žádné KO[2] kritéria nebo naopak "dobrý" dojem z klienta.

(b) pokud rozhodnutí se provádí na základě charakteristik $\boldsymbol{Z}$, pak v $\boldsymbol{Z}$ musí být zahrnuté všechny proměnné z $\boldsymbol{X}$, žádná z nich nesmí být opomíjená nebo nahrazena jinou.

Pokud tyto podmínky nebudou splněné, pak model postavený jenom na přijatých klientech bude vychýlený. Abychom se tomuto vychýlení vyhnuli, je nutné začlenit i zamítnuté klienty do stavby modelu.

Existuje řada technik, které využívají zamítnuté klienty. Zmíníme se o dvou nejrozšířenějších, které se často využívají v praxi, ačkoliv jejich skutečný přínos není jednoznačně prokázán.

Augmentation

Jedná z metod, která řeší otázku *reject inference*, je metoda *augmentation* nebo taky *reweighting*. Metoda začíná vývojem nového skóringového modelu s využitím charakteristik $\boldsymbol{Z}$. Důležitým předpokladem je, že $\boldsymbol{Z}$ nezahrnuje v sobě všechny proměnné $\boldsymbol{X}$, které se používali při rozhodování o přijetí nebo zamítnutí v předchozím modelu. Model ohodnotí přijaté a zamítnuté klienty z minula na základě charakteristik $\boldsymbol{Z}$. Díky předpokladu $\boldsymbol{Z}$ není stejné jako $\boldsymbol{X}$ nastává situace, kdy pro každé konkrétní $\boldsymbol{z}$ máme jak přijaté, tak i zamítnuté klienty. Následně se pomocí přijatých klientů odhadne podmíněná pravděpodobnost $P(Y = 1|\boldsymbol{z})$. Hustota $p(\boldsymbol{z})$ pak slouží jako váhy.

Pro lepší pochopení se podívejme na tabulku (3.1). Představuje roztřídění klientů do n intervalů podle skóre. Pro každý interval j máme počet přijatých

[2]KO kritéria jsou informace, díky kterým se klient zamítne, ačkoliv jeho charakteristiky nejsou špatné. Jde například o vysoký příjem, ale pocházející z trestní činnosti atd.

klientů A_j, počet zamítnutých R_j. To odpovídá předpokladu, že pro každé konkretní $\boldsymbol{z}$ máme množství přijatých a zamítnutých klientů. Dále pro každé A_j označíme g_j jako počet dobrých klientů a b_j jako špatných klientů.

Interval(j)	Počet dobrých	Počet špatných	Počet přijatých	Počet zamítnutých	Váha v intervalu
1	g_1	b_1	$A_1 = g_1 + b_1$	R_1	$\frac{R_1+A_1}{A_1}$
2	g_2	b_2	$A_2 = g_2 + b_2$	R_2	$\frac{R_2+A_2}{A_2}$
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
n	g_n	b_n	$A_n = g_n + b_n$	R_n	$\frac{R_n+A_n}{A_n}$

Tabulka 3.1: Re-weighting

Důležitý předpoklad je, že v každém intervalu je stejná pravděpodobnost defaultu, a to jak u přijatých, tak i u zamítnutých klientů.

$$P(Y = 1|S_j, A) = P(Y = 0|S_j, R), \tag{3.1}$$

kde S_j je skóre intervalů. Z toho plyne, že

$$\frac{g_j}{A_j} = \frac{g_j^r}{R_j} \tag{3.2}$$

kde g_j^r je počet odhadovaných dobrých klientů mezi zamítnutými.

Naším cílem je, aby nový model byl vybudován i s ohledem na tyto zamítnuté případy, proto zvážíme přijaté klienty inverzní hodnotou odhadu pravděpodobnosti, že klient bude přijatý, tedy A_j+R_j/A_j. Díky tomu přijaté klienti v intervalu j budou moci reprezentovat i ty zamítnuté ve stejném intervalu. Nový model se pak postaví na takto zvážených přijatých klientech.

Augementation má velký nedostatek, a to pravě v svém klíčovém předpokladu, protože rovnost $P(Y = 1|S_j, A) = P(Y = 0|S_j, R)$ obecně nemusí platit. Naopak, pokud původní model měl nezanedbatelní diverzifikační schopnosti, pak tento předpoklad nebude splněn.

Iterative reclassification

Jiný přistup než augmentation nabízí *iterative reclassification*. Princip, na jakém pracuje, se dá popsat následujícími čtyřmi kroky :

(1) Klasifikujeme zamítnuté na základě pravidla vybudovaného na přijatých, t.j. označíme zamítnuté jako dobré nebo špatné, v závislosti na jejich zpětně vypočítané pravděpodobnosti defaultu. Nechť m_0 a m_1 jsou počty zamítnutých nově přiřazených do skupiny špatných a dobrých klientů.

(2) Znovu vybudujeme model, využívaje přitom $n_0 + m_0$ pozorování ze skupiny špatných, a $n_1 + m_1$ ze skupiny dobrých. Počet dobrých, respektive špatných klientů mezi přijatými značíme jako n_1, resp. n_0.

(3) Znovu rozdělíme populaci na přijaté a zamítnuté na základě jejich nové pravděpodobnosti defaultu, kterou jsme dostali z nového modelu. Tím pádem získáme nové hodnoty n_1, n_0.

(4) Opakujeme tento postup, dokud se nepřestanou měnit odhadované parametry v modelu.

Táto technika má několik nedostatků. Zaprvé, neustálým opakováním uvedeného postupu nemusíme dojít do stavu, kdy se ustálí počet zamítnutých a přijatých, čili hodnoty n_1, n_0, m_0, m_1 se budou neustále měnit. Zadruhé, nemáme jistotu, že model, který postavíme na přijatých, bude dobře vystihovat i chování zamítnutých případů. Protože čím lepší diverzifikační schopnost bude mít původní model, tím víc se od sebe budou lišit přijatí a zamítnutí klienti.

3.3 Metody s dodatečnými informacemi

Jak jsme už několikrát opakovali, obtíže při tvorbě nového skóringového modelu by se dáli z velké části vyřešit, pokud bychom znali platební chování zamítnutých klientů. Je proto zřejmé, že nejpřímočařejší způsob, jak vyřešit problém reject inference, je získat skutečné, a ne jenom odhadnuté informace o těchto klientech.

První způsob, který nás může napadnout, je poskytovat úvěr každému žadateli. Po nějaké době velice spolehlivě získáme profil klienta, u kterého s největší pravděpodobností nastane default. Ale v praxi je to velice obtížně prověditelné, proto se v Parnitzke (2005) objevuje modifikace této metody, kterou autor pojmenovává *enlargement*. Její princip je následující: ze skupiny klientů, kteří by byli za normálních okolnosti zamítnutí, se vybere určité množství, kterým bude úvěr poskytnut. Tím se skupina přijatých zvětší, z toho plyne i název metody, enlargement (zvětšení). Tato metoda nám dodává data navíc, ale je velice riskantní, protože úvěr může být poskytnut klientům s vysokou pravděpodobností defaultu.

Kapitola 4

Chybějící data

Protože nemáme údaje o tom, zda by zamítnutý klient byl ve skutečnosti dobrý nebo špatný, tak označíme tyto údaje za *chybějící data*. Proto hodně autoru v poslední době označuje problém *reject inference* jako problém chybějících dat. V této kapitoly se budeme věnovat popisu vztahu mezi *reject inference* a chybějícími daty a taky se zmíníme o několika způsobech, pomocí nichž můžeme analyzovat data.

4.1 Mechanismus chybějících dat

Jeden z nejdůležitějších pojmů, který se vyskytuje v souvislosti s chybějícími daty, je *mechanismus chybějících dat*. Mechanismus chybějících dat je rozdělení *nepřítomnosti* dat. Nepřítomnost dat rozumíme náhodnou veličinu, která označuje, zda data chybí či ne. Abychom pojem mechanismus správně zavedli, odkážeme se na práci Little, Rubin (2002).

Nechť $Y = (y_{ij})$ je matice pozorování o rozměrech $(n \times k)$, kde sloupce představují vysvětlující navzájem nezávisle proměnné a řadky jednotlivá pozorování. Definujeme dále matici indikátorů $R = (r_{ij})$ o rozměrech $(n \times k)$ takovou, že $r_{ij} = 0$, pokud hodnota y_{ij} chybí, a $r_{ij} = 1$, pokud hodnota y_{ij} je pozorována. Mechanismus chybějících dat je definovan jako *podmíněné rozdělení* R za podmínky Y. Označíme to jako $P(R|Y, \psi)$, kde ψ jsou neznámé parametry. Mechanismy rozlišujeme na :

1. *Chybějící zcela náhodně* (*missing completly at random, MCAR*), pokud rozdělení nepřítomnosti dat nezávisí na hodnotách Y, ať už jsou či nejsou pozorovaná. To jest $P(R|Y, \psi) = P(R|\psi)$, pro všechny hodnoty v Y a ψ.

2. *Chybějící náhodně* (*missing at random, MAR*), pokud rozdělení nepřítomnosti závisí pouze na pozorovaných hodnotách Y, (ozn. jako Y_{obs}), a ne chy-

bějících hodnotách Y, (ozn. jako Y_{mis}). To jest, $P(R|Y,\psi) = P(R|Y_{obs},\psi)$ pro všechny Y_{mis} a ψ.

3. *Chybějící nenáhodně* (*missing not at random, MNAR*), pokud rozdělení nepřítomnosti závisí na Y_{mis}, a případně i na Y_{obs}. To jest $P(R|Y,\psi) \neq P(R|Y_{obs},\psi)$.

Little,Rubin (2002) přirovnávají tyto mechanismy ke schopnosti předpovídat hodnoty matice R, t.j. které hodnoty Y budou chybět.

4.2 Chybějící data a reject inference

V případě *reject inference* máme vektor $\boldsymbol{X}$ nezávislých proměnných $X_1,\ldots,X_k$. Předpokládáme, že hodnoty $X_1,\ldots,X_k$ nechybí pro žádné pozorování. Jak už víme z úvodu, skóringový model přiřadí každému pozorování i (=žadatel o úvěr) skóre S_i. T.j. $S_i = f(\boldsymbol{x}_i)$, kde $f(.)$ je skóringová funkce. Dále máme stanovenou prahovou hodnotu c, pro kterou platí, že pokud $S_i \geq c$, pak je úvěr poskytnut, v opačném případě je žadatel zamítnut. Nechť A je indikátor, který označuje $A_i = 1$, pokud je úvěr poskytnut i-tému klientovi , a $A_i = 0$ pokud ne. Označíme Y jako pozorovaný výsledek přijatého klienta, $Y = 0$ označuje default, jinak $Y = 1$. Taky označíme $Y =?$ jako výsledek zamítnutých žadatelů. Pro jednoduchost seřadíme jednotlivé klienty dle skóre od S_{min} po S_{max}, a uspořádáme do podoby, jakou vidíme v tabulce (4.1) (pro $\forall\epsilon > 0$).

	X_1	X_2	...	X_k	S	Y	A
1					S_{max}	.	1
2					.	.	1
.					.	.	.
.					.	.	.
.					c	.	1
.					$c-\epsilon$	?	0
.					.	?	0
.					.	.	.
n					S_{min}	?	0

Tabulka 4.1

Jednotlivé mechanismy pro problém reject inference budou následující.

1. MCAR - pravděpodobnost, že bude Y pozorováno, t.j. pravděpodobnost, že $A = 1$, nezávisí na ani na hodnotě Y, ani na $\boldsymbol{x}$.

$$P(A = 1|\boldsymbol{x}, Y) = P(A = 1) \tag{4.1}$$

Tento mechanismus odpovídá situaci, kdy úvěr pro klienta schvaluje náhodnou, což se z pochopitelných důvodů neděje. V každém případě, pokud mechanismus chybějících dat je MCAR, pak reject inference není potřeba, neboť díky náhodnosti výběru máme k dispozici klienty z celé populace.

2. MAR - přijetí klienta závisí na $\boldsymbol{x}$, ale ne na Y.

$$P(A = 1|\boldsymbol{x}, Y) = P(A = 1|\boldsymbol{x}). \tag{4.2}$$

Tuto rovnici můžeme taky napsat ve tvaru

$$P(Y = 1|\boldsymbol{x}, A = 1) = P(Y = 1|\boldsymbol{x}, A = 0) = P(Y = 1|\boldsymbol{x}), \tag{4.3}$$

ze které vidíme, že v případě MAR je pravděpodobnost, že klient je dobrý, nezávisí na tom, zda je přijatý nebo zamítnutý.

3. MNAR - přijetí klienta závisí jak na $\boldsymbol{x}$, tak i na Y.

$$P(A = 1|\boldsymbol{x}, Y) \neq P(A = 1|\boldsymbol{x}), \tag{4.4}$$

a tedy i

$$P(Y = 1|\boldsymbol{x}, A = 1) \neq P(Y = 1|\boldsymbol{x}, A = 0). \tag{4.5}$$

Hand,Henley (1993) uvádějí, že rozhodnutí o přijetí či nepřijetí záleží jenom na $\boldsymbol{Z}$ a ne na "skrytých" informacích, a pokud $\boldsymbol{Z}$ zahrnuje všechny charakteristiky $\boldsymbol{X}$, podle kterých se rozhodoval předcházející model, pak platí $P(Y = 1|\boldsymbol{x}, A = 1) = P(Y = 1|\boldsymbol{x}, A = 0)$., což, jak vidíme, je přesně definice mechanismu MAR. Takže pokud předpokládáme, že mechanismus chybějících dat je typu MAR, pak metody, jako například logistická regrese, nám dají nevychýlený odhad pravděpodobnosti, že klient s charakteristikami $\boldsymbol{x}$ je dobrý, i když budeme budovat model jenom na přijatých případech.

Mechanismus MAR je běžným předpokladem pro mnoho technik reject inference, například augmentation. Ale ve většině případů je to předpoklad nereálný, což způsobuje, že techniky reject inference nejsou v praxi příliš účinné.

V poslední době se problém reject inference snaží řešit pomocí technik z oblastí metody analýzy chybějících dat. Důležitá otázka, kterou je potřeba v tom případě zodpovědět, je, o jaký mechanismus se jedná. Pokud znovu předpokládáme, že mechanismu bude typu MAR, pak nelze očekávat, že si poradí s problémem zamítnutých žadatelů lépe než techniky reject inference nebo obyčejné použití přijatých případů.

Nyní si uvedeme přehled metod analýzy chybějících dat. Existuje velké množství používaných technik, ale my budeme věnovat pozornost především těm, které se používají pro řešení problému zamítnutých klientů.

4.3 Analýza chybějících dat

Dle Fogarty(2006) můžeme rozdělit metody zpracování ztracených dat do následujících čtyř kategorií [1]:

1. Complete-case analýza
2. Available-case analýza
3. Vážící procedury
4. Imputation-based procedury

Complete-case analýza

Complete-case analýza je zaměřená na případy, které neobsahují žádnou chybějící informaci. Všechny neúplné případy se vyloučí a dále se s nimi nepracuje. Tento postup má dvě hlavní výhody, a to (1) jednoduchost provedení a (2) možnost použití klasických statistických metod, protože se s daty pracuje jako s běžným výběrem. Nevýhody této techniky pramení z vyloučení neúplných případů. Ztráta údajů má dva aspekty: ztráta přesnosti a vychýlení (*bias*), a to v případě, že mechanismus chybějících dat není MCAR.

Tento postup se vyplatí v případě, že užitek informací z neúplných případů je minimální a nehrozí velká ztráta na přesnost nebo výrazné vychýlení. Jinými slovy, tuto techniku můžeme použít, pokud počet vyloučených případů není příliš vysoký, nebo pokud se od sebe příliš neliší úplná a neúplná data.

Jako příklad můžeme uvést techniku *listwise deletion*, která děla přesně to, co bylo popsáno, neboli vyloučí všechny neúplné případy. V souvislosti s kreditním scóringem tato technika přesně odpovídá tomu, že pro tvorbu nového skóringového modelu vyloučíme údaje o všech klientech, které nedostali úvěr.

Available-case analýza

Complete-case analýza často zbytečně plýtvá informacemi v případě analýzy střední hodnoty nebo marginálního rozdělení jednotlivých proměnných. Protože vyloučí případ, který obsahuje byť jedinou chybějící proměnnou, tak při větším množství vysvětlujících proměnných to může být velký problém. Například při počtu 20 proměnných a pravděpodobnosti 10%, že proměnná bude chybět, pravděpodobnost úplného případu je $0.9^{20} \cong 0.12$. Takže po vyloučení všech neúplných případů máme jenom 12% případů z celkového množství.

[1] Fogarty(2006) čerpá z Little, Rubin, (2002), kde techniky na zpracování chybějících dat jsou rozdělení jinak: první a druhá kategorie jsou spojené v jednu, čtvrtá je naopak rozdělená na dvě

Naproti tomu *available-case* analýza ponechává všechny dostupné informace. Pouze pokud při výpočtu charakteristiky, která nás zajímá, najde chybějící údaj, tak ho vyloučí. Ačkoliv je mnohem šetrnější než *casewise deletion*, tak i tento způsob může vést k vychýleným výsledkům, a to v případě, že údaje nejsou ztracené náhodně, čili jsou MCAR.

Pro nás ale rozdíl mezi complete-case analýzou a available-case analýzou není až tak podstatný. V případě kreditního skóringu nás zajímá především to, zda klient je anebo není dobrý, údaje, jako např. průměrný příjem nebo průměrný věk žadatele pro nás nejsou důležité. Takže pokud vyloučíme všechny případy, u kterých nám údaj o splacení chybí, tak obdržíme znovu jenom přijaté případy.

Vážicí procedury

Další kategorii jsou tzv. *vážící procedury*. Techniky z této kategorie využívají *vah* k analýze náhodného výběru. Tyto váhy jsou inverzní hodnotou k pravděpodobnosti zahrnutí do výběru, v případě, že neobsahují žádné pozorování bez odpovědi. Uvedeme si příklad použití vážících procedur. Nechť y_i označuje hodnotu proměnné Y pro i-tou jednotku populace. Pak střední hodnota Y v populaci je často odhadována pomocí Horvitz-Tompsnonová odhadu :

$$\left(\sum_{i=1}^{n} \pi_i^{-1} y_i\right) \left(\sum_{i=1}^{n} \pi_i^{-1}\right)^{-1} \tag{4.6}$$

kde se sčítá přes všech n jednotek výběru, a π_i je pravděpodobnost zahrnutí do výběru pro jednotku i. Pokud se, ale v populaci objeví jednotky bez výstupní závislé proměnné, pak se odhad (4.6) nahradí

$$\left(\sum_{i=1}^{n} (\pi_i \hat{p}_i)^{-1} y_i\right) \left(\sum_{i=1}^{n} (\pi_i \hat{p}_i)^{-1}\right)^{-1} \tag{4.7}$$

kde se sčítá přes všechny jednotky, u kterých známe y_i, a $\hat{p}_i$ je pravděpodobnost, že u i-té jednotky budeme znát výstupní proměnnou.

Do této kategorie můžeme zařadit techniku augmentation, kterou jsme popsali v předchozí kapitole.

Imputation-based procedury

Tato kategorie je nejrozsáhlejší a pro řešení problému reject inference má největší význam. Tyto procedury se začaly pro účel reject inference používat relativně nedávno. Matematický popis těchto metod ale přesahuje rámec této práce, podrobnější informace lze najít v Little, Rubin (2002). Imputation-based procedury se snaží nahradit chybějící údaj nějakou vhodnou hodnotou. Oproti ostatním metodám mají řadu výhod. Předně je zde menší vychýlení než u jednodušších technik typu listwise deletion. Další výhodou je, že pokud jsou chybějící data nahrazená, pak můžeme používat standardní statistické metody (např.

logistickou regresi) na celý soubor dat. Imputation-based procedury jsou taky univerzálnější než techniky rejet inference, protože nejsou tak zatížený různými předpoklady, obzvlášť častým předpokladem, že mechanismus chybějících dat je MAR.

Na základě způsobu, jakým procedury určí náhradní hodnotu za tu chybějící, je dělíme na dvě skupiny : metody *non-model based* a metody *model based*.

1. **Non-model based procedury**

   Jak vyplývá z názvu, techniky této procedury nevyužívají k odstranění chybějících hodnot žádný model ani informace jiných zdrojů, pracují jenom s hodnotami, které mají k dispozici. Nejběžnější technikou tohoto typu je *mean imputation*, která nahrazuje chybějící data střední hodnotou. Mean imputation je rozšířená v různých komerčních analýzách, ale v případě reject inference se tato technika nepoužívá. Data, která nám zde chybí, jsou závislé proměnné, označující, zda klient je dobrý nebo špatný, a ty nabývají pouze dvou hodnot.

2. **Model based procedury**

   Existují dva typy model-based procedur. Rozlišujeme je dle typu modelu, na kterém jsou založené. Tyto modely mohou být *implicitní* a *explicitní*.

   Implicitní model - pracuje na principu využití souboru dat nějakým určitým algoritmem. Patří sem například techniky *hot* a *cold deck*. Hlavní myšlenka těchto dvou technik je nahrazení chybějící hodnoty skutečnou hodnotou z podobného pozorování. Skutečná hodnota patří pozorovaní, které je bude v současné době ve stejném souboru dat, který analyzujeme, nebo je nahrazena hodnotou z jiného zdroje, například z předchozího průzkumu. V prvním případě to označujeme jako metodu hot deck, v druhém jako cold deck. Hot deck se používá i pro řešení problému reject inference. Hlavní výhodou této metody je to, že chybějící údaj je nahrazen reálnou, a ne pouze teoretický odhadnutou hodnotou. Nevýhodami jsou pak obtížně odhadnutelná chyba a problém najít podobné pozorování při větším počtu vysvětlujících proměnných.

   Explicitní model - metody z této kategorie předpokládají, že data mají nějaký model. Chybějící údaje jsou pak dopočítávané z tohoto modelu. Patří sem například *regression imputation*, *stochastic regression*, *composite methods*. Pro problém reject inference se nejčastěji používá metoda *maximální věrohodnosti*.

Do kategorie model-based procedury spadá i již zmiňovaná technika extrapolace z přijatých, o které jsme se zmiňovali v minulé kapitole, protože využívá model vyvinutý na přijatých klientech, aby ho použila k odhadu pravděpodobnosti splacení u zamítnutých klientech.

Kapitola 5

Testování

V této kapitole se pokusíme pomocí testu na reálných datech zodpovědět otázku, jak moc ovlivňuje dostupnost informací o zamítnutých klientech diverzifikační schopností skóringového modelu. K tomu použijeme statistický software *R Project*(http://cran.r-project.org/). Naše testování nebude zaměřeno na vyzkoušení některé z popsaných technik reject inference. Hlavním důvodem, proč je nebudeme testovat, je poměrně velké množství článků, které se zabývají podrobně účinností těchto technik v praxi. Jako příklad můžeme uvést Banasik, Crook (2004), věnovaný testu techniky *augmentation*. Námět pro náš test nám poskytla práce Van den Poel, Verstraeten (2004).

Máme k dispozici údaje o 1000 klientech. Tyto data pocházejí ze www.stat.uni-muenchen.de. Každý klient má uvedené svoje osobní charakteristiky a taky informaci o tom, zda se jedná o dobrého či špatného klienta. Test bude probíhat tak, že budeme stavět skóringové modely na tréninkovém vzorku a ověřovat jejich diverzifikační schopnosti na vzorku validačním. Složení obou vzorků se bude měnit, abychom vyzkoušeli různé situace. Důležité je to, že pokud tréninkový vzorek bude obsahovat zamítnuté klienty, pak u těchto klientů budeme skutečně vědět, zda tento klient je dobrý, nebo špatný. V tom se naše situace liší od skutečnosti, kdy tento stav u zamítnutého klienta pouze odhadujme. Jinými slovy, pokud bychom na odhad skutečného stavu zamítnutých klientů použili nějakou techniku reject inference, pak by byla 100% úspěšná.

Konkretně se pokusíme zodpovědět tři otázky.

1. Pokusíme se zjistit, zda námi postavený model pracuje dobře jak na přijatých, tak i na zamítnutých klientech. Vyzkoušíme to tak, že aplikujeme model vyvinutý na tréninkovém vzorku, na dva validační vzorky. Jeden takový vzorek se bude skládat výhradně z přijatých klientů, druhý ze zamítnutých.

2. Druhá otázka je podstatnější. Vyvineme dva modely, jeden na tréninkovém vzorku složeného pouze z přijatých klientů, a druhý na tréninkovém vzorku,

který bude obsahovat jak přijaté, tak i zamítnuté. Oba dva pak vyzkoušíme na validačním vzorku, který bude obsahovat jak zamítnuté, tak i přijaté klienty. Tímto testem se pokusíme zjistit, zda skutečně potřebujeme zamítnuté klienty.

3. Jako poslední věc si vyzkoušíme, jak moc ovlivní znalost zamítnutých případů diverzifikační schopnosti skóringového modelu, pokud se změní vysvětlovací proměnné.

Nyní si konkrétněji popíšeme testovací postup.

5.1 Rozdíly mezi přijatými a zamítnutými

Popis testu

Jak již bylo řečeno, máme k dispozici údaje o tisíci klientů. Je mezi nimi 700 dobrých, a 300 špatných. Ačkoliv budeme pracovat se skóringovými modely, nebudeme využívat skóre, kterého jednotliví klienti dosáhnou, ale jejich odhadnutou pravděpodobnost splatit úvěr. Děláme jednak proto, že skóre přímo souvisí s pravděpodobností splatit úvěr, a jednak proto, že díky zmenšenému počtu proměnných (z 20 na 7), rozpětí skóre nebylo velké, proto bylo obtížné stanovit 300 nejhorších klientů dle skóre. Použité vysvětlující proměnné jsou uvedené v tabulce (5.1):

název	popis	počet kategorii
účet	množství prostředků na účtu	4
splatnost	doba splatnosti (v měsících)	10
morálka	splácení předchozích úvěrů	5
výše	výše úvěrů	10
úspory	výše úspor a cenných papírů	5
dobazam	doba současného zaměstnání (roky)	5
stav	pohlaví a rodinný stav	4
poměr	poměr výše splátky ku příjmu	4

Tabulka 5.1: Vysvětlující proměnné

Tyto proměnné byly zvolené z důvodu, že se ukázaly jako nejvýznamnější pro vysvětlení vztahu individuálních charakteristik klienta a jeho schopnost splatit úvěr.

Další věc, kterou musíme popsat, je stanovení hranice, která bude rozdělovat přijaté a zamítnuté. Nejdříve si na začátku zvolíme, kolik procent bude v celkové

populaci přijatých a zamítnutých. Vyzkoušíme několik takových hranic. Konkrétněji se jedná o rozdělení populace na přijaté ku zamítnutým, a to v poměru 55 : 45, 70 : 30, 85 : 15. Vyzkoušíme dva různé přístupy, jak určit, zda klient bude přijatý, či zamítnutý. Tyto dva přístupy se budou od sebe podstatně lišit. První bude pracovat náhodně, čili náhodně vybere např. 70% populace a označí je jako přijaté. Zbývající 30% se budou brát jako zamítnutí. Druhý vybere přijaté na základě jejich odhadnuté pravděpodobnosti splácení, t.j. seřadí klienty dle odhadu pravděpodobnosti na všech datech, a označí za přijaté 70% nejlepších z nich.

Teď se budeme věnovat postupu při hledání odpovědi na první otázku. Popíšeme postup při stanoveném poměru 70 : 30 a nenáhodném výběru přijatých, resp. zamítnutých. Při jiné hranici a náhodném výběru je postup analogický. Rozdělíme populaci na přijaté a zamítnuté. Máme 700 přijatých a 300 zamítnutých případů. Nejprve si vytvoříme tréninkový vzorek. V tomto případě se tréninkový vzorek skládá výhradně z přijatých případů. Proto náhodně rozdělíme 700 na dvě skupiny po 350 případech. Na tréninkovém vzorku, pomocí logistické regrese, vyvineme skóringový model. Tento model pak vyzkoušíme na validačním vzorku. Diverzifikační schopnost tohoto modelu určíme pomocí kritéria AUC, a pro srovnání uvedeme i Giniho koeficient. Poté vyzkoušíme vyvinutý model na skupině, kterou jsme označili jako zamítnuté klienty, a určíme AUC. Pro větší spolehlivost necháme tento cyklus proběhnout 1750 krát, t.j. 1750 krát náhodně vytvoříme tréninkový a validační výběr. Ze spočítaných AUC uděláme průměr.

Celý tento postup zopakujeme pro různé hranice i pro druhý přistup k určovaní přijatých, resp. zamítnutých.

Zajímá nás, jaké diverzifikační schopnost model postavený na přijatých bude mít. Nejdřivé ho aplikujeme na validační vzorek tvořený přijtými, poté na validační vzorek tvořený zamítnutými klienty. Předpokládáme, že tyto schopnosti budou stejné v případě, kdy o přijetí nebo zamítnutí klienta rozhoduje náhoda. A naopak, v případě, že o přijetí klienta rozhoduje odhadnutá pravděpodobnost splácení, by se diverzifikační schopnosti měly lišit. Tuto domněnku ověříme srovnáním AUC kritérií, respektive Giniho koeficientů.

Výsledky

Tabulka číslo (5.2) udává rozložení klientů do jednotlivých vzorků. Je stejná jak pro náhodný výběr zamítnutých a přijatých klientů, tak i pro nenáhodný. Řádky představují jednotlivé vzorky, sloupce pak hranici. Čísla v tabulce udávají, kolik klientů je v jednotlivých vzorcích při různých hranicích. Tréninkový vzorek, na něž byl postaven model, označme jako ACT1, validační vzorek skládající se z přijatých případů jako ACV, a validační vzorek ze zamítnutých jako REV.

Začneme s výsledky testu při náhodné volbě zamítnutých a přijatých. Model byl postaven na vzorku ACT1, a vyzkoušen na vzorcích ACV a REV. Toto bylo zopakováno pro všechny tři zvolené hranice, a bylo spočítáno AUC a Giniho koeficient. Řádek 'rozdíl' znamená rozdíl hodnoty AUC při aplikaci na ACV a na

	Poměr přijatých ku zamítnutým v %		
Výběr	85 : 15	70 : 30	55 : 45
ACT1	425	350	275
ACV	425	350	275
REV	150	300	450

Tabulka 5.2: Rozdělení do vzorku

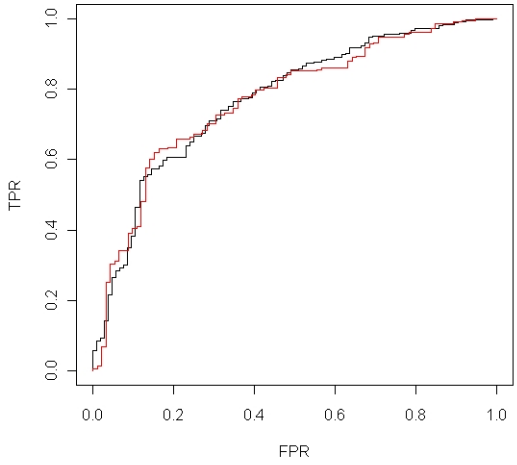
REV.

	Poměr přijatých ku zamítnutých v %					
Výběr	85:15		70:30		55:45	
	AUC	Gini	AUC	Gini	AUC	Gini
ACV	0.77271	0.54542	0.77205	0.54411	0.76827	0.53654
REV	0.77179	0.54358	0.77161	0.54322	0.76878	0.53755
rozdíl	0.00092	0.00184	0.00044	0.00089	-0.00051	-0.00101

Tabulka 5.3: Výsledky

Na první pohled vidíme, že tyto rozdíly jsou zanedbatelné, čili model je stejně účinný jak na přijatých klientech, tak i na zamítnutých. Není to až tak překvapující, protože rozdělení klientů na přijaté a zamítnuté bylo prováděno náhodným způsobem. Tento model má relativně velkou hodnotu AUC, a to díky náhodnému rozdělení klientů. Takže ačkoliv jsme stavěli model na přijatých případech, tak ve skutečnosti se tréninkový vzorek odpovídal svému složení populaci. Nicméně tato situace není v praxi obvyklá, protože nepředpokládáme, že náš předcházející model má nezanedbatelnou diverzifikační schopnost. Podíváme se ještě na ilustrační graf ROC křivek.

Graf č.1 : ROC křivky ACV, REV, náhodný výběr



Zvolili jsme graf při hranici 70:30. Černá křivka ROC odpovídá testu modelu na vzorku, který tvoří přijatí klienti, červená pak zamítnutí. Graf nám potvrdil předpoklad. Plocha pod křivkami je téměř stejná a samotné křivky se na několika místech protínají, tak že nemůžeme tvrdit, že model předpověděl chování přijatých klientů lepé než zamítnutých.

Nyní se podíváme, jaká je situace v případě, že rozdělení na přijaté a zamítnuté neprobíhá náhodně, ale na základě odhadnuté pravděpodobnosti splacení.

	Poměr přijatých ku zamítnutých v %					
Výběr	85:15		70:30		55:45	
	AUC	Gini	AUC	Gini	AUC	Gini
ACV	0.72020	0.44041	0.68651	0.37301	0.60656	0.21313
REV	0.50434	0.00867	0.60487	0.20974	0.61237	0.22473
rozdíl	0.21587	0.43174	0.08164	0.16328	-0.00581	-0.01161

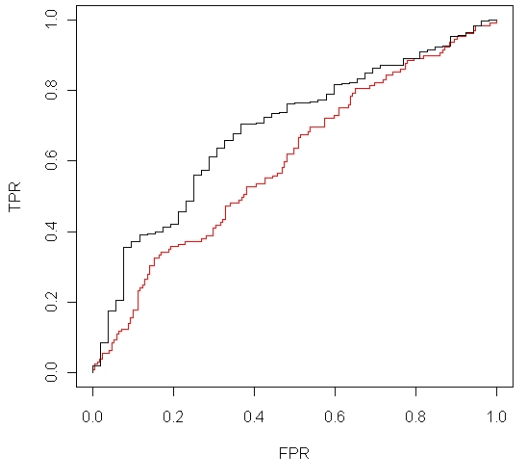
Tabulka 5.4: Výsledky

V tomto případě je situace podstatně jiná než v předchozím. Předně si všimneme, že AUC při aplikaci na ACV roste se zvyšujícím se počtem přijatých. To je způsobeno tím, že s rostoucím množstvím přijatých klientů v populaci, roste i velikost tréninkového vzorku, takže máme více informací ke stavbě modelu. Ale zároveň AUC u REV se zmenšuje. Tento fakt nám potvrdil další předpoklad - čím lepší diverzifikační schopnost model má, tím více se liší přijaté a zamítnuté případy.

Jsou tady dvě zajímavosti, kterých si všimneme když se podíváme na hranici

85:15 a 55:45. Na té první je AUC u REV pouze 0.50434. Tato hodnota odpovídá náhodnému rozhodování, a znamená, že pravděpodobnost, že model označí klienta, jako dobrého, za podmínky, že je špatný, je stejná jako pravděpodobnost, že model označí klienta za dobrého, za podmínky, že je skutečně dobrý. Neboli FPR = TPR. Na hranici 55:45 je rozdíl mezi AUC u ACV a u REV jenom -0.005803. Je to způsobeno tím, že množství zamítnutých klientů je v tomto případě dokonce větší než počet klientů v tréninkovém vzorku (450 versus 275). Nastává tedy situace, o které jsme se již zmínili ve 3.kapitole - díky nedostatku informací nejsme schopni postavit kvalitní model.

Graf č.2 : ROC křivky ACV,REV, nenáhodný výběr



Graf se značně liší od předchozího. Křivka ROC při aplikaci na vzorek z přijatých (černá) je takřka po celé délce výrazně nad křivkou ROC při aplikaci na zamítnuté (červená), což je grafické potvrzení toho, že model postavený na přijatých, odhaduje lepé platební chování přijatých klientů.

5.2 Význam zamítnutých klientů

Popis testu

Postup se částečně shoduje s tím, který jsme popsali v předchozí části. Stanovíme hranici a určíme počet přijatých a zamítnutých. Tentokrát se budeme rozhodovat jenom na základě odhadnuté pravděpodobnosti splacení.

Zadefinujeme *populační vzorek*. Pod tímto pojmem dále budeme rozumět skupinu klientů, která obsahuje jak přijaté, tak i zamítnuté klienty, a to v námi

stanoveném poměru. Čili v případě hranici 70 ku 30 to znamená, že v populačním vzorku bude 70% přijatých a 30% zamítnutých.

Poté, co máme zadefinován populační vzorek, vytvoříme tři vzorky. Jeden se bude skládat pouze z přijatých, zbylé dva budou populační. Nejprve vezmeme tréninkový vzorek z přijatých klientů. Vyvineme model a vyzkoušíme ho na validačním vzorku, který svým složením bude odpovídat populačnímu. Poté vezmeme poslední populační vzorek, prohlásíme ho za tréninkový, a vyvineme na něm další model. I ten pak vyzkoušíme na stejném validačním vzorku. Tento postup necháme proběhnout 1750 krát a spočítáme průměrné hodnoty AUC.

Odpověď na tuto otázku je poměrně důležitá. Zajímá nás, zda je model, který je postaven na populačním vzorku, má lepší diverzifikační schopnosti, než model postavený na přijatých klientech. V případě, že se ukáže byt lepším, je podstatné, zda dosažené zlepšení je významné. Pokud se ukáže, že tento rozdíl není výrazný, pak to znamená, že modely, které jsou vyvíjené u pouze na přijatých klientech, jsou dostatečně efektivní a v takové situaci zamítnutí klienti nejsou potřeba.

Výsledky

Znovu pomocí tabulky (5.5) ukážeme počet pozorování v jednotlivých vzorcích. Pod označením ACT2 rozumíme tréninkový vzorek složený jenom z přijatých klientů, POT je pak tréninkový vzorek skládající se z populačního vzorku. POV je validační vzorek, který je taky populačním. U vzorků skládajících se z populace je v závorce uvedeno, kolik klientů pochází ze zamítnutých.

	Poměr přijatých ku zamítnutým v %		
Výběr	85 : 15	70 : 30	55:45
ACT2	425	350	275
POT1	425(63)	350(105)	275(123)
POV	425(63)	350(105)	275(123)

Tabulka 5.5: Rozdělení do vzorku

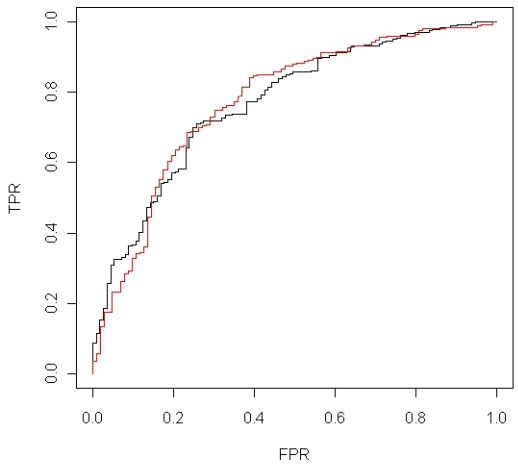
Tabulka (5.6) představuje shrnutí výsledku. Je obdobná té z minulé sekce.

	Poměr přijatých ku zamítnutých v %					
Vyber	85:15		70:30		55:45	
	AUC	Gini	AUC	Gini	AUC	Gini
ACT2 na POV	0.76888	0.53774	0.76676	0.53652	0.74963	0.49927
POT1 na POV	0.77226	0.54451	0.77124	0.54249	0.76734	0.53468
rozdíl	0.00368	0.00677	0.00448	0.00897	0.01770	0.03540

Tabulka 5.6: Výsledky

Tyhle výsledky jsou zajímavé, protože rozdíly mezi dvěma modely jsou relativně malé. Největší rozdíl je při hranici 55:45. Zamítnutí klienti tvoří v tomto případě skoro polovinu populace, takže jejich význam roste. Na druhé straně čím je podíl zamítnutých klientů na populaci menší, tak tím více se rozdíl mezi dvěma modely stírá. Připomeňme ještě, že vycházíme ze situace, kdy všechny výsledky zamítnutých klientů známe skutečně, a ne jenom odhadujeme. Čili jak už bylo řečeno na začátku předchozí kapitoly, námi "použitá" technika reject inference byla zcela úspěšná. Ale 100%-ní úspěch je velice nepravděpodobný, proto rozdíl mezi oběma modely bude ještě nepatrnější. Podíváme se ještě na graf, kde červená ROC křívka představuje použití modelu se zamítnutými klienty, černá pak použiti modelu s přijatými.

Graf č.3 : ROC křivky ACT2,POT1



Křivky na grafu se v několika místech protínají, takže z toho vyplývá, že model postavený s využitím zamítnutých klientů, nemá významně lepší diverzifikační schopnosti, než model vyvinutý jenom na přijatých klientech.

5.3 Změna vysvětlujících proměnných

Popis testu

Jak už jsme několikrát uvedli, tak v Hand, Henley (1993) stojí, že pokud se nový model buduje na přijatých případech a s použitím stejných vysvětlujících proměnných, na kterých bylo učiněno rozhodnutí o přijetí/zamítnutí, pak zamítnutí klienti nepřináší žádné nové informace. Tuto myšlenku jsme potvrdili

předchozím testem. Nyní si nasimulujeme situaci, kdy rozhodnutí o rozdělení klientů na přijaté a zamítnuté bylo učiněno s pomocí jiných charakteristik, než se pak využijí ke stavbě nového modelu.

Postup se shoduje s předcházejícím testováním, až na poslední krok - místo námi používaných proměnných (účet, splatnost, morálka, výše, úspory, doba, stav) použijeme nyní (účet, splatnost, morálka, úspory, poměr). Táto záměna proměn nebyla provedena náhodou - charakteristiky (výše, dobazam, stav) se ukázaly při stavbě modelu na tréninkových vzorcích jako nejméně významné, proto byly z modelu vyloučený, a nahrazený proměnnou (poměr).

Protože jsme provedli takovou výměnu, očekáváme zhoršení predikčních schopností modelů postaveného jak na vzorku tvořeným přijatými případy, tak i na vzorku tvořeným populaci.

Výsledky

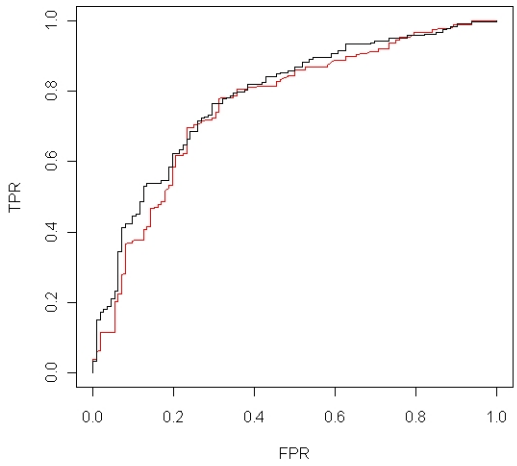
Výsledky budeme znovu prezentovat formou tabulky, kde budeme mít spočítané hodnoty AUC a Giniho koeficient. Vidime, že se prohloubil rozdíl mezi modelem postaveném na přijatých (ACT3) a na populaci (POT2). Tedy při výměně vysvětlujících proměnných hrají zamítnutí větší roli, než v případě, že proměnné zůstávají stejné. Poznamenejme si, že POV je validační vzorek z testovací části 5.2.

	Poměr přijatých ku zamítnutých v %					
Vyber	85:15		70:30		55:45	
	AUC	Gini	AUC	Gini	AUC	Gini
ACT3 na POV	0,67655	0,35309	0,73257	0,46517	0,70915	0,41829
POT2 na POV	0,68329	0,36658	0,74175	0,48349	0,71606	0,43211
rozdíl	0,00674	0,01349	0,00916	0,01832	0,00691	0,01382

Tabulka 5.7: Výsledky

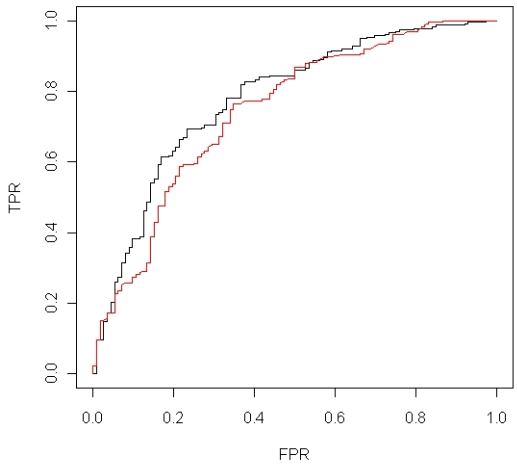
Následující graf je srovnání ROC křivek modelu postavených na populaci. Černá ROC křivka patří modelu POT1, červená pak POT2.

Graf č.4 : ROC křivky POT1, POT2



Poslední graf je srovnání ROC křivek ACT2 (černá barva), a ACT3 (červená).

Graf č.5 : ROC křivky ACT2,ACT3



Poslední dva grafy jsou velice podobné. Jako závěr z této části testu vyvodíme to, že sice zamítnuti klienti hrají roli při stavbě modely, ale testy ukázali, že táto role není tak významná, jak by se mohlo zdát.

Kapitola 6

Závěr

Cílem této práce bylo popsat problém reject inference, jak vzniká a jak se dá řešit. Abychom se mohli bavit o reject inference, potřebovali jsme nejdříve zavést pojem skóringový model. Protože jsme v praktické části práce několik skóringových modelů postavili, bylo nutné se v krátkosti zmínit i o jedné z nejběžnějších metod výstavby, logistické regresi.

Ačkoliv o problému se zamítnutými klienty se začalo mluvit už od počátku vzniku skóringu, do dnešní doby nebyla vydaná žádná rozsáhlejší práce, která by tento problém zevrubně zkoumala. Mnohé z technik reject inference byly vyvíjené, aniž by byly podrobně prověřeny teoretické podklady pro jejich použití. I toto je jeden z důvodů, proč většina technik není v praxi o moc účinnější, než obyčejné využití přijatých klientů.

V posledních 15 letech se problém reject inference začal spojovat s problémem chybějících dat. Důležitou otázkou se stál odhad mechanismu chybějících dat. Jak jsme ukázali předpoklad MAR, který je často využíván metodami analýzy chybějících dat, není správný a takové metody nevedou k významnému zlepšení. Opravdu účinná technika reject inference, respektive analýza chybějících dat, by se měla poradit s chybějícími daty typu MNAR. Jako slibný směr se jeví výzkum metod typu imputation-based. Ale bude potřeba dalších teoretických i praktických výzkumů.

V praktické častí jsme potvrdili některá z tvrzení z Hand,Henley (1993). Konkrétně jsme ukázali, že pokud všechny vysvětlující proměnné pro rozhodnutí přijmout/zamítnout zůstanou zachované i pro tvorbu nového modelu, pak využití výsledků zamítnutých klientů nevede k výraznému zlepšení. Opačná situace je v případě, že se některé proměnné vypustí a nahradí jinými. Význam zamítnutých klientů vzrostl, a naopak model postavený jenom na přijatých případech se výrazně zhoršil.

Výsledky testu byly sice zřejmé, ale mohlo je ovlivnit několik faktorů. Prvním

je volba proměnných. Protože účelem práce nebylo sestavit nejlepší a nejúčinnější model, význam jednotlivých proměnných nebyl podrobně zkoumán a testován. To mohlo samozřejmě ovlivnit jak počáteční rozdělení klientů na zamítnuté a přijaté, tak hlavně stavbu nového modelu. Další faktor, který měl vliv na výsledky, byla poměrně málo rozsáhlá databáze údajů. I ten největší tréninkový vzorek, který zde byl vytvořen, byl v řadech stovek pozorování. Avšak běžné bankovní databáze obsahují informace o desítkách tisíc klientů. To vše mohlo ovlivnit správnost výsledků. Proto, jak již bylo řečeno, je v otázce reject inference nutné dalších výzkumů, především ve směru využití metod analýzy chybějících dat.

Literatura

[1] Åstebro T., Chen G. (2005): *A Maximum Likelihood Approach for Reject Inference*,

[2] Banasik J., Crook J. (2004): *Does Reject Inference Really Improve the Performance of Application Scoring Models?*,Journal of Banking and Finance.

[3] Feelders A.J. (2003): *An Overview of Model Based Reject Inference for Credit Scoring*

[4] Fogarty D.J. (2006): *Multiple Imputation as a Missing Data Approach to Reject Inference on Consumer Credit Scoring*, InterStat, No. 9,

[5] Hand, D.J. , Henley, W.E. (1993): *Can Reject Inference Ever Work?*, IMA Journal of Mathematics Applied in Business and Industry 5, pp. 45-55.

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

[7] Joanes, D.N. (1993): *Reject Inference Applied to Logistic Regression for Credit Scoring*, IMA Journal of Mathematics Applied in Business and Industry, 5(1), pp. 35-43

[8] Little, R.J.A., Rubin D.B. (2002): *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., 3-20

[9] Parnitzke T. (2005): *Credit scoring and the sample selection bias*, Am. J. Phys. 69

[10] Rychnovský M. (2008): *Postupná výstavba modelu ohodnocení kreditního rizika*, MFF UK

[11] Van den Poel D., Verstraeten G.(2004): *The Impact of Sample Bias on Consumer Credit Scoring Performance and Profitability*, 2004