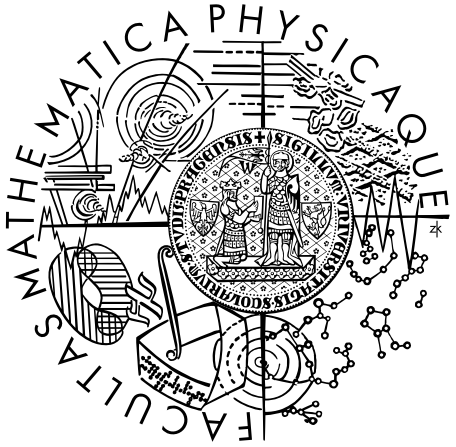
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Karel Komorád

Statistické metody klasifikace a jejich využití pro kreditní skórování

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	Doc. Petr Volf, CSc.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie
Studijní plán:	Matematická statistika

Poděkování

Na tomto místě bych chtěl poděkovat všem svým učitelům a vědeckým pracovníkům z Katedry pravděpodobnosti a matematické statistiky na MFF UK, kteří mi otevřeli dveře do světa matematické statistiky, dále svým spolužákům z téže fakulty za četné diskuze a připomínky, zvláštní dík pak patří vedoucímu mé diplomové práce, doc. Petru Volfovi, CSc. za jeho nekonečnou trpělivost a poskytnutou odbornou pomoc, se kterou mě při psaní této práce provázel.

Tato práce se rodila dlouho a bolestně. Můj velký dík patří rovněž zaměstnancům infekčního oddělení Fakultní Thomayerovy nemocnice v Krči za velkou péči, se kterou se o mě starali. Chtěl bych také poděkovat Prof. Dr. Wolfgangu Härdlemu za zapůjčení zkoumaných dat, která byla poskytnuta z databáze MD*BASE. Nemohu opomenout ani své rodiče, kteří mi umožnili studovat a vždy mě v mém úsilí podporovali. Nakonec připojuji díky své sestře Kateřině za lekce českého jazyka.

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

V Praze dne 27.července 2004 Karel Komorád

Obsah

Název práce: Statistické metody klasifikace a jejich využití pro kreditní skórování

Autor: Karel Komorád

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Doc. Petr Volf, CSc.

E-mail vedoucího: volf@utia.cas.cz

Abstrakt: V naší práci jsme provedli nejprve analýzu empirického datového souboru a následně podrobnou případovou studii využití statistických klasifikačních metod pro kreditní skórování. V porovnání logistického regresního modelu, metody klasifikačních stromů a metody nejbližších sousedů se nejlépe osvědčila logistická regrese, která navíc jako jediná umožnila použití měr podobnosti pro srovnání kvality klasifikačních pravidel. Dále jsme ukázali, že používání standardizovaných nákladů je v případě kreditního skórování nevhodné a může vést až k akceptaci všech žadatelů o úvěr. Rovněž jsme si všimli, že klasifikační stromy mohou postrádat dostatečnou rozlišovací schopnost v případě velice nevyvážených dat.

Klíčová slova: klasifikace, diskriminační analýza, kreditní skórování

Title: Statistical Classification Methods and Their Application in Credit Scoring

Author: Karel Komorád

Department: Department of Probability and Mathematical Statistics

Supervisor: Doc. Petr Volf, CSc.

Supervisor's e-mail address: volf@utia.cas.cz

Abstract: In our thesis we carry out an empirical data set analysis and a thorough case study of statistical classification techniques in credit scoring. For our data set the logistic regression model appears to be the most suitable classification method in comparison with classification trees and k-nearest neighbours method. Moreover, only the logistic regression allows us to use similarity measures for comparison of classifiers. Further we show that the usage of standardized costs is inappropriate in the case of credit scoring and might lead to acceptance of all applicants for a credit. We also figure out that for strongly unbalanced data the classification trees might be lacking in discrimination power.

Keywords: classification, diskriminant analysis, credit scoring

Často používané značení

$\mathcal{K}$	množina zkoumaných objektů (klientů)
$\mathcal{C} = \{1, \ldots, J\}$	množina všech tříd, do nichž patří objekty z $\mathcal{K}$
$\mathcal{X} \subseteq \mathbb{R}^r$	výběrový prostor
$\mathbf{X} = (X_1, \ldots, X_r)^\top$	(sloupcový) náhodný vektor charakterizující zkoumané objekty
$\mathbf{x}$	realizace náhodného vektoru $\mathbf{X}$
Y	náhodná veličina určující příslušnost objektu k jedné z J tříd
$u(\mathbf{x})$	klasifikační pravidlo
$c_{j\|i}$	náklady, které provázejí zařazení objektu třídy i do třídy j
L	střední hodnota ztráty vzniklé při klasifikaci
π_1, π_2	apriorní pravděpodobnost příslušnosti do třídy 1, resp. 2
$f_1(\mathbf{x}), f_2(\mathbf{x})$	hustota rozdělení náhodného vektoru $\mathbf{X}$ objektů třídy 1, resp. 2
$g_1(a), g_2(a)$	hustoty náhodné veličiny $s(\mathbf{X})$ pro objekty třídy 1, resp. 2
$G_1(a), G_2(a)$	distribuční funkce náhodné veličiny $s(\mathbf{X})$ pro objekty třídy 1, resp. 2
$\mathcal{V}_1, \mathcal{V}_2$	vývojový, resp. testovací vzorek
R	míra chybovosti klasifikačního pravidla
$\mathcal{I}$	indikátorová funkce
$s(\mathbf{x})$	skóringová funkce
$\bar{a}$	prahový bod pro skóringovou funkci
$\bar{a}^*$	optimální prahový bod pro skóringovou funkci
$[0, 1]$	uzavřený interval, nebo uspořádaná dvojice bodů

Kapitola 1

Úvod

Mnohé praktické úkoly, které před nás aplikovaná statistika staví, vedou na úlohu klasifikace. Klasifikační (jinak též diskriminační) úlohy obecně předpokládají, že je dáno J známých a poměrně stejnorodých tříd určitých objektů. Cílem je sestrojení klasifikačního pravidla, které dané objekty na základě jejich charakteristik zařadí do příslušných tříd a přitom minimalizuje očekávanou ztrátu způsobenou zařazením objektů do nesprávných tříd.

Z praktických aplikací uveďme např. snahu o stanovení pacientovy diagnózy dle pozorovaných symptomů, úkol na roztřídění námořních plavidel do několika skupin podle záznamu radarových vln, odhad úrovně ozónu pro následující den podle aktuálního stavu atmosféry nebo stanovení přítomnosti daného prvku v chemické sloučenině podle spektrálního rozkladu odraženého světla. Praktické úlohy jsou charakterizovány daty se stále komplexnější strukturou. Kromě vysoké dimenze se jedná o směs kategorických a měřitelných datových typů, nehomogenitu výběrového prostoru (v různých částech platí různé vztahy) či nestandardní datovou strukturu (vektor pozorování má u různých objektů rozdílný počet složek).

K řešení klasifikačních problémů se používají různé techniky. Ke klasickým statistickým přístupům patří např. lineární a kvadratická diskriminační analýza nebo logistická regrese, z novějších metod jmenujme např. klasifikační stromy. V posledních letech jsou často používány některé techniky, které jsou též označovány jako „černé skříňky". Tyto postupy totiž často neuvádějí žádné kritérium, o jehož optimalizaci usilují, ale jsou definovány nějakým algoritmem, který dává návod, jak postupovat při klasifikování. Velice rozšířenými metodami jsou neuronové sítě různých architektur. Ačkoliv není zcela zřejmé,

zda-li takovýmto technikám přísluší přívlastek „statistické", jejich výsledky jsou mnohdy srovnatelné s výsledky klasických statistických procedur.

Cílem naší práce je porovnání klasifikačních metod v případové studii s reálnými daty z oblasti hodnocení žadatelů o bankovní půjčku. Jmenovitě budou zkoumány tyto metody klasifikace: logistický regresní model, klasifikační stromy, metoda nejbližších sousedů a okrajově zmíníme i metodu neuronových sítí. Základní přehled o kreditním skórování je možno nalézt v Thomas (2000) či Lewis (1994).

Při naší analýze jsme používali volně šiřitelný statistický balík R s celou řadou knihoven[1]. Procedury, které jsme psali sami, je možno nalézt na přiloženém kompaktním disku společně s kompletními výstupy analýz a celou řadou doplňujících obrázků.

V následující kapitole formálně popíšeme zkoumaný problém klasifikace a uvedeme některé základní definice. Třetí kapitola se bude blíže zabývat datovým souborem, který je nosnou páteří celé práce. V další kapitole se již budeme věnovat první analýze – logistické regresi. O klasifikačních stromech pojednává 5. kapitola, na kterou navazuje kapitola využívající metodu nejbližších sousedů. V sedmé kapitole porovnáme dosažené výsledky a práci zakončíme souhrnem a závěrečným komentářem v kapitole osmé. V dodatcích přinášíme podrobnější informace k jednotlivým kapitolám, aby jejich zařazení přímo do textu nenarušovalo plynulost čtení.

[1] *http://www.r-project.org*

Kapitola 2

Klasifikace

2.1 Základní úloha

Nechť $\mathcal{K}$ je neprázdná množina studovaných objektů a předpokládejme, že každý z nich náleží právě do jedné z J různých tříd ($2 \leq J < \infty$). Označme $\mathcal{C}$ množinu těchto tříd, bez újmy na obecnosti je $\mathcal{C} = \{1, \ldots, J\}$. Předpokládejme, že na jednotlivých objektech měříme r znaků a každý ze zkoumaných objektů je charakterizován vektorem pozorování $\mathbf{X} = (X_1, \ldots, X_r)^\top$. Výběrový prostor označíme symbolem $\mathcal{X}$ ($\subseteq \mathbb{R}^r$). Příslušnost objektu do nějaké třídy budeme značit symbolem Y. Každému $K \in \mathcal{K}$ je tedy přiřazen náhodný vektor $(\mathbf{X}_K^\top, Y_K)^\top$ definovaný na nějakém pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathrm{P})$ tak, že $\mathbf{X}_K : \Omega \to \mathcal{X}$ je známý vektor napozorovaných hodnot a $Y_K : \Omega \to \mathcal{C}$ je příslušná (ne nutně nám známá) třída. Naším cílem je nalézt klasifikační pravidlo, tj. systematický předpis, který na základě znalosti vektoru $\mathbf{X}$ určí, do které třídy z $\mathcal{C}$ ten který objekt patří.

DEFINICE 2.1 *Klasifikačním pravidlem nazveme funkci $u : \mathcal{X} \to \mathcal{C}$, která každému prvku z prostoru $\mathcal{X}$ přiřadí právě jednu z J tříd.*
Ekvivalentně definujeme klasifikační pravidlo jako rozklad prostoru $\mathcal{X}$ do J po dvou disjunktních borelovských podmnožin $A_1, \ldots, A_J$, $\mathcal{X} = \bigcup_{j=1}^{J} A_j$ tak, že každé $\mathbf{x} \in A_j$ padne do třídy j, tedy že každá množina rozkladu je dána vztahem $A_j = \{\mathbf{x} \in \mathcal{X}; u(\mathbf{x}) = j\}$.

Snaha o přiřazení jedné z J tříd ke každému objektu z $\mathcal{K}$ (takzvaná klasifikace) je tedy ekvivalentní úloze rozlišující J podmnožin (ne nutně souvislých)

v prostoru $\mathbb{R}^r$. Proto se pro stejnou úlohu používá i termín diskriminační analýza.

Nechť je zařazení objektu třídy i do třídy j provázeno náklady $c_{j|i}$. Předpokládejme, že $c_{j|i} > 0$ pro $j \neq i$ a $c_{i|i} \leq 0$, tedy že nesprávné zařazení s sebou nese nenulovou ztrátu a správné zařazení objektu může přinést i zisk. Předpokládejme dále, že rozdělení vektoru $\mathbf{X}$ má hustotu $f_j(\mathbf{x})$ vzhledem k nějaké σ-konečné míře μ, jestliže se jedná o objekty j-té třídy, $j = 1, \ldots, J$. Patří-li objekt do i-té třídy, bude podmíněná střední hodnota ztráty rovna

$$L_i := \sum_{j=1}^{J} c_{j|i} \int_{A_j} f_i(\mathbf{x}) d\mu(\mathbf{x}) = \sum_{j=1}^{J} c_{j|i} \mathrm{P}_i(A_j) \ .$$

Apriorní pravděpodobnost, že objekt patří do i-té třídy, označme symbolem $\pi_i = \mathrm{P}(Y = i)$ a předpokládejme, že $\pi_i > 0$ pro $i = 1, \ldots, J$. Pak je střední hodnota celkové ztráty $L = \pi_1 L_1 + \cdots + \pi_J L_J$. Jako optimální rozklad se přirozeně nabízí ten, který minimalizuje očekávanou ztrátu L.

Označíme-li

$$\bar{f}_j(\mathbf{x}) = \sum_{i=1}^{J} \pi_i c_{j|i} f_i(\mathbf{x}), \quad j = 1, \ldots, J \ ,$$

nechá se L vyjádřit ve tvaru

$$L = \sum_{j=1}^{J} \int_{A_j} \bar{f}_j(\mathbf{x}) d\mu(\mathbf{x}).$$

Následující lemma dává obecný návod k nalezení optimálního klasifikačního pravidla v případě $J \geq 2$.

VĚTA 2.1 *Nechť $\bar{A}_1, \ldots, \bar{A}_J$ je takový rozklad prostoru $\mathbb{R}^r$, že pro každé $\mathbf{x} \in \bar{A}_t$, kde $t = 1, \ldots, J$, je $\bar{f}_t(\mathbf{x}) \leq \bar{f}_j(\mathbf{x})$, $j = 1, \ldots, J$. Pak platí*

$$L \geq \bar{L} := \sum_{j=1}^{J} \int_{\bar{A}_j} \bar{f}_j(\mathbf{x}) d\mu(\mathbf{x}).$$

DŮKAZ Viz Anděl (1985), lemma 15.

V klasické diskriminační analýze se většinou volí $c_{j|j} = 0$ a $c_{j|i} = 1$ pro $i \neq j$, jelikož pak je minimalizace L ekvivalentní minimalizaci počtu chybně

zařazených objektů. Tato volba, kterou budeme nazývat *standardizovanými náklady*, je typická pro mnoho teoretických statí, v praktických úlohách však bývá jen zřídkakdy splněna. Nicméně označíme-li

$$e(\mathbf{x}) = \sum_{i=1}^{J} \pi_i f_i(\mathbf{x}) \ ,$$

dostaneme

$$\bar{f}_j = e(\mathbf{x}) - \pi_j f_j(\mathbf{x}) \ .$$

Při daném $\mathbf{x}$ je potom vztah $\bar{f}_t(\mathbf{x}) \leq \bar{f}_j(\mathbf{x}),\ j = 1, \ldots, J$ z věty 2.1 splněn právě tehdy, když $\pi_t f_t(\mathbf{x}) \geq \pi_j f_j(\mathbf{x}),\ j = 1, \ldots, J$. Toho využívá následující definice:

DEFINICE 2.2 *Bayesovské diskriminační pravidlo zařadí objekt* $\mathbf{x}$ *do té třídy, pro kterou je maximální součin* $\pi_j f_j(\mathbf{x})$*:* $u_b^*(\mathbf{x}) = argmax_{j \in \mathcal{C}} \pi_j f_j(\mathbf{x})$*. Je-li takovýchto indexů* j *více, zařadíme* $\mathbf{x}$ *do libovolné z maximalizujících tříd. Příslušný rozklad má tvar:*

$$A_j = \{\mathbf{x} \in \mathcal{X}; \pi_j f_j(\mathbf{x}) > \pi_i f_i(\mathbf{x}) \ , i = 1, \ldots, J\} \, .$$

Podle věty 2.1 tedy neexistuje klasifikační pravidlo, které by mělo nižší očekávanou ztrátu nežli bayesovské diskriminační pravidlo.

V následující větě, která zobecňuje větu 12.1 z Härdle & Simar (2002), je uveden optimální rozklad výběrového prostoru pro případ dvou tříd a obecných hodnot ztrátové funkce.

VĚTA 2.2 *Diskriminační pravidlo rozlišující mezi dvěma skupinami, které je dáno rozkladem:*

$$A_1 = \left\{ \mathbf{x} \in \mathcal{X}; \frac{\pi_1 f_1(\mathbf{x})}{\pi_2 f_2(\mathbf{x})} > \frac{c_{1|2} - c_{2|2}}{c_{2|1} - c_{1|1}} \right\} \ , \tag{2.1}$$

$$A_2 = \left\{ \mathbf{x} \in \mathcal{X}; \frac{\pi_1 f_1(\mathbf{x})}{\pi_2 f_2(\mathbf{x})} < \frac{c_{1|2} - c_{2|2}}{c_{2|1} - c_{1|1}} \right\} \ , \tag{2.2}$$

kde f_1 *a* f_2 *jsou hustoty rozdělení objektů třídy 1, resp. 2, minimalizuje očekávanou ztrátu* L *(body na hranici mohou být libovolně rozřazeny do jedné z množin* A_1, A_2*).*

DŮKAZ Viz věta A.1 v dodatku A.

Až doposud uvedené vztahy vyžadovaly znalost rozdělení náhodného vektoru $\mathbf{X}$. Diskriminační analýza se zpravidla opírá o předpoklad normality, v kapitole 3 však ukážeme, že náš datový soubor tuto podmínku nesplňuje. V následující podkapitole proto odvodíme, jak vypadá optimální klasifikační pravidlo sestavené na základě skóringové funkce.

Při naší analýze použijeme často užívaného, i když ne zcela efektivního, přístupu a rozdělíme celý datový soubor na dva podsoubory neboli vzorky:

DEFINICE 2.3 *Vzorkem nazveme M-tici po dvou nezávislých dvojic $(\mathbf{X}_1^\top, Y_1)$, ..., $(\mathbf{X}_M^\top, Y_M)$, kde $\mathbf{X}_m \in \mathcal{X}, Y_m \in \mathcal{C}, m = 1, \ldots, M < \infty$. Tuto M-tici označíme symbolem $\mathcal{V}$. Nadto předpokládáme, že pro každé $j \in \mathcal{C}$ platí, že rozdělení $P(A|j) = P(\mathbf{X}_K \in A|Y_K = j)$ jsou stejná pro všechna $K \in \{1, \ldots, M\}$ a $A \subseteq \mathcal{X}$.*

Jeden z těchto vzorků nazveme *vývojovým* (jinak též učebním) vzorkem, označíme jej symbolem $\mathcal{V}_1$ a počet jeho pozorování označíme jako M_1. Vývojový vzorek se užívá k sestavení modelů a určení klasifikačních pravidel u. Druhý z vytvořených vzorků nazveme *testovací*, označíme jej jako $\mathcal{V}_2$ a počet jeho pozorování označíme symbolem M_2. Jednotlivá klasifikační pravidla sestavená na vývojovém vzorku použijeme na testovací vzorek a takto získané odhady $\hat{Y} = u(\mathbf{X})$ porovnáme se skutečnými hodnotami Y. Míru nesprávně zařazených objektů (tzv. *míru chybovosti*), někdy označovanou jako AER (actual error rate), definujeme vztahem

$$R(u) = \mathrm{P}(u(\mathbf{X}_K) \neq Y_K) \text{ , kde } \mathbf{X}_K \in \mathcal{X}, Y_K \in \mathcal{C}, K \in \mathcal{K}$$

a odhadujeme ji buď na vývojovém nebo na testovacím vzorku. Odhad počítaný na vývojovém vzorku:

$$\hat{R}_1(u) = \frac{1}{M_1} \sum_{(\mathbf{X}^\top, Y)^\top \in \mathcal{V}_1} \mathcal{I}\left(u(\mathbf{X}) \neq Y\right) \ ,$$

kde $\mathcal{I}$ je indikátorová funkce, se též označuje jako APER (apparent error rate) a dává příliš optimistické výsledky. Pro srovnávání různých metod je proto vhodnější používat odhad míry chybovosti založený na testovacím vzorku:

$$\hat{R}_2(u) = \frac{1}{M_2} \sum_{(\mathbf{X}^\top, Y)^\top \in \mathcal{V}_2} \mathcal{I}\left(u(\mathbf{X}) \neq Y\right) \ .$$

2.2 Optimální prahový bod

V této podkapitole, a pokud nebude řečeno jinak i v celém zbytku práce, budeme předpokládat, že závisle proměnná veličina má alternativní rozdělení. Její kategorie nyní označme symboly $1, 2$. Kreditní skórování, jako speciální případ klasifikace, se snaží odlišit od sebe dvě skupiny klientů, žadatelů o úvěr. Jednu skupinu tvoří dobří a spolehliví klienti, kteří nejspíše nebudou mít problémy se splácením poskytnutého úvěru. Do druhé skupiny se řadí špatní a nespolehliví klienti, kteří nejspíše nebudou dodržovat splátkový kalendář a upadnou v platební neschopnost ještě před uplynutím doby splatnosti.

Jako *skóringovou funkci* označíme libovolnou nezápornou konečnou funkci s definovanou na množině $\mathcal{X}$, podle jejíž hodnoty rozřazujeme klienty na spolehlivé a nespolehlivé, příp. je dělíme do několika dalších pásem. Snahou je zavést skóringovou funkci tak, aby její vyšší hodnoty odpovídaly nespolehlivým klientům a nižší hodnoty zase spolehlivým klientům z $\mathcal{K}$. Bez újmy na obecnosti můžeme předpokládat, že skóringová funkce s zobrazuje množinu $\mathcal{X}$ do intervalu $[0, 1]$. Hodnotu skóringové funkce pro nějakého klienta z množiny $\mathcal{K}$ nazveme jeho *skórem*.

DEFINICE 2.4 *Mějme dánu skóringovou funkci s a nechť existují podmíněné hustoty náhodné veličiny $s(\mathbf{X}_K)$ za podmínky $[Y_K = 1]$, resp. $[Y_K = 2]$, které označíme g_1 a g_2. Podmíněné distribuční funkce náhodné veličiny $s(\mathbf{X}_K)$ obdobně označíme $G_j(a) = P[s(\mathbf{X}_K) \leq a | Y_K = j]$, kde $j = 1, 2$.*

Mějme nyní nějakou skóringovou funkci s sestavenou na vývojovém vzorku $\mathcal{V}_1$, podle níž chceme rozdělit prvky z $\mathcal{V}_1$ tak, aby objektům třídy 2 příslušely vyšší hodnoty skóre. Přirozeným způsobem je zvolit takzvanou prahovou hodnotu $\bar{a}$ takovou, že pro $\mathbf{x}$ splňující $s(\mathbf{x}) > \bar{a}$ řekneme, že $\mathbf{x} \in A_2$. V opačném případě uvažujeme, že $\mathbf{x} \in A_1$. Vhodnou volbou bodu $\bar{a}$ zajistíme, že bude minimalizována celková očekávaná ztráta $L = \pi_1 L_1 + \pi_2 L_2$:

VĚTA 2.3 *Předpokládejme, že pro náklady spojené s klasifikací platí:*

$$c_{1|1}, c_{2|2} \leq 0 \qquad c_{1|2}, c_{2|1} > 0 \ .$$

Nechť $s(\mathbf{x})$ je nějaká skóringová funkce definovaná na $\mathcal{X}$, G_1, G_2 jsou distribuční funkce veličiny $s(\mathbf{X})$ podmíněné jevem $[Y = 1]$ resp. $[Y = 2]$ a g_1, g_2

jsou odpovídající podmíněné hustoty, o kterých předpokládáme, že mají v nějakém intervalu $[c, d]$ první derivaci (v krajních bodech uvažujeme jednostranné derivace). Nechť ve vnitřních bodech tohoto intervalu dále platí

$$\pi_2(c_{1|2} - c_{2|2})g_2'(x) > \pi_1(c_{2|1} - c_{1|1})g_1'(x) \ . \tag{2.3}$$

Optimální prahový bod $\bar{a}^*$ minimalizující ztrátovou funkci L je pak implicitně určen vztahem:

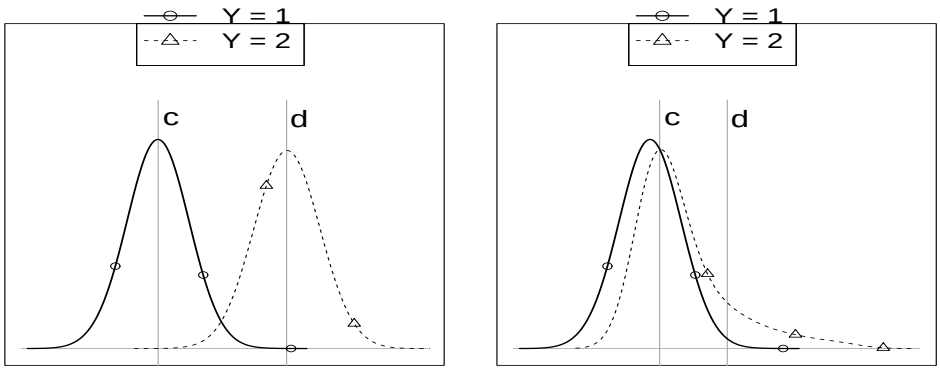
$$\frac{g_1(\bar{a}^*)}{g_2(\bar{a}^*)} = \frac{c_{1|2} - c_{2|2}}{c_{2|1} - c_{1|1}} \frac{\pi_2}{\pi_1} \ . \tag{2.4}$$

DŮKAZ Viz věta A.2 v dodatku A.

Všimněme si, že tato věta hovoří obecně o *jakékoliv* skóringové funkci. Nepříjemnou vlastností ovšem je, že podmíněné hustoty g_1, g_2 neumíme v mnoha případech analyticky vyjádřit. Otázkou by tedy bylo, na jakých odhadech těchto hustot založit výpočet. Ve skutečnosti může být totiž náhodná veličina $s(\mathbf{X})$ značně nespojitá. Dalším problémem je přesné určení nákladů, které nemusejí být nutně konstantní v čase, mohou se lišit objekt od objektu a v některých aplikacích (medicína) může být jejich stanovení více než diskutabilní.

Podmínka 2.3 zajišťuje, že v daném intervalu $[c, d]$ bude $\bar{a}^*$ skutečně *lokálním* minimem funkce L. Kdy je ale tato podmínka splněna? Uvedená věta navíc nehovoří o existenci ani jednoznačnosti minima. Z předpokladů vyplývá, že postačující podmínkou pro vztah 2.3 je: $g_2'(a) \geq 0$ a současně $g_1'(a) \leq 0$, kde alespoň jedna z nerovností je ostrá. Jinými slovy to znamená, že funkce g_2 je rostoucí (neklesající), zatímco funkce g_1 je nerostoucí (klesající) na intervalu $[c, d]$. Na obrázku 2.1 vlevo ilustrujeme průběh g_1, g_2 pro nějakou rozumnou skóringovou funkci včetně intervalu $[c, d]$, kde je tato postačující podmínka splněna. V případě $\pi_1 = \pi_2 = 0,5$ a standardizovaných nákladů je optimální bod $\bar{a}^*$ v průsečíku hustot g_1, g_2. Menší hodnota π_2 bude tento bod posouvat doprava, zatímco větší hodnota rozdílu $c_{1|2} - c_{2|2}$ jej posouvá doleva.

Předpokládejme nyní, že $c_{1|2} > c_{2|1}$ a $c_{1|1} = c_{2|2}$. To znamená, že zařazením objektu třídy 2 do třídy 1 utrpíme větší ztrátu, než kdybychom objekt z třídy 1 zařadili do třídy 2. Tento předpoklad odpovídá situaci v problému



Obrázek 2.1: Různé tvary podmíněných hustot g_1, g_2.

kreditního skórování. Podobně jako v testování hypotéz bychom mohli hovořit o chybném zatřídění I. druhu (utrpíme-li ztrátu $c_{1|2}$) a II. druhu (při ztrátě $c_{2|1}$).

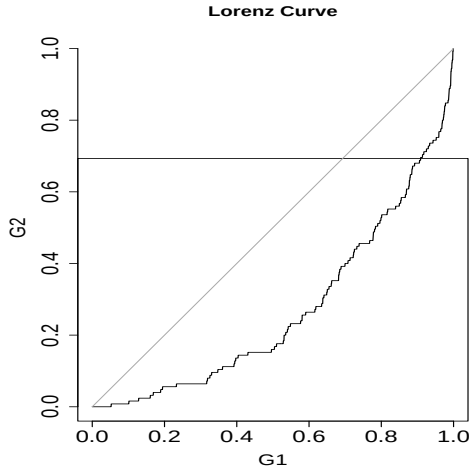
Jak ale uvidíme dále v textu, naše skóringové funkce budou mít spíše průběh podobný situaci na obrázku 2.1 vpravo. V pravém okolí průsečíku hustot g_1, g_2 zřejmě platí $g_1' < g_2' < 0$, přidáme-li další podmínku, pak na základě vlastností spojitých funkcí můžeme vyslovit následující tvrzení, které se pokusíme aplikovat v naší práci.

DŮSLEDEK 2.1 *Nechť je dána skóringová funkce s a nechť pro podmíněné hustoty náhodné veličiny $s(\mathbf{X})$ platí, že v jistém intervalu $[c, d]$ je $g_1 < g_2$ a existují konečné první derivace (v krajních bodech jednostranné) splňující $g_1' < g_2' < 0$. Nechť dále platí následující nerovnosti:*

$$\pi_2(c_{1|2} - c_{2|2}) < \pi_1(c_{2|1} - c_{1|1}) \ , \tag{2.5}$$

$$\frac{g_1(d)}{g_2(d)} \leq \frac{\pi_2(c_{1|2} - c_{2|2})}{\pi_1(c_{2|1} - c_{1|1})} \leq \frac{g_1(c)}{g_2(c)} \ . \tag{2.6}$$

Pak existuje právě jeden bod $\bar{a}^*$ minimalizující celkovou očekávanou ztrátu a je dán vztahem 2.4.



Obrázek 2.2: Lorenzova křivka modelu 4.2.

2.3 Porovnání skóringových funkcí

V této podkapitole uvedeme míry kvality klasifikačních pravidel založené na kvantifikaci očekávané ztráty (*míry nepřesnosti*) i kritéria založená na kvantifikaci shodnosti rozdělení skóringových funkcí pro objekty třídy 1 a 2 (*míry podobnosti*).

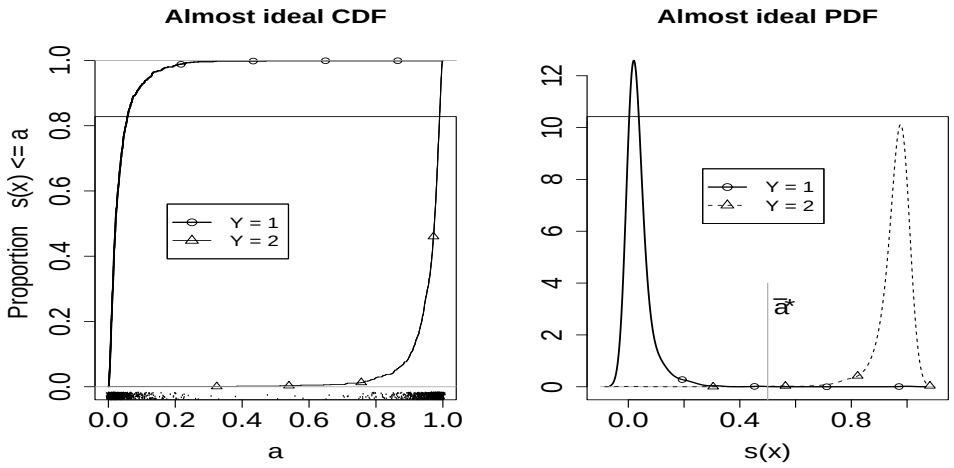
Velice rozšířeným kritériem pro porovnávání kvality klasifikačních pravidel je již dříve zmiňovaná míra chybovosti R. Jedná se o univerzální metodu, která má jasnou interpretaci. V odborné literatuře je však značně kritizována kvůli její citlivosti vůči volbě prahového bodu $\bar{a}$. Podobně je na tom i očekávaná ztráta, která ale navíc rozlišuje špatná zatřídění I. a II. druhu a přihlíží k apriorním pravděpodobnostem. Očekávanou ztrátu odhadujeme, stejně jako míru chybovosti, na testovacím vzorku.

Častou používanou explorativní technikou pro posouzení kvality skóringové funkce je Lorenzova křivka (Hand 1997).

DEFINICE 2.5 *Pro danou skóringovou funkci s definujeme Lorenzovu křivku jako následující množinu bodů z $\mathbb{R}^2$:*

$$LC(s) = \{[G_1(a), G_2(a)] \in [0,1]^2;\ a \in [0,1]\}.$$

Připomeňme, že $G_1(a), G_2(a)$ značí distribuční funkci náhodné veličiny $s(\mathbf{X})$ za podmínky, že $Y = 1$ a $Y = 2$. Pro ilustraci uvádíme na obrázku 2.2



Obrázek 2.3: Podmíněné distribuční funkce G_1, G_2 a hustoty g_1, g_2 pro nějakou hypotetickou, velice dobrou skóringovou funkci.

Lorenzovu křivku modelu 4.2, který popíšeme v podkapitole 4.4. Pro každou rozumnou skóringovou funkci je $G_1(a) \geq G_2(a),\ a \in [0,1]$, a proto Lorenzova křivka leží pod diagonálou. Pro ideální klasifikační pravidlo by Lorenzova křivka vycházela z počátku po ose x až do bodu $[1,0]$, odkud by vedla rovnoběžně s osou y do bodu $[1,1]$. Čím více se však tato křivka blíží diagonále, tím více se sobě vzájemně podobají rozdělení objektů obou tříd. Je zřejmé, že jejich rozlišení bude tím snazší, čím méně se jejich podmíněné hustoty budou překrývat. To ilustrujeme na obrázku 2.3, kde vidíme podmíněné distribuční funkce G_1, G_2 a hustoty g_1, g_2 pro nějakou hypotetickou skóringovou funkci, která by umožnila velice dobře rozlišit objekty třídy 1 a 2. Přímé srovnání hustot bývá ovšem zavádějící, neboť g_1 a g_2 většinou mívají různé nosiče.

Na obrázku 2.3 vpravo rovněž vidíme polohu optimálního prahového bodu $\bar{a}^*$ při standardizovaných nákladech. Zvolením menšího prahového bodu $\bar{a}$ bude klesat počet chybných přiřazení I. druhu, ale poroste počet chybných přiřazení II. druhu. Naopak větší hodnota $\bar{a}$ povede k nárůstu špatně zařazených objektů I. druhu a poklesu špatně zařazených objektů II. druhu.

Uveďme nyní některé míry podobnosti, které úzce souvisejí s Lorenzovou křivkou (Hand 1997).

DEFINICE 2.6 *Nechť G_1, G_2 jsou spojité distribuční funkce z definice 2.4, pak definujeme následující charakteristiky kvality skóringových funkcí:*

Supremální kritérium	$SC(s) = \sup_{a\in(0,1)} \lvert G_1(a) - G_2(a)\rvert$
Giniho koeficient	$GC(s) = 2\int_0^1 \lvert G_1(a) - G_2(a)\rvert dG_1(a)$
c statistika	$c(s) = P[s(\mathbf{X}_{K_1}) \leq s(\mathbf{X}_{K_2}) \lvert Y_{K_1} = 1, Y_{K_2} = 2]$, *kde* $K_1, K_2 \in \mathcal{K}$.

Tato kritéria jistým způsobem kvantifikují podobnost rozdělení náhodných veličin $s(\mathbf{X}_K)$ pro $Y_K = 1$ a $Y_K = 2$. Supremální koeficient je maximální rozdíl distribučních funkcí G_1 a G_2 ve vertikálním směru, Giniho koeficient představuje dvojnásobek plochy mezi distribučními funkcemi G_1 a G_2 vzhledem k míře G_1. A konečně c statistika je pravděpodobnost, že objekt třídy 1 bude mít nižší skóre než objekt třídy 2. Při hledání optimálního klasifikačního pravidla tudíž všechny tři statistiky maximalizujeme. Také si všimněme, že tyto charakteristiky jsou invariantní vůči spojitým monotónním transformacím skóringové funkce.

Různé vztahy vázající se k Lorenzově křivce, Giniho koeficientu a c-statistice shrnuje následující věta.

VĚTA 2.4 *Nechť jsou distribuční funkce G_1, G_2 spojité a $G_1(a) \geq G_2(a)$ pro každé $a \in [0,1]$. Pak platí:*

$$
\begin{aligned}
&(i)\quad GC = 2\int_0^1 \lvert G_1(a) - G_2(a)\rvert dG_2(a)\\
&(ii)\ GC = 2 \text{ plocha mezi } LC \text{ a diagonálou jednotkového čtverce,}\\
&(iii)\quad c = \frac{1}{2}(1 + GC).
\end{aligned}
\tag{2.7}
$$

DŮKAZ Viz Hand (1997).

Z této věty mimo jiné plyne, že $GC \in [0,1]$, $c \in [\frac{1}{2}, 1]$, a z definice 2.6 vyplývá: $SC \in [0,1]$. Na závěr ještě uveďme, že střední čtvercová chyba $\mathrm{E}[Y - s(\mathbf{X})]^2$ se pro porovnávání kvality skóringových funkcí nepoužívá, neboť není invariantní vůči monotónním transformacím funkce s (a z hlediska kreditního skórování je jakákoliv skóringová funkce stejně dobrá jako např. její odmocnina).

Kapitola 3

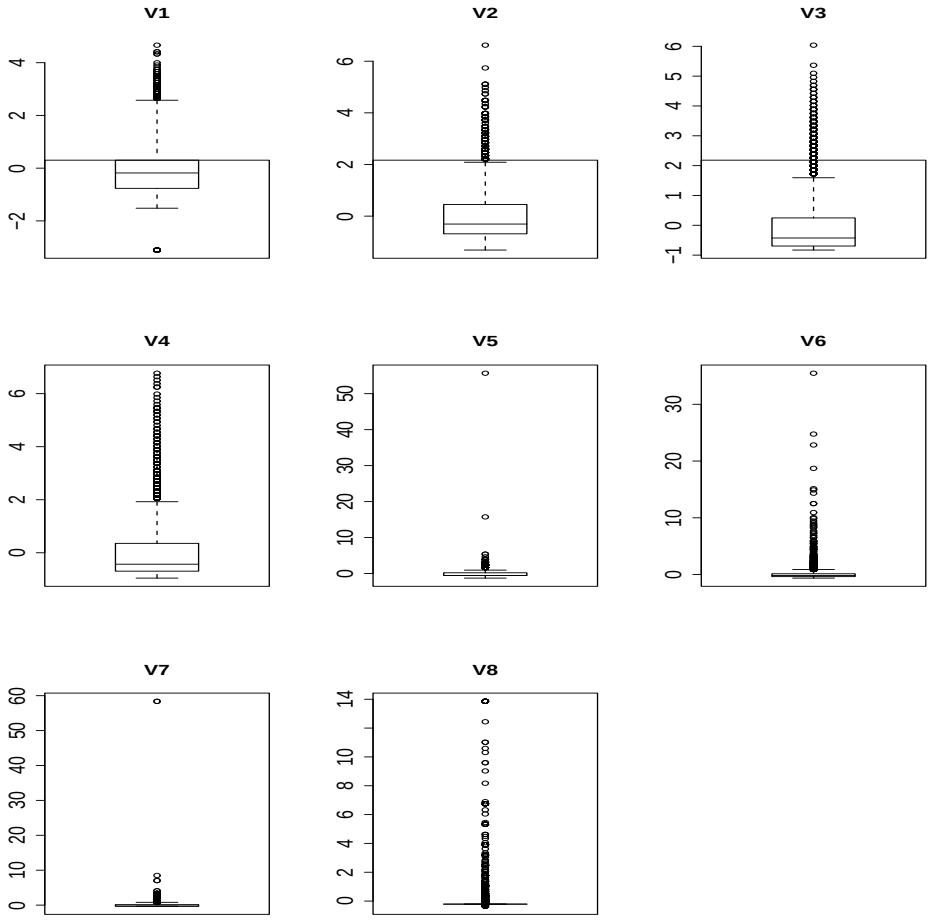
Popis dat

V této kapitole se budeme zabývat strukturou zkoumaných dat. Jelikož se jedná o téměř neupravená reálná data, budeme se muset vypořádat s celou řadou problémů spadajících do různých statistických disciplín. Nejprve popíšeme originální soubor a způsob jeho oříznutí včetně rozdělení na vývojový a testovací vzorek. Blíže se podíváme na jednotlivé proměnné, jejich strukturu a rozdělení v závislosti na hodnotě vysvětlované proměnné.

3.1 Původní datový soubor

Originální data pocházejí z jisté francouzské finanční společnosti a představují záznamy o klientech a jejich platební morálce (samotný datový soubor však pramení z databáze MD*BASE[1]), a proto můžeme předpokládat, že jednotlivá pozorování jsou na sobě nezávislá. Z nám dostupného komentáře pouze víme, že máme k dispozici osm spojitých a patnáct diskrétních nezávisle proměnných veličin. Závisle proměnná nabývá hodnoty 0 a 1, avšak zhruba u čtvrtiny pozorování byla její pravá hodnota změněna pro účely testování na hodnotu 9. Naší snahou bude zkonstruovat takové klasifikační pravidlo, které co nejlépe rozliší dvě skupiny pozorování lišící se hodnotou závisle proměnné veličiny. Používaná data jsou ovšem již odfiltrovaná, neboť část žadatelů o úvěr byla zamítnuta na základě expertního posudku dané banky. Z tohoto důvodu je náš datový soubor v jistém smyslu příliš homogenní, což samozřejmě přináší značné problémy.

[1] *http://www.quantlet.org/mdbase/*



Obrázek 3.1: Krabicové diagramy původních spojitých veličin.

Na rozdíl od předchozí kapitoly budeme značit jednotlivé kategorie jakékoliv L-hodnotové diskrétní veličiny symboly $0, \ldots, L-1$. Bohužel nevíme, jestli se jedná o diskrétní měřitelné, ordinální či nominální veličiny. Spojité náhodné veličiny byly z obavy o únik citlivých informací standardizovány. Kromě toho nám zůstal utajen význam všech proměnných, díky čemuž jsme ochuzeni o možnost interpretace výsledného modelu i o využití logických vazeb mezi jednotlivými proměnnými. Pro základní představu však můžeme uvést, že v praxi bývají zkoumanými veličinami vzdělání, věk, pohlaví, rodinný stav, počet dětí, roční příjem, historie účtů, délka pracovní činnosti, podíl ručitele na úvěru, pracovní sektor apod. (Kaiser & Szczesny 2000).

Proměnná	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
Pozorování $\notin [-3,3]$	60	90	129	107	22	61	16	92
Relativně	0,9%	1,3%	1,9%	1,6%	0,3%	0,9%	0,2%	1,4%

Tabulka 3.1: Počet pozorování podezřelých svými odlehlými hodnotami pro jednotlivé spojité veličiny z celkového počtu 492 odstraněných pozorování.

Pro jednoduchost zápisu označme závisle proměnnou náhodnou veličinu Y, spojité veličiny $V_1, \ldots, V_8$ a diskrétní veličiny $V_9, \ldots, V_{23}$. Nejprve odstraníme všechna pozorování s hodnotou $Y = 9$, neboť jsou pro účely naší analýzy nepoužitelná. Takto nám z původních 8830 zůstane jen 6672 pozorování a nadále budeme pracovat již jen s tímto menším datovým souborem.

Krabicové diagramy jednotlivých spojitých veličin jsou zachyceny na obrázku 3.1. Vidíme, že jejich rozdělení jsou více či méně šikmá. Některá pozorování bychom mohli na první pohled označit za odlehlá (maxima u proměnných V_5 či V_7), správně bychom však měli provést nějaký formální test na odlehlá pozorování. Testy navržené v knize Rönz (1998), jako je Grubbsův, David-Hartley-Pearsonův či Dixonův test, předpokládají normalitu, o níž se v našem případě zjevně opřít nemůžeme. Proto jsme se v této fázi rozhodli postupovat stejně, jako bylo navrženo v práci Härdle a kol. (2001): ořízneme pozorování, pro něž spojité veličiny vybočí z intervalu $[-3, 3]$. Takto získaná data byla rovněž použita v pracích Müller & Rönz (1999) či Komorád (2002), což nám později umožní srovnat dosažené výsledky. Tabulka 3.1 shrnuje údaje o počtu odstraněných pozorování.

Vidíme, že největší počet pozorování, která jsou větší než 3, resp. menší než -3, je u proměnné V_3 a sice 129, což představuje 1,9% procenta pozorování z celého datového souboru. Dohromady jsme takto vyloučili 492 (7,4%) pozorování. Z toho 421 pozorování nabylo hodnoty v absolutní hodnotě větší než tři u jedné spojité proměnné a 71 pozorování alespoň u dvou proměnných (tabulka 3.2).

Za zmínku stojí fakt, že u pozorování číslo 8819 – 8830 z originálních

Nevyhovujících proměnných	1	2	3	5	Celkem:
Počet vyřazených pozorování	421	59	11	1	492 (7,4%)

Tabulka 3.2: Vyřazená pozorování podle počtu nevyhovujících proměnných.

dat (která odpovídají posledním jedenácti pozorováním našeho souboru o 6672 pozorováních) nastává zvlášní situace: vždy alespoň pět spojitých veličin nabývá hodnoty svého minima a pro všechna tato pozorování je proměnná $V_1 = -3,106$ (což znamená, že tato pozorování byla odstraněna). Kromě toho se jejich identifikační číslo liší od ostatních o více než milion, takže usuzujeme, že tato pozorování pochází ze skupiny klientů se zcela jinou strukturou.

Po odstranění pozorování s podezřelými hodnotami spojitých veličin nám tedy pro analýzu zbude 6180 pozorování a všechna vyvinutá klasifikační pravidla budou fungovat pouze pro pozorování se spojitými veličinami z rozmezí $[-3, 3]$. Pozorování, která tuto podmínku nesplňují, by se pak musela řešit individuálně na základě speciálního expertního posudku.

Pro účely nezávislého srovnání různých modelů rozdělíme našich 6180 pozorování náhodně na dva podsoubory v poměru 2:1. Vývojový vzorek $\mathcal{V}_1$ bude obsahovat 4135 pozorování a testovací vzorek $\mathcal{V}_2$ jich bude mít 2045. Rozdělení v tomto poměru se drží návrhu uvedeného v Breimann a kol. (1984), nemá však žádné teoretické opodstatnění. V následujících třech podkapitolách provedeme bližší analýzu vývojového vzorku.

3.2 Nezávisle proměnná

K určení významu kódování této proměnné nám dobře poslouží tabulka četností 3.3. Na základě těchto údajů usuzujeme, že kategorie $Y = 0$ označuje klienty, kteří neměli problémy se splácením poskytnutého úvěru. Naproti

Kategorie	Četnost	Relativně
0	3888	94,0%
1	247	6,0%

Tabulka 3.3: Četnosti nezávisle proměnné veličiny Y.

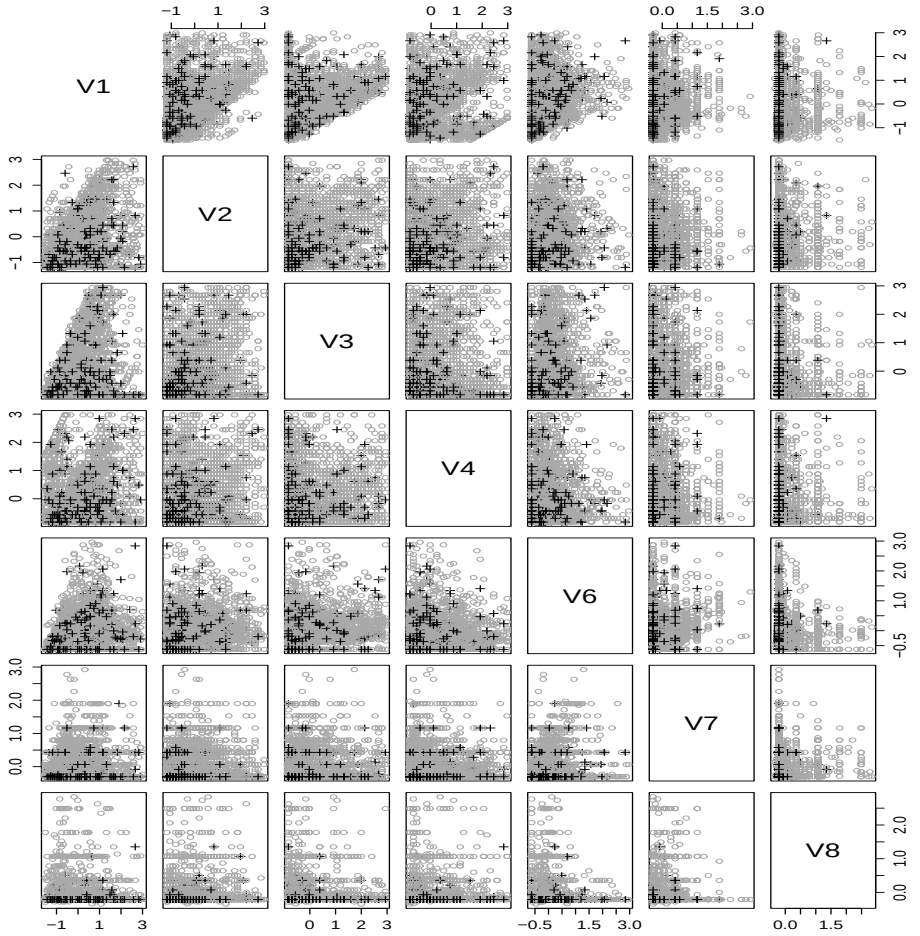
tomu do kategorie $Y = 1$ byli pravděpodobně zařazeni ti klienti, kteří upadli v platební neschopnost. V tom se náš datový soubor shoduje s běžnou mírou nesplácejících klientů, která kolísá mezi jedním až sedmi procenty (Arminger a kol. 1997). Ze statistického hlediska se však jedná o velkou nevyváženost, která je pro tento typ úloh bohužel charakteristická, a která s sebou přináší značné potíže, jak ještě uvidíme v následujících kapitolách.

3.3 Spojité veličiny

Abychom si přiblížili strukturu spojitých veličin vývojového vzorku, podívejme se na obrázky B.1 až B.8 v dodatku B. Ty zobrazují krabicové diagramy, jádrové odhady hustoty pro spolehlivé a nespolehlivé klienty a příslušné normální diagramy. V tabulce pod každým z obrázků jsou uvedeny odpovídající hodnoty extrémů, kvartilů, mediánu a směrodatné odchylky pro populaci dobrých a špatných klientů. Srovnání rozdělení těchto populací můžeme vidět v horních dvou grafech. Levý krabicový diagram odpovídá spolehlivým klientům ($Y = 0$), pravý pak zobrazuje klienty nespolehlivé ($Y = 1$). Jestli se tyto skupiny liší ve střední hodnotě, otestujeme formálně pomocí Wilcoxonova testu. Pro každou náhodnou veličinu testujeme hypotézu rovnosti středních hodnot skupiny spolehlivých a nespolehlivých klientů proti oboustranné alternativě. Výsledky shrnuje tabulka 3.4. Dvě skupiny klientů se na pětiprocentní hladině liší v proměnných V_1, V_2, V_7 a V_8. Výše zmíněné obrázky tyto výsledky ilustrují, nicméně z krabicových diagramů na obrázcích B.7 a B.8 nepoznáme nic, neboť rozdělení veličin V_7 a V_8 je příliš šikmé. Při pohledu na obrázek B.5 si všimneme, že veličina V_5 je diskrétní a nabývá pouze pěti různých hodnot. Při sběru dat zřejmě došlo k omylu a tato veličina byla nejspíše chybně označena za spojitou. V naší práci s ní budeme zacházet jako s diskrétní veličinou a jakou takovou ji budeme značit V_{24} (její kategorie postupně označíme $0, \ldots, 4$). Při bližším ohledání zjistíme, že i proměnné V_7 a V_8 mají jisté náznaky výskytu bodů nespojitosti. Tento dojem je ještě umocněn při pohledu na párový diagram (scatterplot) na obrázku 3.2, kde je zachycen vliv dvojic spojitých veličin na hodnotu závisle proměnné. Šedivá kolečka označují pozorování s hodnotou $Y = 0$, černé křížky příslušejí pozorováním, kde $Y = 1$. Povšimněme si, že Y je rovno jedné spíše u pozorování s nižšími hodnotami veličin V_7 a V_8, což by mohlo indikovat větší význam těchto proměnných pro naši analýzu. Jádrové odhady hustot f_0 pro spolehlivé klienty (šedivě) a f_1 pro klienty nespolehlivé (černě) na obrázcích

Proměnná	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
Statistika $\times 10^3$	441,2	548,9	508,5	458,0	465,3	501,1	565,7	530,3
P-hodnota	0,032	0,000	0,118	0,220	0,345	0,248	0,000	0,000

Tabulka 3.4: Testové statistiky a p-hodnoty pro Wilcoxonův test.



Obrázek 3.2: Párový diagram spojitých veličin.

v dodatku nám dávají tušit, že nebude snadné tyto dvě skupiny od sebe oddělit, neboť jednorozměrná rozdělení spojitých veličin $V_1, \ldots, V_8$ si jsou velice podobná.

Normální diagramy nenaznačují shodu s normálním rozdělením, což formálně otestujeme pomocí Shapiro-Wilkova testu. Ten však zamítá nulovou hypotézu na 5% hladině pro spolehlivé i nespolehlivé klienty u všech spojitých veličin. Protože jádrový odhad hustoty svým tvarem připomíná spíše logaritmicko-normální rozdělení (zejména u veličiny V_4), zkusíme tuto hypotézu otestovat a použijeme Kolmogorov-Smirnovův test. Shodnost s log-normálním rozdělením ale není na 5% hladině průkazná pro žádnou z veličin.

Pro úplnost dodejme, že spojité veličiny nevykazují výraznější známky korelovanosti. Nejextrémnější hodnota korelačního koeficientu je 0,43 mezi veličinami V_1 a V_2. (celá korelační matice je uvedena v tabulce B.1 v dodatku).

3.4 Diskrétní veličiny

Jelikož neznáme typ jednotlivých diskrétních veličin, budeme se všemi formálně zacházet jako s nominálními veličinami. Tím se nedopustíme žádné chyby, naše výsledky ovšem mohou být slabší v případě, že některá z veličin je ordinální či dokonce měřitelná. Přehled absolutních a relativních četností jednotlivých kategorií pro veličiny V_9–V_{24} je uveden v tabulkách C.1, C.2 a C.3 v dodatku. Tamtéž jsou na obrázcích C.1, C.2 a C.3 zachyceny sloupcové diagramy těchto proměnných. Povšimněme si, že struktura klientů s dobrou platební morálkou ($Y = 0$) a klientů v platební neschopnosti ($Y = 1$) se nijak zvlášť neliší, přesto se nám může zdát, že některé kategorie by mohly mít větší vliv. Můžeme kupříkladu nabýt dojmu, že má-li nějaké pozorování hodnotu proměnné $V_{19} = 1$, bude se spíše jednat o klienta v platební neschopnosti. Obdobný vliv by mohly mít i kategorie 7 veličiny V_{22} či kategorie 2 a 8 proměnné V_{23}.

Abychom viděli přesný vliv jednotlivých kategorií na hodnotu nezávisle proměnné, spočteme pravděpodobnosti, že se jedná o nespolehlivého klienta za podmínky, že $i-$tá proměnná nabývá kategorie k, tj. podmíněné pravděpodobnosti $\mathrm{P}(Y = 1|V_i = k)$. K výpočtu využijeme Bayesovu větu. Výsledky jsou uvedeny v tabulce 3.5, kde vidíme, že naší pozornosti unikla kategorie 4 u proměnné V_{20}, která má markantní vliv na naši analýzu: klient s hodnotou veličiny $V_{20} = 4$ bude s pravděpodobností 23,3% nespolehlivý. Můžeme čekat, že čím vzdálenější je podmíněná pravděpodobnost $\mathrm{P}(Y = 1|V_i = k)$ od hodnoty nepodmíněné pravděpodobnosti $\mathrm{P}(Y = 1) = 6\%$ (viz tabulka 3.3), tím je pro nás kategorie k veličiny V_i (a tím i celá veličina) významnější.

3.5 Ztrátová funkce a vyváženost dat

Ztrátovou funkci jsme zavedli na základě následující úvahy: v polovině devadesátých let, tedy v době, ze které pocházejí naše data, se úroková míra

V_i	k	%	V_i	k	%	V_i	k	%	V_i	k	%
V_9:	0	8,3	V_{17}:	0	5,6		2	4,0		7	0,9
	1	5,1		1	5,4		3	7,8		8	10,5
V_{10}:	0	13,4		2	5,2		4	23,3		9	5,3
	1	4,4		3	6,7		5	4,1	V_{23}:	0	7,2
V_{11}:	0	6,9	V_{18}:	0	4,1	V_{21}:	0	4,1		1	5,4
	1	5,8		1	7,0		1	3,9		2	3,3
V_{12}:	0	5,5		2	5,2		2	4,7		3	5,6
	1	6,7		3	9,5		3	10,4		4	4,0
V_{13}:	0	8,8		4	7,5		4	6,9		5	4,8
	1	5,8		5	6,1		5	5,3		6	7,8
V_{14}:	0	5,8	V_{19}:	0	4,6		6	7,6		7	6,2
	1	10,6		1	8,4	V_{22}:	0	4,2		8	9,8
V_{15}:	0	4,1		2	5,9		1	3,8		9	1,5
	1	4,9		3	8,6		2	7,9		10	7,5
	2	6,9		4	4,7		3	3,0	V_{24}:	0	5,7
V_{16}:	0	5,9		5	5,8		4	7,6		1	6,8
	1	7,5	V_{20}:	0	10,7		5	5,4		2	4,8
	2	4,9		1	7,5		6	9,6		3	8,1
										4	9,5

Tabulka 3.5: Podmíněné pravděpodobnosti $\mathrm{P}(Y = 1|V_i = k)$.

spotřebitelských úvěrů[2] pohybovala mezi 16 a 17%. V případě, že akceptovaný klient upadne v platební neschopnost, utrpí věřitel ztrátu, která se rovná přibližně šestinásobku toho, co by vydělal, kdyby klient svůj závazek splatil ($100/16 \doteq 6$, tato úvaha však nebere v potaz časovou cenu peněz). V případě zamítnutí spolehlivého klienta utrpí banka ztrátu rovnou tomu, co by vydělala při jeho akceptaci. Zamítnutím nespolehlivého klienta neutrpí banka žádnou ztrátu, ani nic nevydělá. Proto volíme ztrátovou funkci, která se dá zapsat do následující matice:

$$C = \begin{pmatrix} c_{1|1} & c_{1|2} \\ c_{2|1} & c_{2|2} \end{pmatrix} = \begin{pmatrix} -1 & 6 \\ 1 & 0 \end{pmatrix} . \tag{3.1}$$

Snadno spočítáme, že podmínka 2.5 z důsledku 2.1 bude v takovém případě splněna.

[2] www.banque-france.fr/gb/stat

V následujících kapitolách uvidíme, že naše klasifikační pravidla chybně zařazují až příliš velké procento nespolehlivých klientů. Jeden z možných důvodů je skutečnost, že ve vývojovém vzorku jsou tito klienti zastoupeni ve velice malém počtu. Bohužel nepomohlo ani zavedení vah. Např. jejich použitím u klasifikačních stromů se buď rapidně snižoval počet uzlů stromu, až všechna pozorování skončila v kořeni, nebo bylo naopak dosaženo maximální povolené hloubky stromu (32) a výpočetní software skončil chybou. Metoda nejbližších sousedů (procedura `knn` z balíku `class`) použití vah dokonce vůbec neumožňuje. Proto jsme se pokusili překonat problém nevyváženosti dat a malého zastoupení objektů třídy 1 vytvořením náhradního vývojového vzorku. Do něj jsme zařadili všechny nespolehlivé klienty z $\mathcal{V}_1$, tj. 247 pozorování, a k nim náhodně vybrali nejprve 988 a posléze 494 spolehlivých klientů. Tím jsme získali vývojový vzorek, kde byl poměr špatných vůči dobrým klientům 1:4, resp. 1:2. Dobrovolně jsme se tak vzdali značného množství informací o spolehlivých klientech. Cena za větší vyváženost dat byla však příliš vysoká. Dosáhli jsme sice nižšího počtu chybně označených klientů I. druhu, značně se ale zvýšil počet chybně označených klientů II. druhu, díky čemuž se očekávaná ztráta téměř nezměnila. Giniho koeficient a supremální kritérium byly dokonce nepatrně nižší. Proto jsme nakonec zůstali u původního vývojového vzorku, který zachovává reálný poměr spolehlivých a nespolehlivých klientů. Můžeme však tvrdit, že třídy spolehlivých a nespolehlivých klientů se značně prolínají a nejsou námi používanými vysvětlujícími proměnnými dostatečně dobře rozlišitelné.

3.6 Doplňující poznámky

V praktickém kreditním skórování se vedle oklasifikování jednotlivých subjektů určuje explicitně i pravděpodobnost, že daný subjekt patří do třídy 1, tj. že se jedná o špatného klienta (tzv. pravděpodobnost *defaultu*). Z podstaty problému známe hodnotu závisle proměnné Y pouze u těch klientů, jejichž žádost o úvěr byla akceptována. U zamítnutých klientů tato informace přirozeně chybí. Jelikož žádosti nejsou zamítány náhodně, je zřejmé, že odhad pravděpodobnosti defaultu založený pouze na datech akceptovaných žádostí, nebude nestranný. Odhad pravděpodobnosti defaultu na základě zamítnutých dat se v odborné literatuře označuje jako *reject inference*.

Kapitola 4

Zobecněný lineární model

V této kapitole nejprve stručně shrneme základy teorie zobecněných lineárních modelů, abychom si vytvořili vhodné prostředí pro využití jedné z nejčastěji používaných metod pro kreditní skórování – logistického regresního modelu. Následující dvě podkapitoly přináší pouze základní přehled zobecněných lineárních modelů a logistické regrese. Vycházejí z poznatků uvedených v knihách Zvára (2003) a Härdle a kol. (2000), kde je možné nalézt podrobnější vysvětlení, případně odkaz na další literaturu. V dalších podkapitolách odhadneme model obsahující všechny proměnné bez interakcí, vybereme jednodušší podmodel, podíváme se na modely s interakcemi a použijeme rovněž zobecněný aditivní model pro přiblížení tvaru spojitých vysvětlujících veličin. Na závěr srovnáme vhodnost volby jednotlivých modelů na testovacím vzorku.

4.1 GLM a přehled označení

Mějme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, $X_{n\times r}$ danou matici a β_0, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top$ neznámé parametry. Předpokládejme, že náhodné veličiny Y_1, $\ldots$, Y_n jsou nezávislé a jejich rozdělení závisí prostřednictvím parametrů β_0, $\boldsymbol{\beta}$ na známých vektorech $\mathbf{x}_{k\bullet} = (x_{k1}, \ldots, x_{kr})$, $k = 1, \ldots, n$. Matice

$$(\mathbf{1}, X_{n\times r}) = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1r} \\ 1 & x_{21} & \ldots & x_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nr} \end{pmatrix}$$

má hodnost d. Pod pojmem *zobecněný lineární model* (generalized linear model – GLM) rozumíme takový model, ve kterém platí:

1. Podmíněnou střední hodnotu veličin $Y_1, \ldots, Y_n$ lze zapsat ve tvaru:

$$\mathrm{E}[Y_k|\mathbf{x}_{k\bullet}] = G(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_{k\bullet}) \ ,$$

   kde G je známá ryze monotónní *spojovací funkce*[1] se spojitou druhou derivací,

2. Y_k má rozdělení exponenciálního typu[2], tj. jeho hustotu lze zapsat ve tvaru:

$$f_k(y_k, \theta_k, \phi) = \exp\left\{\frac{y_k\theta_k - b(\theta_k)}{a(\phi)} + c(y_k, \phi)\right\},$$

   kde $c(y, \phi)$ nezávisí na parametru θ, $a(\phi) > 0$ a $b(\theta)$ má spojitou druhou derivaci.

Symbolem μ_k označíme podmíněnou střední hodnotu $\mathrm{E}[Y_k|\mathbf{x}_{k\bullet}]$, η bude značit součet $\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$, takže spojovací funkce vyjadřuje vzah mezi μ a η: $\mu = G(\eta)$. Ke každému rozdělení exponenciálního typu závisle proměnné existuje jedna speciální spojovací funkce, pro kterou platí $\eta = \theta$ a kterou nazýváme *kanonický link*. Parametr ϕ většinou bývá pouze rušivý, parametr θ označujeme jako kanonický. Předmět našeho zájmu, tedy parametr $\boldsymbol{\beta}$ odhadujeme metodou maximální věrohodnosti. K řešení věrohodnostní rovnice

$$\frac{\partial \log L(\mathbf{Y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{k=1}^{n} \frac{\partial \ell(Y_k, \theta_k, \phi)}{\partial \boldsymbol{\beta}} = 0$$

se používají iterativními postupy, např. Newton-Raphsonův algoritmus či Fisherův skórový algoritmus. Výsledný odhad $\hat{\boldsymbol{\beta}}$ je nestranný a má asymptoticky normální rozdělení.

4.2 Logistická regrese

Má-li závisle proměnná veličina alternativní rozdělení a je-li spojovací funkce dána vztahem:

$$\mathrm{E}[Y|\mathbf{x}] = G(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}) = \frac{\exp\{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}\}}{1 + \exp\{\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}\}} \ , \tag{4.1}$$

[1] Někdy se jako spojovací funkce označuje G^{-1}.

[2] Exponenciálního typu je např. binomické, poissonovo, normální či gamma rozdělení.

hovoříme o logistickém regresním modelu. Funkci $\eta(\mu) = \log \frac{\mu}{1-\mu}$ říkáme *logit*, tato funkce je kanonickým linkem pro závisle proměnnou s alternativním rozdělením. Pro nula-jedničkovou vysvětlovanou veličinu se někdy též jako spojovací funkce používá distribuční funkce standardního normálního rozdělení. Takový model pak nazýváme probitovým.

Maximální hodnotu logaritmu věrohodnostní funkce v saturovaném modelu budeme značit symbolem ℓ_{max}. Pro běžný model, jehož odhad parametrů je $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, definujeme vztahem:

$$D(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = 2(\ell_{max} - \ell(\hat{\beta}_0, \hat{\boldsymbol{\beta}}))$$

takzvanou *devianci*, která se používá pro testování podmodelů. Pro diagnostické účely se v logistické regresi nejčastěji používají takzvaná *devianční rezidua*, která jsou definována vztahem:

$$u_k^{dev} = \sqrt{-2\left(Y_k \log \hat{\mu}_k + (1 - Y_k) \log(1 - \hat{\mu}_k)\right)}\, \mathrm{sign}(Y_k - \hat{\mu}_k), \tag{4.2}$$

kde $\hat{\mu}_k$ je odhad pravděpodobnosti $\mu_k = \mathrm{P}[Y_k = 1|\mathbf{x}_{k\bullet}] = \mathrm{E}[Y_k|\mathbf{x}_{k\bullet}]$.

Harrell (2002) navrhuje pro kvantifikování kolinearity jednotlivých regresorů (zobecněného) lineárního modelu použít charakteristiku známou pod jménem variance inflation factor (VIF), tedy jakýsi faktor navýšení rozptylu. Ten se počítá ze vztahu $VIF_j = \frac{1}{1-r_j^2}$, kde r_j^2 je čtverec výběrového koeficientu mnohonásobné korelace mezi j-tým sloupcem matice X a zbylými sloupci. Jinými slovy VIF říká, kolikrát se zhorší rozptyl odhadu koeficientu j-tého regresoru z důvodu jeho korelovanosti s ostatními regresory (ideální by tedy bylo VIF rovné 1).

Pro posouzení vlivu jednotlivých pozorování použijeme takzvanou *Cookovu vzdálenost*. Označme $\hat{\boldsymbol{\beta}}_{(t.)}$ odhad vektoru parametrů ve výše zmíněném modelu, avšak s vynecháním t-tého pozorování, a nechť $\hat{Y}_{(t.)}$ značí odhad $\hat{\mathbf{Y}}_{(t.)} = G\left(\hat{\beta}_{0(t.)} + X^\top \hat{\boldsymbol{\beta}}_{(t.)}\right)$. Cookova vzdálenost je definována vztahem:

$$D_C = \frac{1}{rS^2} ||\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{t}.)}||^2 \,,$$

kde r je počet sloupců matice X a S^2 je reziduální rozptyl.

V našem označení bude tedy $\mathbf{x}_{k\bullet}$ značit vektor pozorování k-tého klienta, tj. $\mathbf{x}_{k\bullet} = (V_{k,1}, \ldots, V_{k,4}, V_{k,6}, \ldots, V_{k,24})$, $X = (\mathbf{x}_{1\bullet}^\top, \ldots, \mathbf{x}_{4135\bullet}^\top)^\top$ bude matice dimenze 4135×23 a $\mathbf{Y} = (Y_1, \ldots, Y_{4135})^\top$ bude vektor napozorvaných hodnot závisle proměnné. Dále $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4, \beta_6, \ldots, \beta_{24})^\top$ je vektor neznámých parametrů a β_0 je neznámá regresní konstanta.

Veličina	D.F.	Dev.	AIC
V_{10}	1	1802,6	1806,6
V_{20}	5	1806,4	1818,4
V_{22}	9	1826,8	1846,8
V_{23}	10	1833,4	1855,4
V_{21}	6	1843,0	1857,0
V_7	1	1851,2	1855,2

Tabulka 4.1: Proměnné s nejnižšími hodnotami deviance a Akaikeho informačního kritéria při jednoduché regresi. Druhý sloupec udává počet stupňů volnosti.

4.3 Všechny proměnné bez interakcí

Na úvod ukážeme výsledky jednoduché regrese. Tabulka 4.1 shrnuje veličiny, jejichž deviance a *Akaikeho informační kritérium* (AIC)[3] dosáhly nejnižších hodnot. Tato tabulka neříká vlastně nic zásadního, hodnoty Akaikeho kritéria můžeme později porovnat s hodnotami AIC složitějších modelů, povšimněme si ale, že čelní místa obsadily samé diskrétní veličiny (až V_7 je spojitá).

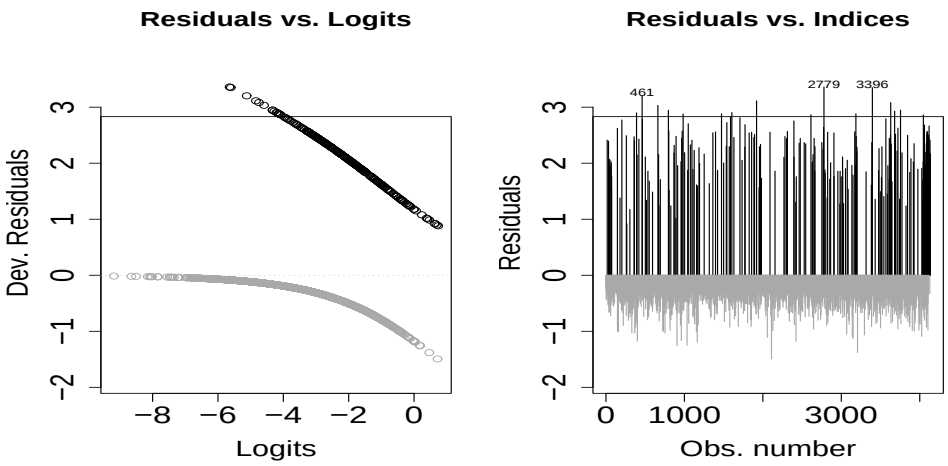
Pro model obsahující všechny proměnné (bez interakcí) uvádíme v tabulce 4.2 odhady koeficientů, jejich standardní chyby a hladiny spolehlivosti. Na tento model budeme v dalším odkazovat jako na model 4.1. Jedna hvězdička značí koeficienty signifikantní na 5% hladině, dvě hvězdičky na 1% a tři hvězdičky jsou signifikantní na 0,1% hladině. Pro diskrétní veličiny je jako referenční kategorie vzata první kategorie (s hodnotou 0), takže např. $V_{20}3$ značí čtvrtou kategorii (s hodnotou 3) proměnné V_{20}. Reziduální deviance tohoto modelu je rovna 1560,3 na 4070 stupních volnosti, Akaikeho informační kritérium (AIC) nabývá hodnoty 1690. Z tabulky vyčteme, že koeficienty jsou pouze u 14 proměnných (každou indikátorovou veličinu počítáme zvlášť) signifikantně odlišné od nuly na 5% hladině. Pro výběr vhodného podmodelu použijeme krokovou regresi, kterou popíšeme v následující podkapitole.

Obrázek 4.1 vlevo zobrazuje rezidua proti logitům, tj. proti hodnotám $G^{-1}(\hat{Y}) = \log \frac{\mu}{1-\mu}$, kde G je spojovací funkce ze vztahu 4.1. Vpravo pak vidíme rezidua jednotlivých pozorování. V obou případech se jedná o devianční rezidua definovaná vztahem 4.2. Šedivá rezidua odpovídají spolehlivým kli-

[3] AIC $= D(\mathbf{b}) + 2$(počet parametrů)

	$\hat{\beta}$	Std.Er.	p		$\hat{\beta}$	Std.Er.	p
Icpt	−2,511	0,730	0,001***	$V_{20}3$	−0,281	0,285	0,323
V_1	0,204	0,129	0,115	$V_{20}4$	1,082	0,385	0,005**
V_2	−0,326	0,101	0,001**	$V_{20}5$	−1,012	0,176	0,000***
V_3	−0,058	0,106	0,582	$V_{21}1$	0,052	0,334	0,877
V_4	0,038	0,100	0,705	$V_{21}2$	0,150	0,272	0,581
V_6	0,042	0,182	0,817	$V_{21}3$	0,842	0,257	0,001**
V_7	−0,891	0,223	0,000***	$V_{21}4$	0,392	0,238	0,099
V_8	−0,912	0,409	0,026*	$V_{21}5$	0,338	0,293	0,249
$V_9$1	−0,311	0,176	0,077	$V_{21}6$	0,646	0,313	0,039*
$V_{10}1$	−0,954	0,155	0,000***	$V_{22}1$	−0,224	0,408	0,582
$V_{11}1$	0,019	0,212	0,930	$V_{22}2$	0,493	0,318	0,121
$V_{12}1$	0,237	0,166	0,152	$V_{22}3$	−0,411	0,541	0,447
$V_{13}1$	−0,494	0,274	0,072	$V_{22}4$	0,522	0,419	0,213
$V_{14}1$	0,767	0,301	0,011*	$V_{22}5$	−0,090	0,456	0,844
$V_{15}1$	−0,110	0,438	0,802	$V_{22}6$	0,712	0,426	0,095
$V_{15}2$	0,248	0,229	0,279	$V_{22}7$	−1,453	0,732	0,047*
$V_{16}1$	0,016	0,218	0,940	$V_{22}8$	0,652	0,553	0,238
$V_{16}2$	−0,287	0,241	0,235	$V_{22}9$	0,119	0,356	0,739
$V_{17}1$	−0,020	0,362	0,957	$V_{23}1$	−0,347	0,377	0,357
$V_{17}2$	−0,244	0,282	0,386	$V_{23}2$	−0,631	0,342	0,065
$V_{17}3$	0,089	0,186	0,633	$V_{23}3$	−0,035	0,409	0,931
$V_{18}1$	0,372	0,239	0,119	$V_{23}4$	−0,501	0,437	0,252
$V_{18}2$	0,363	0,291	0,212	$V_{23}5$	−0,512	0,585	0,382
$V_{18}3$	0,725	0,318	0,023*	$V_{23}6$	−0,071	0,349	0,839
$V_{18}4$	0,342	0,437	0,433	$V_{23}7$	−0,195	0,338	0,564
$V_{18}5$	0,909	0,281	0,001**	$V_{23}8$	0,201	0,296	0,495
$V_{19}1$	0,537	0,310	0,083	$V_{23}9$	−1,542	1,036	0,137
$V_{19}2$	0,023	0,309	0,941	$V_{23}10$	0,139	0,322	0,666
$V_{19}3$	0,564	0,363	0,120	$V_{24}1$	0,305	0,225	0,176
$V_{19}4$	−0,132	0,311	0,671	$V_{24}2$	0,001	0,265	0,998
$V_{19}5$	0,244	0,319	0,445	$V_{24}3$	0,566	0,295	0,055
$V_{20}1$	−0,370	0,284	0,193	$V_{24}4$	0,540	0,513	0,293
$V_{20}2$	−0,892	0,318	0,005**				

Tabulka 4.2: Odhady koeficientů, jejich standardní chyby a hladiny spolehlivosti proměnných modelu 4.1. Icpt značí absolutní člen.



Obrázek 4.1: Struktura reziduí modelu 4.1.

entům, černá pak nespolehlivým klientům. Díky nevyváženosti dat je model přiléhavější pozorováním, která odpovídají spolehlivým klientům. Rezidua u_k^{dev} těch pozorování, kde $Y_k = 1$ jsou až několikanásobně větší než rezidua pro $Y_k = 0$.

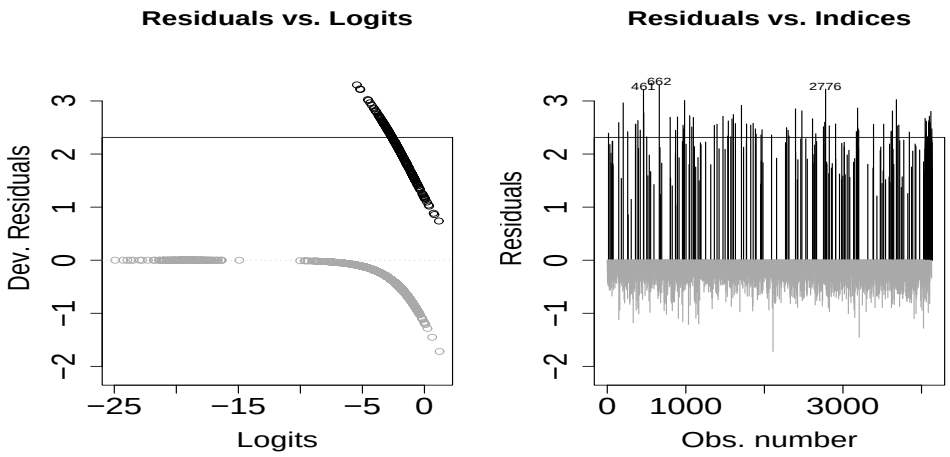
4.4 Kroková regrese bez interakcí

Nyní se ještě jednou podíváme na model bez přítomnosti vzájemných interakcí mezi vysvětlujícími proměnnými a pokusíme se jej zjednodušit. K nalezení vhodného modelu použijeme obousměrnou krokovou regresi s využitím Akaikeho informačního kritéria. Ve výsledném modelu je navrženo ponechat proměnné V_1, V_2, V_7 – V_{10}, V_{12} – V_{15} a V_{18} – V_{22}. Reziduální deviance tohoto modelu dosáhla hodnoty 1587,2 se 4093 stupni volnosti, AIC je rovno 1671,2.

Maximální hodnota VIF tohoto modelu je 4,53 (u koeficientu $V_{22}2$), což ale nepovažujeme za extrémně vysoké číslo, a veličinu V_{22} v analýze ponecháme.

Odstranili jsme 6 pozorování s největší hodnotou Cookovy vzdálenosti, a získali tak model s deviancí 1522,0 na 4087 stupních volnosti a AIC rovným 1606. Odhady parametrů pro tento výsledný model, který budeme dále značit 4.2, jejich standardní chyby a hladiny spolehlivosti jsou uvedeny v tabulce 4.3.

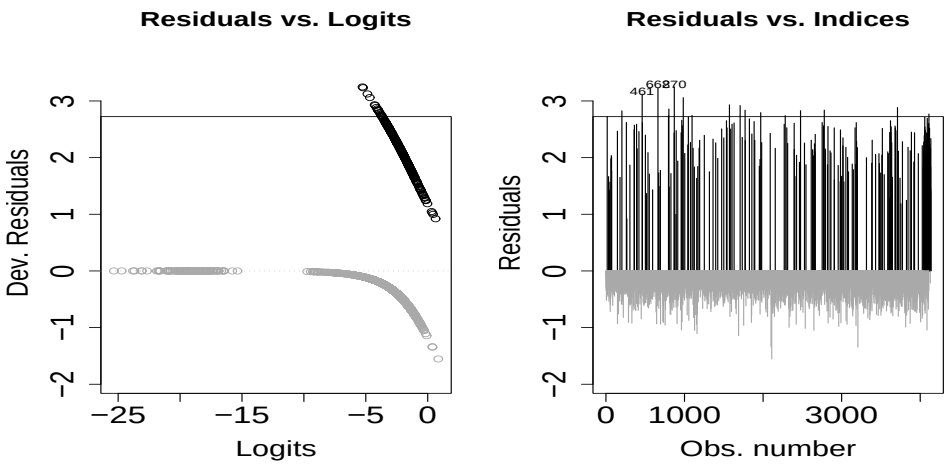
Obrázek 4.2 ukazuje devianční rezidua proti logitům a rezidua jednotlivých pozorování. Všimněme si, že oproti modelu 4.1 se zde oddělila skupina



Obrázek 4.2: Struktura reziduí modelu 4.2.

	$\hat{\beta}$	Std.Er.	p		$\hat{\beta}$	Std.Er.	p
Icpt	−3,071	0,616	0,000***	$V_{19}5$	0,347	0,324	0,284
V_1	0,219	0,120	0,067.	$V_{20}1$	−0,380	0,282	0,178
V_2	−0,390	0,101	0,000***	$V_{20}2$	−0,981	0,321	0,002**
V_7	−1,019	0,236	0,000***	$V_{20}3$	−0,387	0,288	0,179
V_8	−1,789	0,630	0,005**	$V_{20}4$	1,113	0,383	0,004**
$V_9$1	−0,375	0,176	0,033*	$V_{20}5$	−1,022	0,175	0,000***
V_{10}1	−1,034	0,155	0,000***	$V_{21}1$	−0,016	0,341	0,962
V_{12}1	0,284	0,165	0,085.	$V_{21}2$	0,024	0,277	0,931
V_{13}1	−0,480	0,278	0,085.	$V_{21}3$	0,853	0,256	0,001***
V_{14}1	0,736	0,299	0,014*	$V_{21}4$	0,359	0,237	0,130
V_{15}1	0,279	0,349	0,424	$V_{21}5$	0,329	0,292	0,260
V_{15}2	0,617	0,177	0,001***	$V_{21}6$	0,623	0,312	0,046*
V_{18}1	0,350	0,239	0,142	$V_{22}1$	−0,033	0,394	0,932
V_{18}2	0,256	0,277	0,355	$V_{22}2$	0,617	0,305	0,043*
V_{18}3	0,555	0,312	0,075.	$V_{22}3$	−0,435	0,558	0,435
V_{18}4	0,334	0,426	0,433	$V_{22}4$	0,689	0,409	0,092.
V_{18}5	0,652	0,241	0,007**	$V_{22}5$	0,112	0,422	0,790
V_{19}1	0,647	0,314	0,039*	$V_{22}6$	0,902	0,408	0,027*
V_{19}2	0,224	0,308	0,467	$V_{22}7$	−1,507	0,397	0,097
V_{19}3	0,680	0,367	0,064.	$V_{22}8$	0,501	0,540	0,354
V_{19}4	−0,054	0,314	0,863	$V_{22}9$	0,306	0,341	0,368

Tabulka 4.3: Odhady koeficientů, jejich standardní chyby a hladiny spolehlivosti proměnných modelu 4.2. Icpt značí absolutní člen.



Obrázek 4.3: Struktura reziduí modelu 4.3.

spolehlivých klientů majících hodnoty logitů v intervalu $[-25; -15]$. Jedná se o pozorování, u kterých je $\hat{Y}$ velmi blízké nule.

4.5 Model s interakcemi

Nyní budeme uvažovat i tu možnost, že mezi jednotlivými vysvětlujícími proměnnými mohou nastat interakce (pouze do druhého řádu). Vzhledem k rozsáhlosti datového souboru a velkému počtu vysvětlujících proměnných není možné vyhodnotit model obsahující všechny interakce. K nalezení vhodného modelu nejprve vyšetříme několik modelů obsahujících různé interakce. V každém z nich provedeme sestupný výběr na přítomné proměnné, jejich druhé mocniny a interakce, abychom do závěrečného modelu zahrnuli ty proměnné a interakce, které byly obsaženy alespoň v jednom z doporučených modelů. Druhý, alternativní postup je popsán v podkapitole 4.6, kde použijeme zobecněný aditivní model.

V prvním kroku jsme tedy odhadli celkem 31 modelů tak, aby se každá z možných interakcí a druhých mocnin vyskytla právě v jednom z nich. Tyto modely navrhly použít některou z těchto proměnných: V_7, V_{22}, $V_1 * V_9$, $V_2 * V_{12}$, $V_8 * V_{10}$ a $V_{11} * V_{20}$, kde poslední zápis značí proměnnou V_{11}, V_{20} a jejich vzájemnou interakci (kterou též budeme značit $V_{11} : V_{20}$). Z důvodu vysokých hodnot VIF (10,09 až 36,50) jsme vypustili interakce V8:V10 a V11:V20. Po vypuštění šesti pozorování s největší hodnotou Cookovy vzdálenosti získáme

	$\hat{\boldsymbol{\beta}}$	Std.Er.	p		$\hat{\boldsymbol{\beta}}$	Std.Er.	p
Icpt	−1,799	0,399	0,000***	$V_{20}5$	−0,925	0,172	0,000***
V_1	−0,046	0,147	0,755	$V_{22}1$	−0,286	0,390	0,464
V_2	−0,236	0,120	0,048*	$V_{22}2$	0,399	0,302	0,186
V_7	−0,951	0,227	0,000***	$V_{22}3$	−0,784	0,549	0,154
V_8	−1,868	0,629	0,003**	$V_{22}4$	0,396	0,408	0,331
$V_9 1$	−0,375	0,172	0,029*	$V_{22}5$	−0,314	0,410	0,444
$V_{10}1$	−1,057	0,152	0,000***	$V_{22}6$	0,615	0,398	0,122
$V_{11}1$	−0,046	0,205	0,824	$V_{22}7$	−1,556	0,401	0,097
$V_{12}1$	0,151	0,187	0,419	$V_{22}8$	0,275	0,538	0,609
$V_{20}1$	−0,241	0,275	0,381	$V_{22}9$	−0,044	0,334	0,894
$V_{20}2$	−1,039	0,315	0,001***	$V_1 : V_9 1$	0,321	0,164	0,050*
$V_{20}3$	−0,407	0,284	0,152	$V_2 : V_{12}1$	−0,630	0,211	0,003**
$V_{20}4$	1,245	0,376	0,001***				

Tabulka 4.4: Odhady koeficientů, jejich standardní chyby a hladiny spolehlivosti proměnných modelu 4.3. Icpt značí absolutní člen.

model, ke kterému budeme odkazovat jako k modelu 4.3, jenž má devianci rovnou 1571,4 na 4104 stupních volnosti a Akaikeho kritérium 1621.

Odhady pro tento model jsou uvedeny v tabulce 4.4. Obrázek 4.3 ukazuje opět strukturu reziduí, která se však příliš neliší od struktury reziduí předchozího modelu 4.2.

4.6 Zobecněný aditivní model

Doposud jsme předpokládali pouze lineární vztahy mezi vysvětlujícími veličinami. V této podkapitole se za použití *zobecněného aditivního modelu* (generalized additive model – GAM) pokusíme zachytit i případné vztahy nelineární. Předpokladem našeho modelu GAM bude, že nezávisle proměnná ve střední hodnotě splňuje:

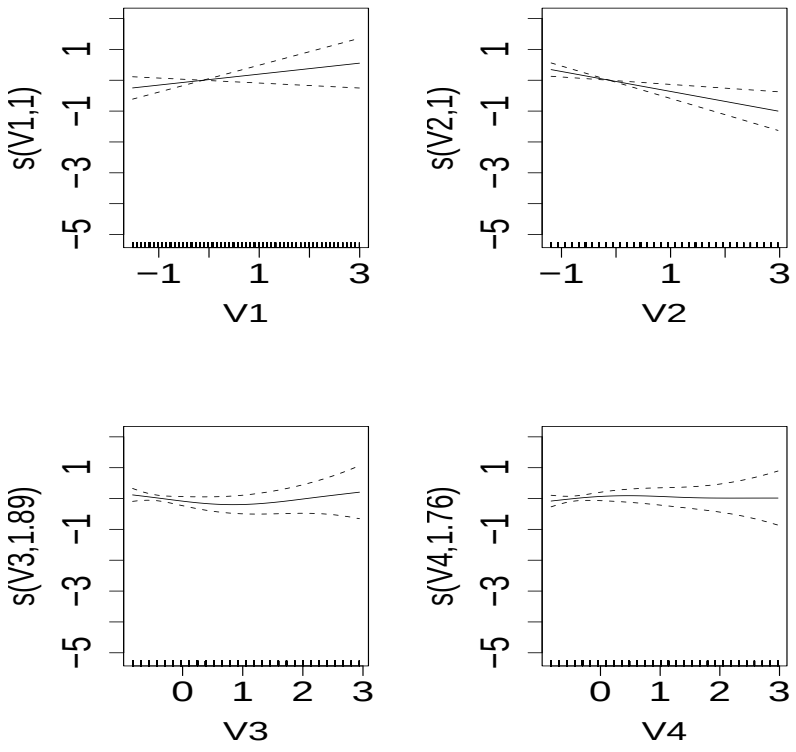
$$\mathrm{E}[Y|V_1,\ldots,V_4,V_6,V_7,V_8] = G\left(c + \sum_{i\in\{1,\ldots,8\}}^{i\neq 5} g_i(V_i)\right) ,$$

kde G je známá spojovací funkce ze vztahu 4.1, c je neznámá konstanta a g_i jsou neznámé neparametricky odhadované funkce.

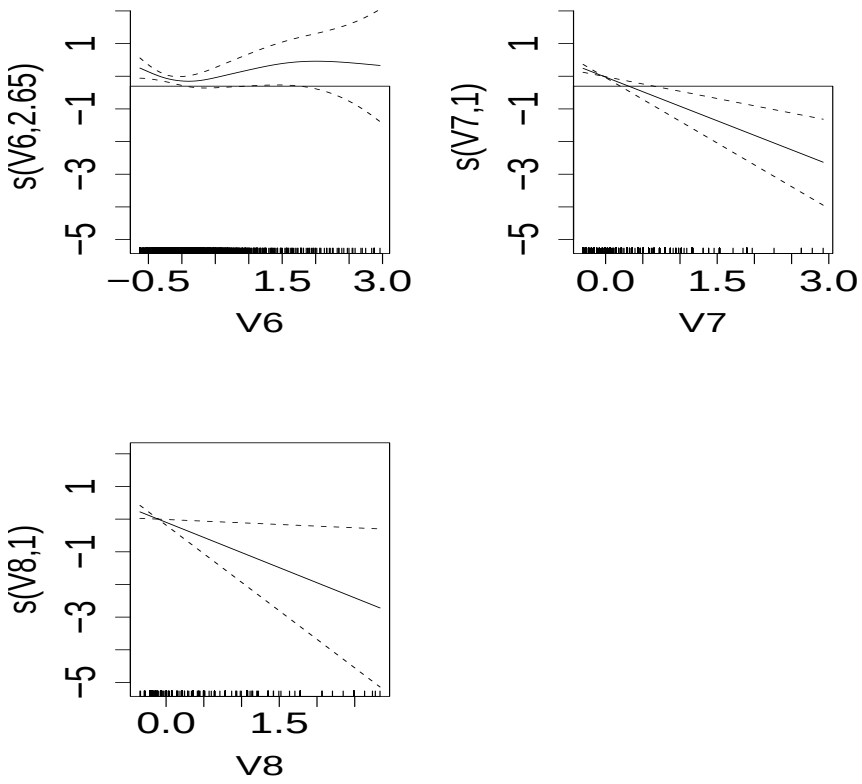
Manuál S-PLUS (2001) navrhuje pro odkrytí nelineární struktury spojitých veličin použít nejprve zobecněný aditivní model a následně jeho výsledky zpětně aplikovat do zobecněného lineárního modelu. Autoři doporučují modelovat (pokud je to možné) spíše lineárním nežli aditivním modelem, neboť druhý jmenovaný je komplikovanější a dochází u něj k větším ztrátám v přesnosti.

V použitém aditivním modelu jsme vyhlazovali měřitelné proměnné pomocí kubického splinu. Výsledky jsou na obrázcích 4.4 a 4.5. Jak můžeme vidět, kromě proměnných V_3 a V_6 není důvod předpokládat jiný průběh než lineární. Pro proměnnou V_3 by byla přiléhavější kvadratická křivka a pro V_6 taktéž, avšak jen do bodu 1,5, dále již může být modelována lineárně.

Nyní sestavíme zobecněný lineární model (jedná se spíše o model semiparametrický) obsahující všechny proměnné, kde V_3 a V_6 budou vyhlazeny



Obrázek 4.4: Neparametricky odhadnuté regresní křivky spojitých veličin ze zobecněného aditivního modelu.



Obrázek 4.5: Neparametricky odhadnuté regresní křivky spojitých veličin ze zobecněného aditivního modelu.

kubickým splinem. Navíc přidáme interakce použité v modelu 4.3 (tj. $V_1 : V_9$, $V_2 : V_{12}$, $V_8 : V_{10}$ a $V_{11} : V_{20}$). Tento model nejprve zjednodušíme obousměrnou krokovou regresí a posléze z důvodu vysokých hodnot VIF (11,06 – 35,29) vypustíme interakce $V_8 : V_{10}$ a $V_{11} : V_{20}$. Výsledný model obsahuje proměnné V_1, V_2, V_7, ..., V_{15}, V_{18}, ..., V_{22}, interakce $V_1 : V_9$ a $V_2 : V_{12}$ a spline proměnné V_3. Pro příliš velkou Cookovu vzdálenost ještě vypustíme šest pozorování a získáme tak model 4.4, který dosahuje reziduální deviance rovné 1511,1 se 4082 stupni volnosti a Akaikeho informační kritérium má rovné 1605. Odhady koeficientů tohoto modelu, jejich standardní chyby a hladiny spolehlivosti jsou uvedeny v tabulce 4.5.

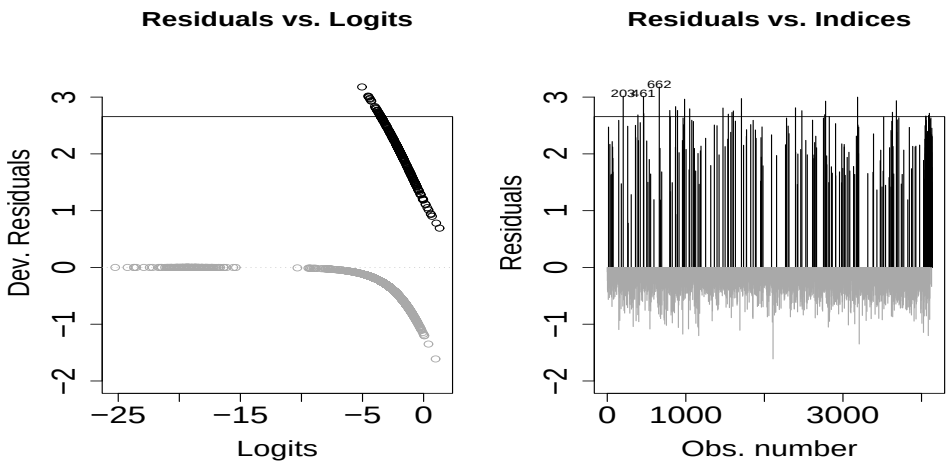
Obrázek 4.6 ukazuje strukturu reziduí, která se opět příliš neliší od reziduí předešlých dvou modelů.

	$\hat{\boldsymbol{\beta}}$	Std.Er.	p		$\hat{\boldsymbol{\beta}}$	Std.Er.	p
Icpt	−3,051	0,652	0,000***	$V_{19}5$	0,342	0,326	0,295
V_1	−0,027	0,163	0,867	$V_{20}1$	−0,331	0,282	0,239
V_2	−0,189	0,123	0,125	$V_{20}2$	−0,993	0,323	0,002**
V_3	−2,456	5,506	0,656	$V_{20}3$	−0,463	0,295	0,116
V_3^2	10,634	4,852	0,028*	$V_{20}4$	1,198	0,388	0,002**
V_7	−0,887	0,231	0,000***	$V_{20}5$	−1,022	0,176	0,000***
V_8	−1,799	0,630	0,004**	$V_{21}1$	−0,038	0,342	0,910
$V_9$1	−0,389	0,175	0,026*	$V_{21}2$	0,038	0,276	0,891
V_{10}1	−1,045	0,155	0,000***	$V_{21}3$	0,831	0,257	0,001**
V_{11}1	−0,028	0,210	0,894	$V_{21}4$	0,309	0,239	0,195
V_{12}1	0,095	0,185	0,608	$V_{21}5$	0,321	0,293	0,273
V_{13}1	−0,474	0,280	0,090.	$V_{21}6$	0,662	0,314	0,035*
V_{14}1	0,697	0,301	0,021*	$V_{22}1$	0,057	0,407	0,889
V_{15}1	0,253	0,351	0,471	$V_{22}2$	0,713	0,317	0,024*
V_{15}2	0,635	0,178	0,000***	$V_{22}3$	−0,349	0,570	0,541
V_{18}1	0,384	0,240	0,110	$V_{22}4$	0,610	0,417	0,144
V_{18}2	0,256	0,279	0,359	$V_{22}5$	0,193	0,430	0,654
V_{18}3	0,615	0,309	0,047*	$V_{22}6$	1,000	0,418	0,017*
V_{18}4	0,287	0,428	0,502	$V_{22}7$	−1,501	0,396	0,097
V_{18}5	0,639	0,245	0,009**	$V_{22}8$	0,474	0,560	0,397
V_{19}1	0,698	0,314	0,026*	$V_{22}9$	0,358	0,354	0,313
V_{19}2	0,247	0,309	0,425	$V_1 : V_9$	0,357	0,164	0,030*
V_{19}3	0,642	0,369	0,082.	$V_2 : V_{12}$	−0,527	0,202	0,009**
V_{19}4	−0,037	0,315	0,906				

Tabulka 4.5: Odhady koeficientů, jejich standardní chyby a hladiny spolehlivosti proměnných modelu 4.4. Icpt značí absolutní člen.

Model	Deviance	D.F.	AIC	tabulka	strana
4.1	1560,3	4070	1690	4.2	26
4.2	1522,0	4087	1606	4.3	28
4.3	1571,4	4104	1621	4.4	31
4.4	1511,1	4082	1605	4.5	33

Tabulka 4.6: Přehled modelů logistické regrese.



Obrázek 4.6: Struktura reziduí modelu 4.4.

4.7 Testovací vzorek

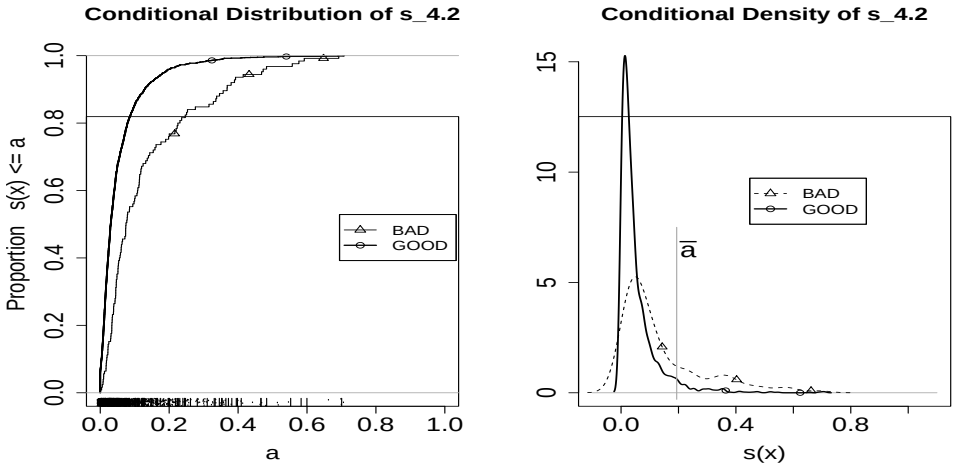
Nyní použijeme odhadnuté modely na testovací vzorek. Důležitým předpokladem (který je však díky způsobu vytvoření těchto vzorků splněn) je to, že data v obou vzorcích jsou nezávislá a podmíněně stejně rozdělená.

Tabulka 4.6 shrnuje doposud odhadnuté modely. Z hlediska hodnoty Akaikeho kritéria jsou nejvhodnějšími modely 4.4 a 4.2, pak následuje 4.3 a až daleko za ním zaostává model 4.1.

Obrázek 4.7 zobrazuje v levém panelu empirické distribuční funkce $\hat{G}_0$ a $\hat{G}_1$ modelu 4.2, pravý panel pak zobrazuje jádrové odhady hustot $\hat{g}_0, \hat{g}_1$ (normální jádro, šířka okna rovna směrodatné odchylce) téhož modelu počítané na testovacím vzorku. Pro jiné modely se tyto grafy příliš neliší, a proto je uvádíme až v dodatku D. Distribuční funkce mají očekávaný průběh, funkce $\hat{G}_0$ spolehlivých klientů leží nad funkcí $\hat{G}_1$, nicméně obě si jsou velmi podobné, tudíž plocha mezi těmito distribučními funkcemi není příliš velká a takovéto klasifikační pravidlo má do ideálního ještě daleko.

Z grafu hustot vidíme, že rozdělení $s(\mathbf{x})$ pro spolehlivé a nespolehlivé klienty se značně překrývají, tudíž jejich rozlišení na základě skóringové funkce nebude pořádně dost dobře možné.

Připomeňme, že $\bar{a}$ značí prahový bod takový, že je-li $s(\mathbf{x}_K) > \bar{a}$, označíme klienta K za nespolehlivého a v opačném případě za spolehlivého. Pro nalezení optimálního $\bar{a}^*$ se pokusíme použít důsledek 2.1. V podkapitole 3.5



Obrázek 4.7: Empirické distribuční funkce $\hat{G}_0, \hat{G}_1$ a odhady hustot $\hat{g}_0, \hat{g}_1$ pro model 4.2 ($\bar{a} = 0,194$).

jsme řekli, že podmínka 2.5 je splněna, zbývá tedy ověřit podmínku 2.6. Pravá nerovnost je splněna ($0,36/1,88 \doteq 0,19 < 1$), avšak problém nastává se splněním levé nerovnosti. Zřejmě totiž nenajdeme bod d takový, že $\hat{g}_0(d)/\hat{g}_1(d) < 0,19$, aby současně na celém intervalu $[c, d]$ platilo $\hat{g}_0' < \hat{g}_1' < 0$. V následujících kapitolách ještě uvidíme, že ani u jiných modelů nejsou předpoklady nutné pro použití našeho důsledku 2.1 splněny, a proto si naši situaci zjednodušíme a zvolíme prahový bod na základě následující úvahy: v případě standardizovaných nákladů ideální skóringová funkce s maximalizuje c-statistiku, a tudíž všem dobrým klientům přiřadí nižší skóre než klientům špatným. Takže v případě standardizovaných nákladů bude optimálním prahovým bodem 94% kvantil náhodné veličiny $s(\mathbf{x}_K)$, kde $K \in \mathcal{V}_1$.

Model	Suprem.	c-stat.	Gini	L
4.1	0,340	0,725	0,450	−0,566
4.2	0,355	0,731	0,461	−0,585
4.3	0,310	0,707	0,413	−0,557
4.4	0,361	0,716	0,432	−0,573

Tabulka 4.7: Srovnání modelů pomocí supremálního kritéria, c statistiky, Giniho koeficientu a očekávané ztráty.

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1823	**97**	5,05%	194
1	**97**	28	77,60%	(9,49%)

Tab. 4.8a: Model 4.1

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1834	**86**	4,48%	180
1	**94**	31	75,20%	(8,80%)

Tab. 4.8b: Model 4.2

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1823	**97**	5,05%	197
1	**100**	25	80,00%	(9,63%)

Tab. 4.8c: Model 4.3

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1830	**90**	4,69%	187
1	**97**	28	77,60%	(9,14%)

Tab. 4.8d: Model 4.4

Tabulka 4.8: Pozorované vs. předpovězené hodnoty s příslušnými mírami chybovosti.

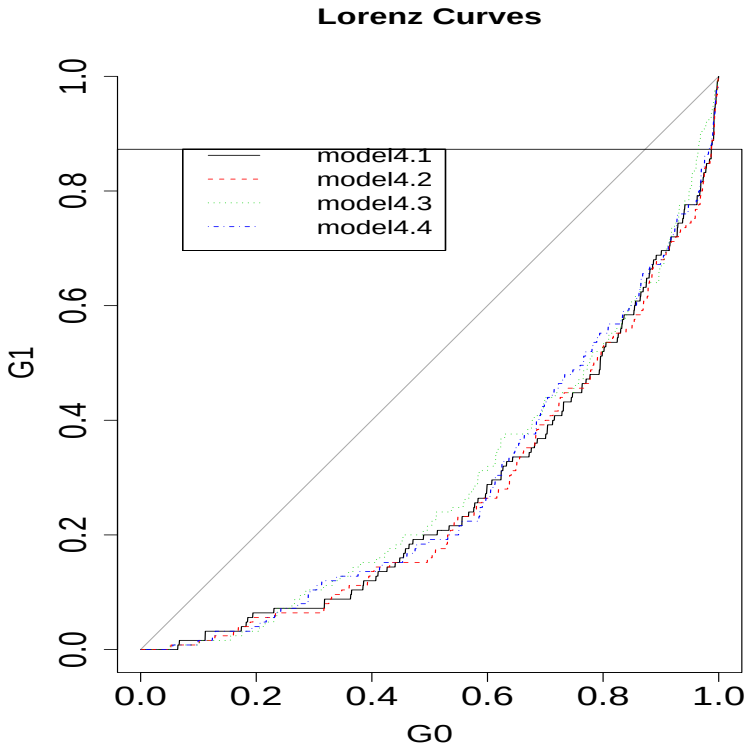
Charakteristiky kvality klasifikačních pravidel jsou uvedeny v tabulce 4.7. Podle Giniho koeficientu se na druhém místě překvapivě umístil model 4.1, který byl z hlediska Akaikeho kritéria nejhorší. Model 4.3 však potvrdil, že pro klasifikaci není příliš vhodný. To vidíme i z tabulky 4.8, ve které je zachycena celková míra chybovosti a její rozlišení na I. a II. druh pro jednotlivé modely. Model 4.3 chybně označuje 9,63% klientů, což je nejhorší výsledek mezi všemi modely. Podobně má i nejhorší míru chybovosti I. druhu: rovných 80% nespolehlivých klientů označuje jako spolehlivé. Na druhou stranu jednoznačně nejlepší z hlediska míry chybovosti je model 4.2 s 8,8% nesprávně zařazených klientů.

Z tabulek chybovostí můžeme pro jednotlivé modely spočítat celkovou očekávanou ztrátu podle vzorce

$$L = \sum_{i=0}^{1} \pi_i \sum_{j=0}^{1} c_{j|i} \mathrm{P}_i[\hat{Y} = j] \ .$$

Výsledky jsme rovněž shrnuli do tabulky 4.7. Nejnižší očekávané ztráty dosahuje opět model 4.2. Jelikož jeho ztráta činí $-0,585$, říkáme, že tento model dosahuje nejvyššího očekávaného zisku.

Všimněme si, že míra chybně zařazených nespolehlivých klientů v modelech 4.1 – 4.4 dosahuje 75,0 – 80,0%, což jsou velmi vysoké hodnoty, které by v praxi byly nepřijatelné. Tyto žalostné výsledky jsou nejspíše způsobeny malým zastoupením pozorování s hodnotou kategorie $Y = 1$. Naproti tomu



Obrázek 4.8: Lorenzova křivka pro každý z modelů 4.1 – 4.4.

chyby druhého druhu, tedy bezproblémoví klienti označení za nespolehlivé, dosahují velice příznivých hodnot 4,5 – 5,0%.

Obrázek 4.8 ukazuje na závěr Lorenzovy křivky modelů 4.1 – 4.4. Tyto křivky si jsou velice podobné a nemůžeme říci, že by jeden model byl výrazně lepší nežli modely jiné. Jako nejhorší se ale jeví model 4.3, neboť jemu odpovídající křivka se nejvíce blíží diagonále. To znamená, že empirické distribuční funkce $\hat{G}_0$ a $\hat{G}_1$ si jsou (z těchto čtyř modelů) nejvíce podobné.

Podíváme-li se na všechna srovnávací kritéria, vidíme, že jako jednoznačně nejhorší vychází model 4.3. Nejlepší výsledky vykazuje model 4.2. – pouze supremální kritérium by upřednostňovalo spíše model 4.4. Ten měl také nižší hodnoty deviance i Akaikeho kritéria.

Kapitola 5

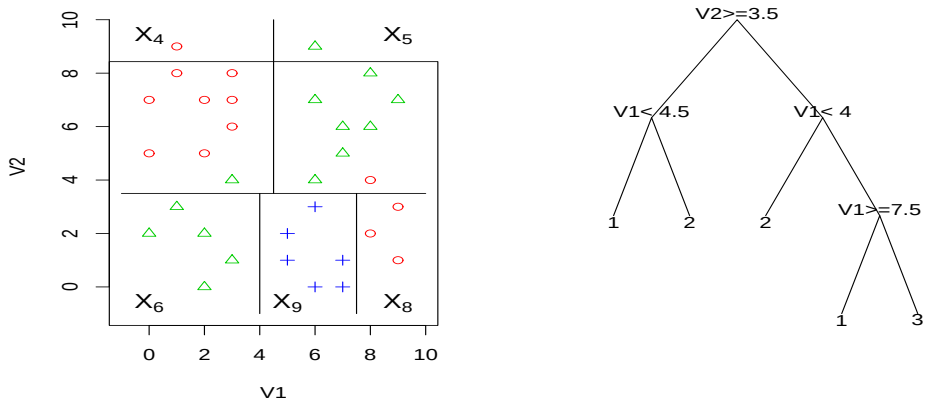
Klasifikační stromy

V této kapitole se budeme zabývat technikou binárních stromů používaných ke klasifikaci – neparametrickou metodou běžně známou pod zkratkou CART, která v sobě skrývá označení pro klasifikační a regresní stromy. Základním principem této metody je rekurzivní dělení prostoru vysvětlujících proměnných do stále menších množin. Technika klasifikačních stromů umožňuje zpracování dat vysoké dimenze, přirozeným způsobem zachází s měřitelnými i kategorickými proměnnými, je robustní vůči odlehlým pozorováním a nabízí sofistikovaný způsob řešení případů s chybějícími pozorováními.

Pro utvoření základní představy o fungování této metody věnujeme celou následující podkapitolu jejímu popisu. Nebudeme však zacházet příliš do detailů, celá metodika je podrobně popsána v monografii Breimann a kol. (1984), ze které ostatně vychází i náš souhrn. V dalších podkapitolách již použijeme klasifikační stromy k řešení naší úlohy.

5.1 Popis metody

Klasifikační stromy jsou konstruovány postupným rozkladem výběrového prostoru $\mathcal{X}$. Postup procedury si ilustrujeme na malém hypotetickém příkladě. Nechť je výběrový prostor $\mathcal{X} = \mathbb{R}^2$, vysvětlující proměnné označme V_1 a V_2 a nechť ve vývojovém vzorku je dáno 33 pozorování, která patří do jedné ze tří tříd. Tato pozorování vidíme v levé polovině obrázku 5.1. Tamtéž je již také znázorněn rozklad prostoru $\mathbb{R}^2$, který je výsledkem klasifikačního stromu z pravé poloviny obrázku.



Obrázek 5.1: Hypotetický příklad klasifikačního stromu.

Celý výběrový prostor $\mathcal{X}$, kterému pro konzistenci označení přiřadíme též symbol $\mathcal{X}_1$, rozdělíme na dvě disjunktní podmnožiny $\mathcal{X}_2$ a $\mathcal{X}_3$, kde $\mathcal{X}_2 = \mathbb{R} \times [3,5;\infty)$ a $\mathcal{X}_3$ je její doplněk v $\mathbb{R}^2$. V každém dalším kroku pak podobně rozdělíme některou ze známých množin $\mathcal{X}_i$ na dvě neprázdné disjunktní podmnožiny. Množinu $\mathcal{X}_2$ tak rozdělíme na podmnožiny $\mathcal{X}_4 = (-\infty;4,5) \times [3,5;\infty)$ a $\mathcal{X}_5 = [4,5;\infty) \times [3,5;\infty)$, které již dále nedělíme. Obdobným způsobem pokračujeme s množinou $\mathcal{X}_3$, kterou nejprve rozdělíme na množiny $\mathcal{X}_6$, $\mathcal{X}_7$, a druhou jmenovanou posléze ještě rozdělíme na množiny $\mathcal{X}_8$ a $\mathcal{X}_9$.

Klasifikační stromy jsou vlastně jen schématem znázorňujícím příslušný rozklad výběrového prostoru. Množina $\mathcal{X}_1$ představuje kořen stromu. Z kořene odchází všechna pozorování, pro něž je hodnota $V_2 \geq 3,5$, do levého dceřinného uzlu $\mathcal{X}_2$ a ostatní pozorování jdou do pravého dceřinného uzlu $\mathcal{X}_3$. Každý uzel tedy symbolizuje rozdělení množiny, kterou reprezentuje, do dvou disjunktních podmnožin podle hodnoty jedné nezávisle proměnné. Koncové uzly, které se již dále nedělí, nazveme listy, v grafu u nich udáváme příslušnost k jedné z daných tříd. Všechny listy jsou po dvou disjunktní a dohromady tvoří rozklad celého prostoru $\mathcal{X}$. Jedné třídě pak může příslušet i více listů. V našem případě je:

$$A_1 = \mathcal{X}_4 \cup \mathcal{X}_8, \qquad A_2 = \mathcal{X}_5 \cup \mathcal{X}_6, \qquad A_3 = \mathcal{X}_9.$$

Při klasifikování objektu neznámé třídy procházíme podle předepsaných dělení stromem postupně od kořene dolů, až skončíme v nějakém listu. Třídu,

která je přiřazena tomuto listu, pak přiřadíme i nově zkoumanému objektu.

Při právě popsané konstrukci konkrétního stromu jsme nezdůvodňovali jednotlivé kroky. Obecně je třeba ke konstrukci jakéhokoliv klasifikačního stromu znát odpovědi na následující otázky:

1. Jak bude vypadat množina všech možných dělení?
2. Jak rozdělit ten který uzel?
3. Které třídě přiřadit koncový list?
4. Kdy skončit dělení?

Než přistoupíme k jejich zodpovězení, zavedeme alespoň základní označení. Obecný binární strom budeme značit symbolem T, množinu všech jeho listů pojmenujeme $\tilde{T}$ a libovolný uzel označíme písmenem t. Strom T', který rozděluje výběrový prostor $\mathcal{X}$ stejně jako strom T a liší se od něho pouze tím, že končí proces dělení některých uzlů dříve, nazveme jeho podstromem, píšeme $T' < T$. Naopak uzel t společně s právě všemi svými následovníky patřící stromu T nazveme větví T_t. Pro každý uzel t označíme symbolem $N(t)$ počet pozorování $\mathbf{x}$ z vývojového vzorku $\mathcal{V}_1$ takových, že $\mathbf{x} \in t$, dále $N_j(t)$ bude značit počet objektů třídy j přítomných v uzlu t. Počet všech objektů třídy j ve vývojovém vzorku označíme N_j. Připomeňme, že π_j značí apriorní pravděpodobnost příslušnosti do j-té skupiny. Pravděpodobnost, že objekt třídy j bude přítomen v uzlu t, značenou $p(j,t)$, budeme odhadovat ze vztahu

$$p(j,t) = \pi_j \frac{N_j(t)}{N_j} \ .$$

Pravděpodobnost, že nějaké pozorování bude patřit do uzlu t, odhadneme jako

$$p(t) = \sum_{j=1}^{J} p(j,t) \ .$$

Podmíněná pravděpodobnost, že objekt z uzlu t je z j-té třídy, je pak dána vztahem

$$p(j|t) = \frac{p(j,t)}{p(t)} \ , \quad \text{tj.} \ \sum_{j=1}^{J} p(j|t) = 1 \ .$$

Jestliže pro všechny apriorní pravděpodobnosti platí $\pi_j = N_j/N$, pak podmíněné pravděpodobnosti $p(j|t)$ představují podíl objektů třídy j v uzlu t:

$$p(j|t) = \frac{N_j(t)}{N(t)} \ .$$

Nyní zavedeme základní pojem *nečistoty stromu*, ze které celá konstrukce klasifikačních stromů vychází.

DEFINICE 5.1

1. *Mírou*[1] *nečistoty budeme rozumět jakoukoliv nezápornou funci* ϕ *definovanou na množině všech uspořádaných* J*-tic* $(p_1, \ldots, p_J)$ *splňujících* $p_j \geq 0$ *a* $\sum_j p_j = 1$*, která vyhovuje následujícím předpokladům:*
   (a) ϕ *je symetrická funkce, tj.* $\phi(p_1, \ldots, p_J) = \phi(p_{i_1}, \ldots, p_{i_J})$*, kde* $(i_1, \ldots, i_J)$ *je libovolná permutace množiny* $\{1, \ldots, J\}$ *,*
   (b) ϕ *dosahuje svého maxima právě v bodě* $(\frac{1}{J}, \ldots, \frac{1}{J})$ *,*
   (c) $\phi(1, 0, \ldots, 0) = \phi(0, 1, 0, \ldots, 0) = \cdots = \phi(0, \ldots, 0, 1) = 0$ *a* $\phi > 0$ *jinde.*
2. *Nečistota uzlu* t *je funkce* $i(t) = \phi(p(1|t), \ldots, p(J|t))$ *pro nějakou danou míru nečistoty* ϕ*.*
3. *Nečistotu stromu* T *definujeme vztahem*

$$I(T) = \sum_{t \in \tilde{T}} i(t)p(t) = \sum_{t \in \tilde{T}} i(t) \sum_{j=1}^{J} p(j, t) \ .$$

Nečistota stromu je tedy střední hodnotou nečistoty koncových listů. Ačkoliv je funkce nečistoty základním stavebním kamenem pro konstrukci klasifikačních stromů, není její volba pro kvalitu výsledného klasifikačního pravidla klíčová (Breimann a kol. 1984). Často používanými funkcemi nečistoty je entropie, deviance či Giniho index různorodosti, který je definován vztahem:

$$i(t) = \sum_{j \neq i} p(j|t)p(i|t) = 1 - \sum_{j=1}^{J} p^2(j|t) \ . \tag{5.1}$$

Můžeme jej interpretovat následujícím způsobem: vybereme-li náhodně nějaký prvek z uzlu t, bude s pravděpodobností $p(i|t)$ náležet do třídy i. Pravděpodobnost, že se ale ve skutečnosti jedná o prvek třídy j, činí $p(j|t)$. Celkový odhad pravděpodobnosti špatného zatřídění pak bude právě Giniho index $\sum_{j \neq i} p(j|t)p(i|t)$.

[1] Slovo míra v tomto kontextu není exaktně definovaným matematickým pojmem.

5.1.1 Množina možných dělení

Množinu všech možných dělení označíme symbolem $\mathcal{D}$, její prvky budeme značit D. Nechť je dána nějaká podmnožina výběrového prostoru (tj. nějaký uzel) $\mathcal{X}_u$, kterou chceme rozdělit na dvě podmnožiny $\mathcal{X}_{u_l}$ a $\mathcal{X}_{u_r}$ podle nějaké náhodné veličiny V_i. Pokud bude tato veličina měřitelná, mající hodnoty v intervalu $[a_i, b_i)$, kde $-\infty \leq a_i < b_i \leq \infty$, pak bude množina $\mathcal{X}_{u_l}$ tvořena pozorováními, pro něž je $V_i \in [a_i, c_i)$, a množina $\mathcal{X}_{u_r}$ bude tvořena zase těmi pozorováními, kde $V_i \in [c_i, b_i]$, pro nějaké c_i. Bude-li veličina V_i kategorická, mající l kategorií, které si označíme pro jednoduchost symboly $1, 2, \ldots, l$, pak nějaká její neprázdná podmnožina a její doplněk budou tvořit množiny možných hodnot veličiny V_i pro pozorování z $\mathcal{X}_{u_l}$, resp. z $\mathcal{X}_{u_r}$. Množina všech možných dělení $\mathcal{D}$ je tvořena sjednocením všech možných dělení pro jednotlivé proměnné $V_i, i = 1, \ldots, r$.

Poznamenejme ještě, že možných dělení je jen konečně mnoho, neboť předpokládáme, že vývojový vzorek je konečná množina. Nabývá-li měřitelná veličina v našem vývojovém vzorku L různých hodnot, pak existuje $L - 1$ možných dělení podle této veličiny[2]. Má-li naopak nějaká kategorická veličina l různých hodnot, pak pro ni existuje celkem 2^{l-1} dělení.

5.1.2 Pravidlo dělení uzlu

Základní myšlenka dělení dat spočívá v tom, že dceřiné uzly jsou vždy čistší nežli uzel rodičovský. Jestliže se uzel t při nějakém dělení D rozdělí do dceřinných uzlů t_l, t_r a označíme-li symboly q_l, q_r podíly objektů z t jdoucí do uzlu t_l, resp. do t_r, pak definujeme kvalitu tohoto dělení $\varphi(D, t)$ jako pokles funkce nečistoty:

$$\varphi(D, t) = \Delta i(D, t) = i(t) - q_l i(t_l) - q_r i(t_r) \ .$$

Zcela přirozeně pak při rozhodování o nejvhodnějším dělení vybereme to dělení, které maximalizuje $\varphi(D, t)$.

Míru špatného zatřídění objektů z uzlu t odhadneme vztahem $\hat{r}(t) = 1 - \max_j p(j|t)$. Odhad celkové míry chybovosti stromu T určíme ze vztahu $\hat{R}(T) = \sum_{t \in \tilde{T}} \hat{r}(t) p(t)$.

[2]To platí za předpokladu, že uzel může být tvořen i jen jediným pozorováním. Většinou se stanovuje omezující podmínka na minimální možnou velikost uzlů.

Postup při dělení libovolného uzlu t je následující: procházíme všechny veličiny jednu po druhé, u každé určíme všechna možná dělení, ze kterých vybereme vždy to nejlepší pro danou veličinu, čímž získáme celkem r kandidátů. Z nich nakonec vybereme nejlepší dělení D^*, takže platí:

$$\Delta i(D^*, t) = \max_{D \in \mathcal{D}} \Delta i(D, t)$$

5.1.3 Pravidlo rozřazení do tříd

Naším cílem je strom s minimálním znečištěním, proto každý list přiřadíme té třídě, která minimalizuje danou funkci nečitoty tohoto uzlu. Definujeme-li např. funkci nečistoty uzlu vztahem $i(t) = 1 - \max_j p(j|t)$, získáme tak maximálně věrohodné pravidlo zařazování uzlů do tříd. Každému uzlu (a to obecně nejen koncovému) přiřadíme totiž tu třídu, jejíž hodnota je v daném uzlu nejpravděpodobnější. Takže třída j^* je přiřazena uzlu t, jestliže platí: $p(j^*|t) = \max_j p(j|t)$. Pro takovouto funkci nečistoty pak uzel obsahuje tím více špatně zařazených objektů, čím více je znečištěn.

5.1.4 Konec štěpení

Klasifikační strom můžeme, teoreticky vzato, větvit tak dlouho, dokud nějaké uzly obsahují více než jedno pozorování. Takový strom by zřejmě byl přeparametrizovaný, a proto je třeba najít nějaký vhodný, menší strom. Většinou se postupuje tak, že se nejprve zkonstruuje bohatě rozvětvený strom a v druhém kroku tento strom takzvaně prořežeme, abychom odstranili zbytečná dělení a získali tak strom optimální velikosti. Správně provedené prořezání stromu je klíčovým problémem pro výslednou kvalitu klasifikačního pravidla.

5.1.5 Prořezávání stromu

Čím více uzlů klasifikační strom T obsahuje, tím nižší hodnoty odhadu míry chybovosti $\hat{R}(T)$ dosahuje.

VĚTA 5.1 *Pro každý uzel t, který není listem, a jeho větev T_t platí:*

$$\hat{R}(t) > \hat{R}(T_t) \ .$$

DŮKAZ Viz věta 3.8 v Breimann a kol. (1984).

Při hledání stromu optimální velikosti nezáleží na tom, z jak velkého stromu vycházíme (pokud je dostatečně velký). Jestliže totiž proces prořezávání vycházející ze stromu T_{max} vybere jako optimální nějaký strom T_{opt}, který je podstromem stromu T'_{max}, pak prořezávání vycházející ze stromu T'_{max} vede k témuž optimálnímu stromu T_{opt}.

Nyní popíšeme způsob prořezávání minimalizující takzvanou celkovou míru komplexity stromu.

DEFINICE 5.2 *Komplexitou stromu T rozumíme počet jeho listů $|\tilde{T}|$. Míru komplexity stromu definujeme vztahem:*

$$\hat{R}_\alpha(T) = \hat{R}(T) + \alpha|\tilde{T}| \ , \tag{5.2}$$

kde parametr $\alpha \in \mathbb{R}_0^+$ nazýváme parametrem komplexity,

Touto definicí jsme zavedli jisté ad hoc kritérium, které nám může připomínat AIC či jiná penalizační kritéria. Nyní zavedeme stromy minimální ve smyslu právě uvedeného kritéria.

DEFINICE 5.3 *$T(\alpha)$, nejmenší minimalizující podstrom pro parametr komplexity α, splňuje:*
(i) $\hat{R}_\alpha(T(\alpha)) = \min_{T \leq T_{max}} \hat{R}_\alpha(T)$,
(ii) jestliže platí $\hat{R}_\alpha(T(\alpha)) = \hat{R}_\alpha(T)$ pro nějaký strom T, pak $T(\alpha) \leq T$.

VĚTA 5.2 *Pro každé $\alpha \in \mathbb{R}_0^+$ existuje nejmenší minimalizující podstrom popsaný v definici 5.3*

DŮKAZ Viz věta 3.7 v Breimann a kol. (1984).

Při prořezávání budeme vycházet z nějakého bohatě rozvětveného stromu T_{max}. Nejprve nalezneme nejmenší podstrom $T_1 < T_{max}$, se stejnou mírou chybovosti $\hat{R}(T_1) = \hat{R}(T_{max})$. Ten vznikne oříznutím všech dvojic listů t_l, t_r, pro jejichž otcovský list platí $\hat{R}(t) = \hat{R}(t_l) + \hat{R}(t_r)$.

VĚTA 5.3 *Pro každé dělení uzlu t do dceřinných uzlů t_l, t_r platí: $\hat{R}(t) \geq \hat{R}(t_l) + \hat{R}(t_r)$.*

DŮKAZ Viz věta 2.14 v Breimann a kol. (1984).

Při hledání $T_2 < T_1$ chceme nalézt nejslabší spojení ve stromu T_1, které pak zařízneme. Definujme pro každé $t \in T_1$ funkci $h_1(t)$ vztahem:

$$h_1(t) = \begin{cases} \frac{\hat{R}(t) - \hat{R}(T_t)}{|\tilde{T}_t| - 1} & t \notin \tilde{T}_1 \\ +\infty & t \in \tilde{T}_1 \ . \end{cases}$$

Za nejslabší článek stromu T_1 označíme uzel $\bar{t}_1 := \operatorname{argmin}_{t \in T_1} h_1(t)$ a položíme $\alpha_2 = h_1(\bar{t}_1)$. Budeme-li postupně zvyšovat parametr komplexity α, pak první uzel, pro který nastane rovnost

$$\hat{R}_\alpha(\{t\}) = \hat{R}_\alpha(T_t) \ ,$$

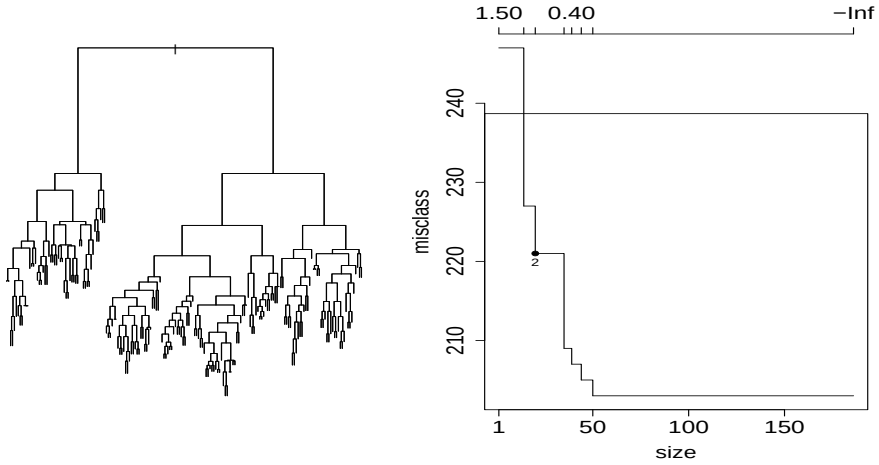
bude právě $\bar{t}_1$, a to pro hodnotu $\alpha = \alpha_2$. Je zřejmé, že samotný list $\{\bar{t}_1\}$ je vhodnější než celá větev $T_{\bar{t}_1}$. Uříznutím větve $T_{\bar{t}_1}$ získáme strom T_2. Obdobným způsobem sestrojíme ze stromu T_2 strom T_3 a pokračujeme dále, až dospějeme k samotnému kořeni $\{t_1\}$. Pokud kdykoliv nastane situace $h_i(\bar{t}_i) = h_i(\bar{t}'_i)$, definujeme $T_{i+1} = T_i - T_{\bar{t}_i} - T_{\bar{t}'_i}$. Právě popsanou konstrukcí získáme posloupnost do sebe vnořených stromů $T_1 > T_2 > \cdots > \{t_1\}$ a odpovídajících hodnot parametrů komplexity $0 =: \alpha_1 < \alpha_2 < \ldots$. Nezodpovězenou otázkou ale zůstává, které α_i (a jemu odpovídající strom T_i) vybrat. Jedno z možných řešení je využití křížového ověřování.

5.1.6 Doplňující poznámky

Při používání metody klasifikačních stromů může být problémem stabilita stromu. Dvě různá dělení některého uzlu mohou způsobit téměř stejný pokles nečistoty. Při malých změnách ve vstupních datech pak může být upřednostněno buď to či ono dělení.

Autoři sice uvádějí, že klasifikační stromy jsou poměrně robustní vůči odlehlým pozorováním, přesto byly dále vyvíjeny různé modifikace klasifikačních stromů, jejichž cílem bylo další zajištění robustnosti.

Dovolujeme si ještě upozornit, že v našem hypotetickém případě by rotace dat pravděpodobně vedla k lepším výsledkům. Princip konstrukce klasifikačních stromů umožňuje totiž dělení výběrového prostoru pouze ve směru rovnoběžném s některou z os. A to není v případě na obrázku 5.1 ten nejvhodnější způsob. Jako alternativu k rotaci nabízejí klasifikační (a regresní)



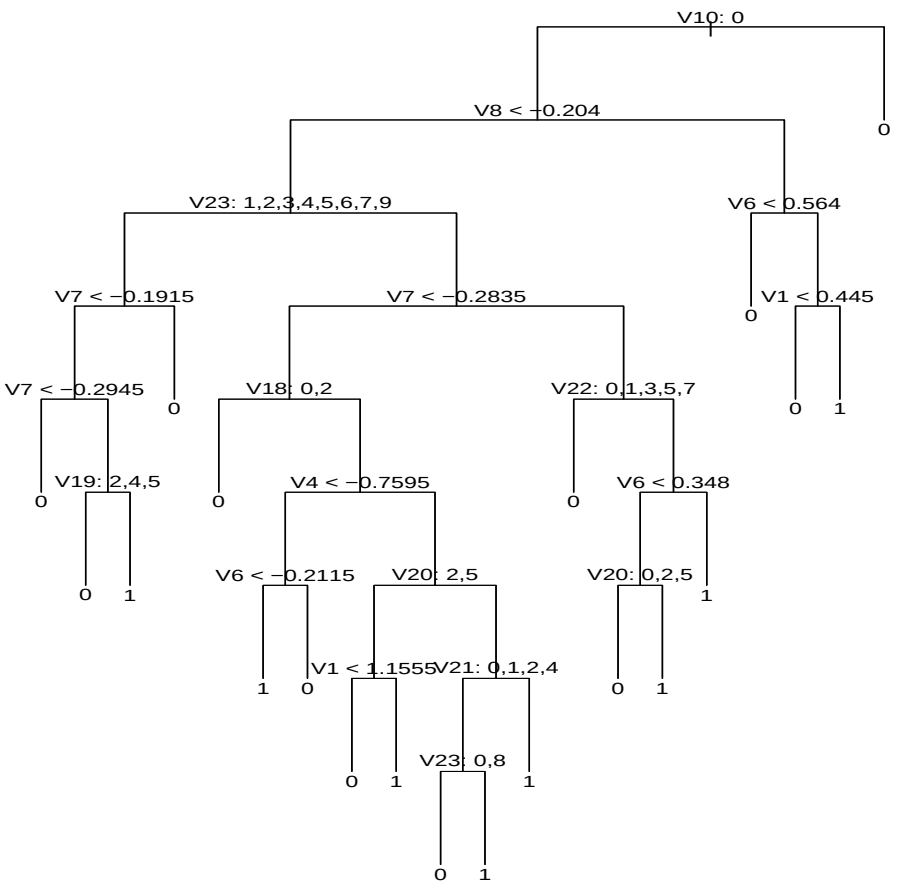
Obrázek 5.2: Příliš rozvětvený klasifikační strom (vlevo) a chybovost v závislosti na velikosti stromu, resp na parametru α (vpravo).

stromy rozšíření množiny možných dělení $\mathcal{D}$. Ta se pak neomezuje pouze na dělení podle jednotlivých proměnných, ale umožňuje i dělení podle lineární kombinace několika proměnných.

5.2 Sestavené modely

Podle výše popsaného postupu sestrojíme nejprve poněkud bohatší klasifikační strom. Ten můžeme vidět v levé polovině obrázku 5.2. Funkcí nečistoty byl zvolen Giniho index definovaný vztahem 5.1, nejmenší počet pozorování, která mohou tvořit uzel stromu, byl 3.

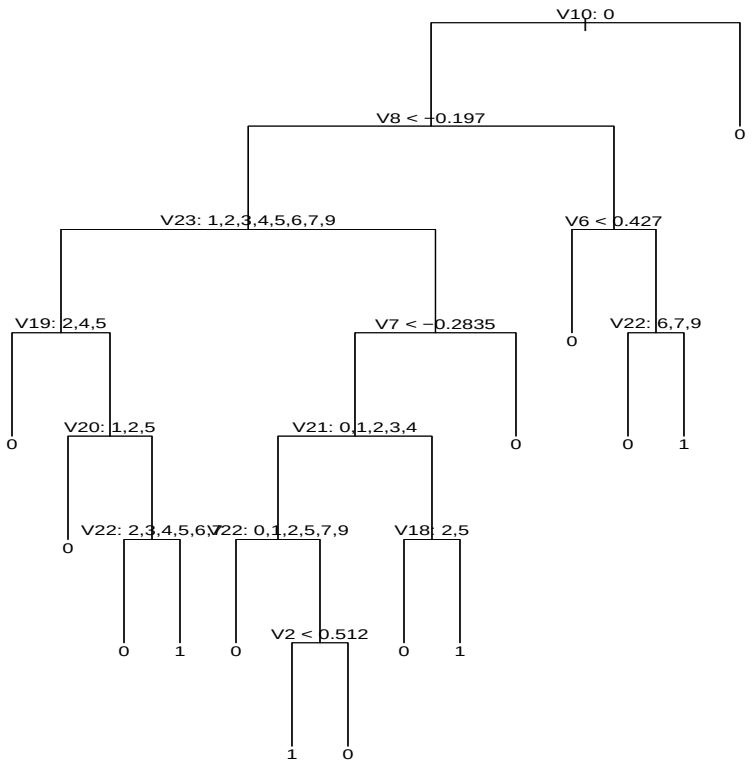
Tento rozsáhlý strom má 186 listů a kromě proměnných V_{13} a V_{16} využívá všech zbylých nezávislých veličin. Dá se očekávat, že tento strom bude přeparametrizovaný, přesto jej pro srovnání použijeme a označíme jako model 5.1. Jeho míra chybovosti je rovna 4,9% (31 špatných klientů a 172 dobrých klientů, kteří byli nesprávně zařazeni). Tento strom se pokusíme prořezat, abychom dostali klasifikační strom optimální velikosti. Na obrázku 5.2 vpravo vidíme počet chybně označených objektů v závislosti na počtu koncových listů podstromů, které vzniknou ořezáváním z našeho původního modelu 5.1. Horní horizontální osa nese hodnoty parametru komplexity α. Protože více než polovinu z celkového rozpětí chybně označených případů správně zařadí



Obrázek 5.3: Strom s dvaceti koncovými listy (model 5.2).

strom mající 20 koncových listů (bod č. 2 na obrázku), vybereme jej jako náš druhý model a označíme 5.2, jemu odpovídající parametr komplexity α je roven 0,80. Celý strom vidíme na obrázku 5.3. Jeho míra chybovosti je 5,3% (17 nespolehlivých a 204 spolehlivých klientů, kteří jsou špatně zatříděni).

Pro výběr optimálně velkého stromu zvolíme ještě jiný přístup. Rozdělíme vývojový vzorek na dvě poloviny, na první polovině budeme konstruovat stromy různých velikostí a na druhé budeme kontrolovat přiměřenost stromu dané velikosti. Dělení vývojového vzorku na poloviny jsme provedli tak, aby



Obrázek 5.4: Strom se čtrnácti koncovými listy (model 5.3).

v obou částech byl zachován poměr špatných a dobrých klientů. První polovina tak obsahuje 2053 klientů, z nichž je 124 nespolehlivých. Tímto postupem jsme jako optimální vybrali strom mající 14 listů, který označíme jako model 5.3 (obrázek 5.4). V první polovině vývojového vzorku dosahuje míry chybovosti 5,3% (100 špatných a 9 dobrých klientů, kteří jsou špatně označeni). Přesná struktura stromů modelu 5.2 a 5.3 je podrobně rozepsána v dodatku E.

5.3 Testovací vzorek

Metoda klasifikačních stromů neurčuje žádnou funkci, podle které by se rozhodovalo o příslušnosti objektů do jedné z možných tříd, ale rovnou zkoumané

Model	Suprem.	c-stat.	Gini	L
5.1	0,126	0,872	0,743	−0,542
5.2	0,141	0,872	0,743	−0,513
5.3	0,136	0,869	0,737	−0,528

Tabulka 5.1: Srovnání modelů pomocí supremálního kritéria, c statistiky, Giniho koeficientu a očekávané ztráty.

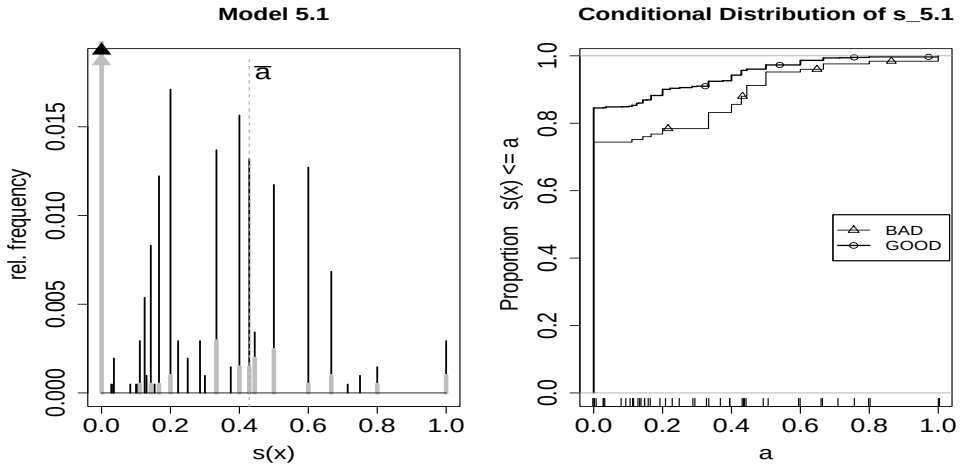
objekty do těchto tříd zařazuje. Abychom mohli použít všechny srovnávací metody popsané výše, definujeme skóringovou funkci příslušnou klasifikačnímu stromu jako pravděpodobnost, že daný objekt K náleží třídě 1:

$$s(\mathbf{x}_K) = \mathrm{P}[Y_K = 1|\mathbf{x}_K] = \mathrm{E}[Y_K|\mathbf{x}_K] \ .$$

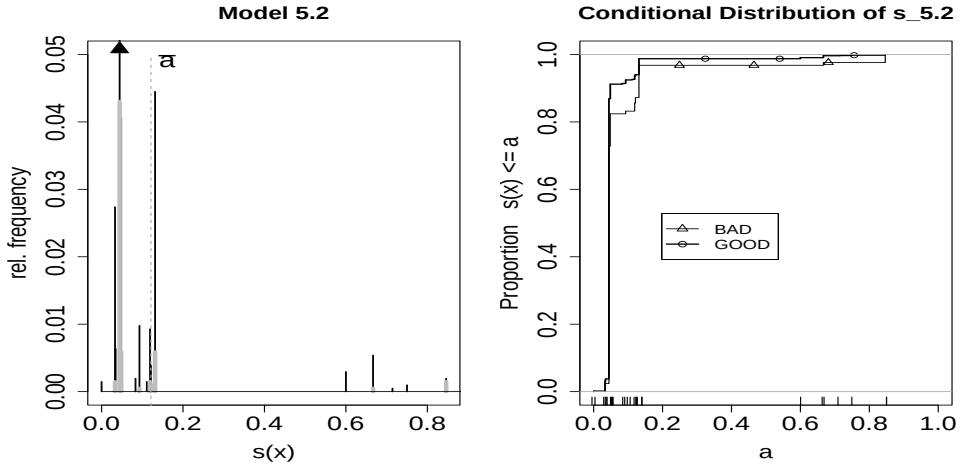
Tak získáme skóringovou funkci s požadovanou vlastností, že vyšší hodnoty skóre budou příslušet jedincům třídy 1. Opět použijeme prahovou hodnotu rovnou 94% kvantilu funkce s na vývojovém vzorku. Tím se ovšem odkláníme od vizualizace problému pomocí stromu, neboť pravidlo přiřazování uzlů některé ze tříd (popsáno v odstavci 5.1.3) pracuje s prahovou hodnotou 0,5. Jinými slovy, klasifikační strom nám posloužil jen jako pomocná metoda k určení skóringové funkce. Nevýhodou takto určené skóringové funkce je, že může nabývat pouze několika málo různých hodnot. V našem případě tomu tak skutečně je – pro model 5.1 má funkce s 30 různých hodnot, pro model 5.2 pouze 16 hodnot a pro model 5.3 již jen 14 různých hodnot.

Hodnoty supremálního kritéria, c-statistiky, Giniho koeficientu a střední ztráty jednotlivých modelů jsou uvedeny v tabulce 5.1. Z pohledu Giniho koeficientu jsou modely 5.1 a 5.2 stejně vhodné (a lepší než model 5.3). Supremální kritérium by z nich upřednostňovalo spíše model 5.2. Nejvyššího očekávaného zisku dosahuje překvapivě model 5.1.

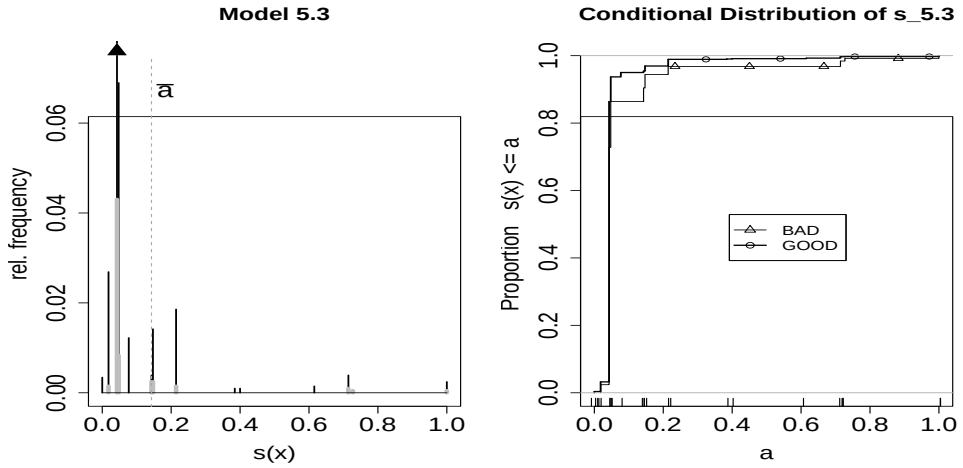
Na obrázku 5.5 vpravo vidíme empirické distribuční funkce modelu 5.1. Podmíněné hustoty neodhadujeme jádrovým odhadem právě kvůli malému počtu různých hodnot náhodné veličiny $s(X)$. Místo toho uvádíme na obrázku 5.5 vlevo sloupcový diagram. Černé sloupce představují spolehlivé a šedivé nespolehlivé klienty. Výška sloupců odpovídá relativnímu zastoupení příslušného skóre ve zkušebním vzorku. Jelikož $s(x) = 0$ pro 1623 spolehlivých (84,5%) a 93 nespolehlivých (74,4%) klientů, není tato kategorie zobrazena celá. Z této informace rovněž vyplývá, že žádnou volbou prahového



Obrázek 5.5: Sloupcové diagramy skóringové funkce pro dobré (černě) a špatné (šedivě) klienty modelu 5.1 ($\bar{a} = 0,429$). Vpravo empirické distribuční funkce $\hat{G}_0, \hat{G}_1$.



Obrázek 5.6: Sloupcové diagramy skóringové funkce pro dobré (černě) a špatné (šedivě) klienty modelu 5.2 ($\bar{a} = 0,121$). Vpravo empirické distribuční funkce $\hat{G}_0, \hat{G}_1$.



Obrázek 5.7: Sloupcové diagramy skóringové funkce pro dobré (černě) a špatné (šedivě) klienty modelu 5.3 ($\bar{a} = 0,143$). Vpravo empirické distribuční funkce $\hat{G}_0, \hat{G}_1$.

bodu nemůžeme snížit míru chybovosti I. druhu pod 74,4%. Empirické distribuční funkce a sloupcové diagramy veličin $s_{5.2}(\mathbf{X})$ a $s_{5.3}(\mathbf{X})$ jsou zachyceny na obrázcích 5.6 a 5.7. Tabulka 5.2 uvádí absolutní i relativní počty chybně zařazených klientů v jednotlivých modelech. Připomeňme, že model 5.1 odpovídá příliš rozvětvenému stromu o 186 listech, o kterém jsme očekávali, že bude přeparametrizovaný. Z hlediska celkové míry chybovosti, očekávané ztráty i tvaru distribučních funkcí G_0, G_1 dává tento strom překvapivě nejlepší výsledky.

5.4 Dodatečná analýza

Ve srovnání s modely předchozí kapitoly dosahují námi sestavené klasifikační stromy mnohem vyšších hodnot míry chybovosti I. druhu. Aby čtenář nenabyl dojmu, že metoda klasifikačních stromů je naprosto nevhodná, provedli jsme dodatečnou analýzu na jiném datovém souboru, který je podrobně popsán v knize Fahrmeir & Hamerle (1984). Jedná se taktéž o kreditní data, tentokrát z německé banky, závisle proměnná je rovněž dvouhodnotová, k dispozici jsou 3 spojité a 17 diskrétních nezávisle proměnných. Celý datový soubor obsahuje 1000 pozorování, z nichž 300 odpovídá špatným klientům. Datový soubor byl

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1837	**83**	4,32%	193
1	**110**	15	88,00%	(9,44%)

Tab. 5.2a: Model 5.1 (186 listů).

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1805	**115**	5,99%	224
1	**109**	16	87,20%	(10,95%)

Tab. 5.2b: Model 5.2 (20 listů).

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1832	**88**	4,58%	201
1	**113**	12	90,40%	(9,83%)

Tab. 5.2c: Model 5.3 (14 listů).

Tabulka 5.2: Pozorované vs. předpovězené hodnoty s příslušnými mírami chybovosti.

opětovně rozdělen na vývojový a testovací vzorek v poměru 2:1 a dále jsme postupovali obdobným způsobem jako v případě francouzských dat.

Fahrmeir & Hamerle (1984) použili nejprve lineární diskriminační analýzu, která chybně označila 28,9% pozorování (28,7%, resp. 29,1% chybných označení I., resp. II. druhu). Autoři rovněž použili kvadratickou diskriminační analýzu, která chybně označila 31,2% pozorování (28,3%, resp. 34,0% chybných označení I., resp. II. druhu). Námi vybraný klasifikační strom chybně označil 31,4% pozorování (28,7%, resp. 32,6% chybných označení I., resp. II. druhu). Z toho usuzujeme, že klasifikační stromy mohou konkurovat klasickým metodám, v případě značně nevyvážených datových souborů ale nemusí mít dostatečnou rozlišovací schopnost.

Breimann a kol. (1984) uvádějí celou řadu příkladů, kdy klasifikační stromy dávají lepší výsledky, než logistická regrese. Nespornou výhodou klasifikačních stromů je skutečnost, že implicitně řeší problém vzájemných interakcí mezi vysvětlujícími veličinami. V zobecněném lineárním modelu je třeba explicitně říci, které interakce budeme uvažovat. Běžně se pracuje s interakcemi druhého či maximálně třetího řádu. V předchozí kapitole jsme ale viděli, že použití všech interakcí druhého řádu nemusí výpočetní software zvládat ani v případě našich dat, která nejsou nijak extrémně velká.

Kapitola 6

Nejbližší sousedé

Metoda nejbližších sousedů je neparametrická technika standardně používaná pro odhadování hustot náhodných veličin, vyrovnávání regresních křivek a ke klasifikaci. Tato metoda nesestavuje explicitně žádný model, určuje ho však implicitně pomocí vývojového vzorku.

Abychom popsali algoritmus této metody, předpokládejme, že je dán vývojový vzorek $\mathcal{V}_1$ o M prvcích. Ke každému novému pozorování $\mathbf{x}_0$, které chceme zařadit do jedné z J tříd nejprve nalezneme jeho k nejbližších sousedů z $\mathcal{V}_1$. Vzdálenost zde měříme nějakou vhodnou metrikou d definovanou na $\mathcal{X}^2$. Prvek $\mathbf{x}_0$ pak přiřadíme té třídě, která je mezi jeho k nejbližšími sousedy nejpravděpodobnější.

Označíme-li prvky vývojového vzorku $\mathbf{x}_1, \ldots, \mathbf{x}_M$ tak, že $d(\mathbf{x}_0, \mathbf{x}_1) \leq d(\mathbf{x}_0, \mathbf{x}_2) \leq \cdots \leq d(\mathbf{x}_0, \mathbf{x}_M)$, pak metoda k-nejbližších sousedů přiřadí každému novému bodu $\mathbf{x}_0$ třídu $u(\mathbf{x}_0)$:

$$u(\mathbf{x}_0) = \text{argmax}_{j \in \mathcal{C}} \mathrm{P}(Y_0 = j | \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}). \tag{6.1}$$

Případně definujeme

$$u(\mathbf{x}_0) = \text{argmax}_{j \in \mathcal{C}} \mathrm{P}(Y_0 = j | \{\mathbf{x}_1, \ldots, \mathbf{x}_{k+z}\}), \tag{6.2}$$

existují-li ve vývojovém vzorku body $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{k+z}$ takové, že $d(\mathbf{x}_0, \mathbf{x}_k) = d(\mathbf{x}_0, \mathbf{x}_{k+1}) = \cdots = d(\mathbf{x}_0, \mathbf{x}_{k+z})$.

Argument maxima nemusí být určen jednoznačně, v takovém případě zvolíme náhodně jednu ze tříd maximalizujících pravděpodobnost v 6.1 (resp. v 6.2).

Číslo k zde funguje jako vyhlazovací parametr – nízká hodnota k bere v úvahu jen malé okolí $\mathbf{x}_0$, klasifikační pravidlo pak bude méně vychýlené, bude mít ale větší rozptyl. Velikost tohoto okolí ovšem závisí na „hustotě" vývojového vzorku v příslušné části výběrového prostoru. Naopak vyšší hodnota k přiřazuje nové objekty té třídě, která je nejpravděpodobnější v širším okolí bodu $\mathbf{x}_0$, snižuje se tedy rozptyl a zvyšuje vychýlení klasifikačního pravidla.

Největším problémem při používání metody nejbližších sousedů není však nalezení konkrétního parametru k, ale volba vhodné metriky. V případě spojitého náhodného vektoru $\mathbf{X}$ se často volí Euklidovská metrika, kterou pro dvě realizace $\mathbf{x}_a, \mathbf{x}_b$ můžeme zapsat:

$$d_2(\mathbf{x}_a, \mathbf{x}_b) = ((\mathbf{x}_a - \mathbf{x}_b)^\top (\mathbf{x}_a - \mathbf{x}_b))^{1/2} \ ,$$

kde jednotlivé veličiny jsou napřed standardizovány. Přes její jasnou interpretaci nemusí být nejvhodnější, nevýhodou může být např. to, že všechny použité proměnné jsou stejně důležité. Tedy i ty, které nemají na diskriminaci eventuelně žádný vliv. Hastie a kol. (2001), Hand & Henley (1996) a Hand (1997) uvádějí celou řadu modifikací metrik pro spojité proměnné.

V případě metriky pro nominální náhodnou veličinu se jedinou rozumnou metrikou (Hand 1997) zdá být:

$$d_0(x_a, x_b) = \begin{cases} 0 & \text{pokud } x_a = x_b \\ +\infty & \text{jinak} \ , \end{cases}$$

neboť měření vzdálenosti mezi nominálními kategoriemi zřejmě nedává smysl. V případě několika nominálních a několika spojitých veličin měříme vzdálenost jen mezi těmi body, jejichž nominální veličiny nabývají stejných hodnot (jinak je jejich vzdálenost rovna nekonečnu).

Ještě komplikovanější situace nastává u ordinálních veličin. Pro ordinální náhodnou veličinu X mající L kategorií můžeme konstruovat metriku následujícím způsobem (Härdle & Simar 2002): zavedeme indikátorové veličiny, tj. získáme celkem $L-1$ binárních proměnných, a náhodnou veličinu X můžeme reprezentovat vektorem $(X_1, \ldots, X_{L-1})$. Pro libovolné dva body

$\mathbf{x}_a = (x_{a1}, \ldots, x_{aL-1})$, $\mathbf{x}_b = (x_{a1}, \ldots, x_{aL-1})$ definujeme:

$$h_1 = \sum_{l=1}^{L-1} \mathcal{I}(x_{al} = x_{bl} = 1) \ , \qquad h_2 = \sum_{l=1}^{L-1} \mathcal{I}(x_{al} = 0, x_{bl} = 1) \ ,$$
$$h_3 = \sum_{l=1}^{L-1} \mathcal{I}(x_{al} = 1, x_{bl} = 0) \ , \quad h_4 = \sum_{l=1}^{L-1} \mathcal{I}(x_{al} = x_{bl} = 0) \ .$$

Metriku pro ordinální veličinu X pak zavedeme vztahem:

$$d(x_a, x_b) = \frac{h_1 + \delta h_4}{h_1 + \delta h_4 + \lambda(h_2 + h_3)} \ ,$$

kde δ, λ jsou váhy. Používané hodnoty jsou například $[0, 1]$, $[1, 1]$ či $[1, 2]$.

Nezodpovězenou otázkou ovšem zůstává, jak kombinovat různé veličiny, tj. řešení situace, kdy je pozorování charakterizováno směsí nominálních, ordinálních a měřitelných veličin.

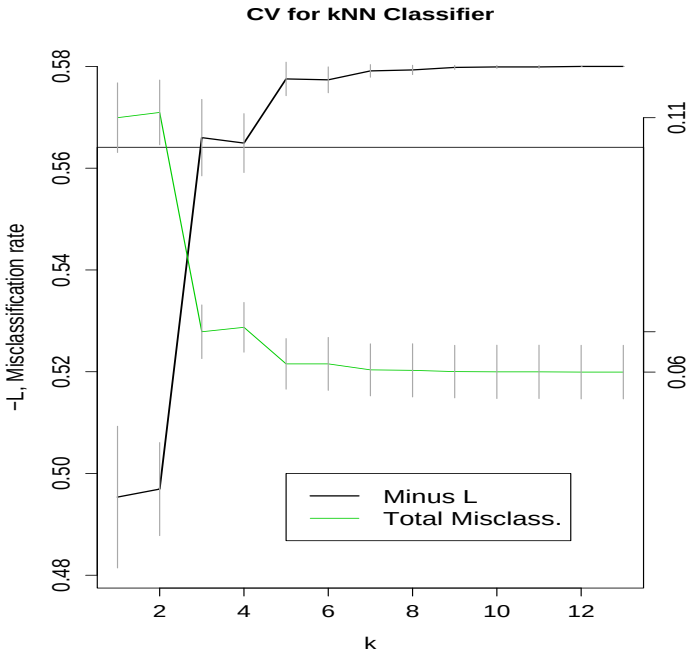
6.1 Sestavení modelu

V případě diskrétních veličin neumíme rozhodnout, který typ metriky použít, jelikož neznáme typ těchto veličin (nominální, ordinální či měřitelné). Z tohoto důvodu jsme se omezili pouze na spojité veličiny a zvolili pro ně Euklidovskou metriku, která je v programu R implementována.

Hodnotu parametru k pro výsledný model jsme určili na základě křížového ověřování. Jeho výsledky jsou shrnuty v tabulce 6.1 a ilustrovány na obrázku 6.1. Vidíme, že ztráta i celková chybovost jsou minimalizovány pro $k \geq 12$. V takovém případě jsou ovšem všichni špatní klienti označeni za dobré. Tato varianta se nezdá být zrovna nejvhodnější, nicméně to je v souladu s výsledky klasifikačních stromů, kde nejnižší míry chybovosti dosahoval strom mající pouze jediný uzel, tj. všem objektům přiřazoval třídu 0.

Takovéto klasifikační pravidlo, které akceptuje všechny klienty dává vlastně stejný výstup, jako expertní systém dané banky (který byl ovšem vytvořen na základě úplnějších dat). Otázkou tedy je, jak by naše klasifikační pravidlo fungovalo na klienty, kterým nebyl úvěr poskytnut, o nichž ale bohužel nemáme k dispozici žádné údaje.

Za model 6.1 zvolíme model určený spojitými proměnnými vývojového vzorku $\mathcal{V}_1$ a algoritmem dvanácti nejbližších sousedů s Euklidovskou metrikou. Abychom ponechali určitý prostor pro alespoň nějaké zařazení do třídy 1,



Obrázek 6.1: Aritmetické průměry očekávaného zisku (hodnoty na levé ose) a míry celkové chybovosti ± směrodatné odchylky (hodnoty na pravé ose) získané z křížového ověřování.

k	$\hat{R}_2$	$\sigma(\hat{R}_2)$	$\hat{L}$	$\sigma(\hat{L})$
1	0,110	0,0069	-0,495	0,0139
2	0,111	0,0064	-0,497	0,0092
3	0,068	0,0053	-0,566	0,0075
4	0,069	0,0049	-0,565	0,0058
5	0,062	0,0050	-0,578	0,0033
6	0,062	0,0052	-0,577	0,0026
7	0,060	0,0051	-0,579	0,0013
8	0,060	0,0052	-0,579	0,0009
9	0,060	0,0052	-0,580	0,0004
10	0,060	0,0053	-0,580	0,0003
11	0,060	0,0053	-0,580	0,0003
12	0,060	0,0053	-0,580	0,0000
13	0,060	0,0053	-0,580	0,0000

Tabulka 6.1: Odhady míry chybovosti a očekávané ztráty včetně směrodatných odchylek založené na křížovém ověřování.

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1920	**0**	0,00%	125
1	**125**	0	100,00%	(6,11%)

Tab. 6.2a: Model 6.1 ($k = 12$).

Y	$\hat{Y}$		špatně	celkem
	0	1		
0	1898	**22**	1,15%	143
1	**121**	4	96,80%	(6,99%)

Tab. 6.2b: Model 6.2 ($k = 3$).

Tabulka 6.2: Pozorované vs. předpovězené hodnoty s příslušnými mírami chybovosti.

použijeme ještě tentýž model využívající ovšem pouze 3 nejbližších sousedů, který budeme značit 6.2.

Pokusili jsme se rovněž využít binární veličiny V_{10}, kterou klasifikační stromy umístili do kořene, která v jednoduché regresi dosahovala nejnižší hodnoty deviance i Akaikeho kritéria a jejíž koeficient byl u všech modelů kapitoly 4 signifikantně různý od nuly na 0,1% hladině. Dosažené výsledky však byly téměř identické s výsledky bez využití informací o V_{10}.

6.2 Testovací vzorek

Jelikož pravděpodobnost náležení k nějaké třídě $\mathrm{P}(Y = j|o(\mathbf{x}))$ v okolí bodu $\mathbf{x}$ nabývá pouze osmi různých hodnot u modelu 6.1 (resp. čtyř hodnot u modelu 6.2), nebudeme v této kapitole konstruovat žádnou skóringovou funkci ani využívat žádné míry podobnosti. Výsledky obou modelů použitých na testovací vzorek jsou shrnuty v tabulce 6.2.

Jak se dalo čekat, model 6.1 označuje všechny klienty za dobré, a dosahuje tak celkové míry chybovosti 6,11% a očekávané ztráty $L = -0,580$. Oproti tomu model 6.2 špatně označuje 6,99% klientů, a dosahuje tak očekávané ztráty $L = -0,570$.

Z pohledu celkové míry chybovosti se jedná o dosud nejlepší výsledky, podle očekávané ztráty vycházel pouze model 4.2 lépe (model 4.4 je lepší nežli 6.2).

Proto jsme se podívali podrobněji na strukturu funkce L. V našem testovacím vzorku se z pohledu střední hodnoty ztráty vyplatí zvýšení chybovosti I. druhu o jednotku, pokud se současně sníží chybovosti II. druhu alespoň o tři. Nechť je dáno klasifikační pravidlo, které špatně označuje v našem testovacím vzorku m_1 špatných a m_2 dobrých klientů a uvažujme jiné klasifikační

pravidlo, které špatně zařadí $m_1 + 1$ špatných a $m_2 - z$ dobrých klientů. Pro ztrátové funkce těchto pravidel platí:

$$L_1 = -0,94\frac{1920 - m_2}{1920} + 0,94\frac{m_2}{1920} + 0,06 \cdot 6\frac{m_1}{125}$$
$$L_2 = -0,94\frac{1920 - m_2 + z}{1920} + 0,94\frac{m_2 - z}{1920} + 0,06 \cdot 6\frac{m_1 + 1}{125}$$

Ztráta druhého klasifikačního pravidla pak bude nižší než ztráta prvního, jestliže bude platit:

$$z \ > \frac{0,36}{125}\frac{1920}{1,88} \doteq 2,94$$

Příčina horších výsledků metody nejbližších sousedů může být i v tom, že na rozdíl od jiných metod jsme v této kapitole využívali pouze informace ze spojitých proměnných. Další prostor pro eventuelní zlepšení je též v použití jiné než Euklidovské metriky.

Kapitola 7

Srovnání s dřívějšími výsledky

Na závěr srovnáme nejlepší modely jednotlivých metod a porovnáme je s výsledky prací Härdle a kol. (2001) a Komorád (2002). První jmenovaná práce používá k sestavení modelů celý datový soubor o 6180 pozorováních (tj. $\mathcal{V}_1 \cup \mathcal{V}_2$) a k testování využívá jiný soubor o 2158 pozorováních. Další odlišností je setrvání autorů v názoru, že V_5 je spojitá veličina, a dále způsob, jakým obešli problém s určením prahového bodu $\bar{a}$ (chybná označení jsou počítána pro pět různých hodnot bodu $\bar{a}$). Zato druhá jmenovaná práce zpracovává naprosto stejný datový zdroj jako tato diplomová práce, díky čemuž jsou výsledky plně srovnatelné. V obou zmíněných pracích se však autoři zabývali z možných kritérií kvality modelů pouze mírou chybovosti. Hodnotu naší ztrátové funkce jsme pro jednotlivé modely dopočítali.

Härdle a kol. (2001) používá logistickou regresi se všemi vysvětlujícími proměnnými, lineární diskriminační analýzu a několik modelů semiparametrických. Z nich vybíráme pouze model, který neparametricky odhaduje proměnné V_1, V_3 a V_4, tj. předpokládá, že:

$$\mathrm{E}[Y|(V_1,\ldots,V_{23})] = G\left(g_1(V_1) + g_3(V_3) + g_4(V_4) + \sum_{i\in\{1,\ldots,23\}}^{i\neq 1,3,4} \beta_i V_i\right) ,$$

kde G je spojovací funkce ze vztahu 4.1. K modelu logistické regrese budeme odkazovat jako k modelu H.1, výsledky lineární diskriminační analýzy označíme H.2 a semiparametrický model budeme značit zkratkou H.3.

Pro základní představu o modelech použitých v druhé jmenované práci nyní stručně popíšeme architekturu vícevrstevného perceptronu (MLP) a neu-

ronových sítí založených na radiálních bázových funkcích (RBF). Jejich podrobnější popis i odkaz na další literaturu je možné nalézt ve zmíněné práci. Model vícevrstevného perceptronu obecně předpokládá, že pro závisle proměnnou Y platí:

$$\mathrm{E}[Y|\mathbf{x}] = F_2\left(w_0^{(2)} + \sum_{j=1}^{h} w_j^{(2)} F_1\left(w_{j0}^{(1)} + \sum_{i=1}^{r} w_{ji}^{(1)} x_i\right)\right) ,$$

kde $w_{ji}^{(1)}$ a $w_j^{(2)}$ jsou váhy (tzv. váhy skryté resp. výstupní vrstvy), r je dimenze vektoru $\mathbf{x}$ a neznámý parametr h značí počet jednotek skryté vrstvy. F_1 a F_2 jsou známé (tzv. transferové) funkce, za které se často volí logistická spojovací funkce. Parametr h se většinou určuje z křížového ověřování, neznámé parametry $w_{ji}^{(1)}$, $w_j^{(2)}$ se odhadují iterativní procedurou nazývanou zpětná propagace.

Architektura radiálních bázových funkcí předpokládá, že pro závisle proměnnou Y platí:

$$\mathrm{E}[Y|\mathbf{x}] = \psi\left(\sum_{j=1}^{h} [w_0 + w_j \phi_j(\mathbf{x})]\right) ,$$

kde h je počet shluků, $w_0, w_1, \ldots, w_h$ jsou neznámé parametry, ϕ_j jsou radiální bázové funkce[1] (každá má neznámé parametry $\mathbf{c}_j = (c_{1j}, \ldots, c_{rj}), \sigma_j$) a ψ je výstupní transferová funkce. Nejčastěji používanou radiální funkcí je r-rozměrná hustota normálního rozdělení (parametr $\mathbf{c}$ pak odpovídá vektoru středních hodnot). V první fázi učení RBF se body prostoru $\mathbb{R}^r$ sdruží do h shluků. Každý shluk je reprezentován jednou bázovou funkcí (tj. v prvním kroku odhadneme parametry $\mathbf{c}_j$ a σ_j). Počet shluků se opět určuje křížovým ověřováním. V druhé fázi se pak iterativně počítají váhy $w_0, w_1, \ldots, w_h$.

K dispozici máme čtyři modely neuronových sítí. Kromě různé architektury se liší rovněž počtem jednotek vstupní vrstvy a počtem jednotek (shluků) skryté vrstvy. Výstupní vrstva má jen jednu jednotku (neboť závisle proměnná veličina je pouze dvouhodnotová). Jednotky vstupní vrstvy odpovídají vysvětlujícím proměnným použitým v daném modelu. Spočteny byly následující modely: MLP o 23 vysvětlujících proměnných a 12 jednotkách ve skryté vrstvě (model K.1), MLP se 17 nezávisle proměnnými a 13

[1] Radiálně symetrická funkce splňuje: je-li $\|x_i\| = \|x_j\|$ pak $\phi(\|x_i\|) = \phi(\|x_j\|)$.

$\bar{a}$	I.typ	II. typ	Celk.	−L
0,05	27,4	34,3	33,9	0,197
0,10	53,1	13,3	15,3	0,499
0,25	83,2	2,5	6,7	0,593
0,50	99,1	0,2	5,3	0,579
0,75	100,0	0,0	5,2	0,580

Tab. 7.1a: Model H.1.

$\bar{a}$	I.typ	II. typ	Celk.	−L
0,05	28,3	32,6	32,3	0,225
0,10	54,0	13,1	15,2	0,499
0,25	85,0	2,0	6,4	0,596
0,50	99,1	0,2	5,4	0,579
0,75	100,0	0,0	5,2	0,580

Tab. 7.1b: Model H.3.

Mod.	I.typ	II. typ	Celk.	−L
H.2	46,0	24,5	25,7	0,314
K.1	88,0	5,5	10,5	0,520
K.2	88,0	5,6	10,6	0,518
K.3	87,2	5,5	10,5	0,523
K.4	84,8	5,4	10,2	0,533

Tab. 7.1c: Další převzaté modely.

Mod.	I.typ	II. typ	Celk.	−L
4.2	75,2	4,5	8,8	0,585
4.4	77,6	4,7	9,1	0,573
5.1	88,0	4,3	9,4	0,542
5.2	87,2	6,0	11,0	0,513
6.1	100,0	0,0	6,1	0,580

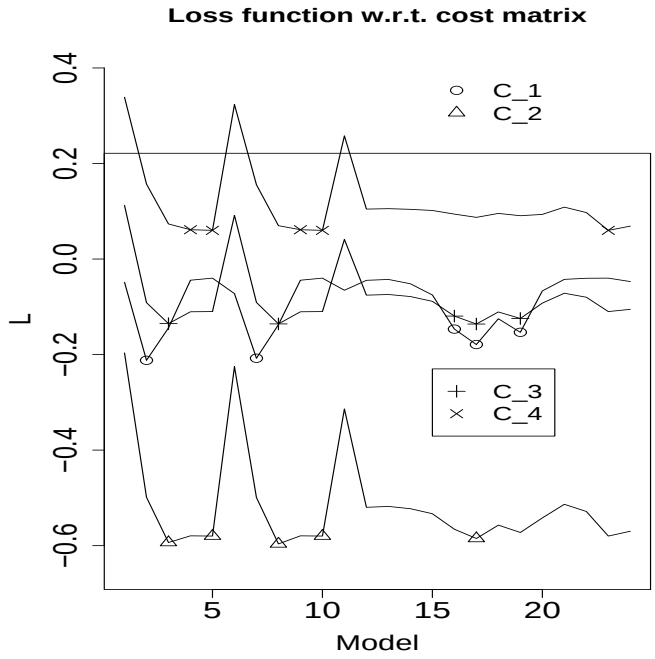
Tab. 7.1d: Výsledky našich modelů.

Tabulka 7.1: Závěrečné srovnání míry chybovosti a očekávané ztráty jednotlivých modelů. Chybovost je uvedena v procentech.

jednotkami skryté vrstvy (model K.2), RBF s 23 vysvětlujícími veličinami a 100 shluky ve skryté vrstvě (model K.3) a konečně RBF se 17 nezávisle proměnnými[2] a 80 shluky ve skryté vrstvě (model K.4).

Jelikož jedině modely 4. kapitoly dávaly dostatečně přesvědčivé hodnoty ukazatelů podobnosti, budeme v této kapitole používat pouze míru chybovosti a očekávanou ztrátu, které jsou uvedeny v tabulce 7.1. Výsledky modelu logistické regrese H.1 a semiparametrického modelu H.3. jsou uvedeny v tabulce 7.1a, resp. 7.1b. První sloupec udává hodnoty prahového bodu $\bar{a}$, který se pohybuje v rozmezí 0,05 – 0,75. Další sloupce uvádějí míru chybovosti I. a II. druhu, celkové míry chybovosti a opačnou hodnotu ztrátové funkce. V tabulce 7.1c jsou výsledky zbylých převzatých modelů a tabulka 7.1d shrnuje nejlepší výsledky našich modelů pro každou z metod. Nejvyššího zisku dosahuje semiparametrický model H.3 s prahovým bodem 0,25 a velice podobně dopadl i logistický model H.1 (s týmž prahovým bodem). Problém je ovšem v tom, že prahový bod je třeba volit dříve, než se model použije na testovací vzorek. Na třetím místě již figuruje náš logistický model 4.2, který má jednu

[2]Jedná se o tyto proměnné: V_1, V_2, V_5, V_7–V_{10}, V_{12}–V_{14}, V_{16} a V_{18}–V_{23}.



Obrázek 7.1: Očekávaná ztráta pro všechny modely při různých nákladových maticích.

z nejnižších hodnot míry chybovosti I. druhu. Těsně jej následuje model 4.4. Rovněž si všimneme, že metodou klasifikačních stromů a nejbližších sousedů dosáhneme jen průměrných, nikoli však nejhorších výsledků.

Poměřování modelů pomocí ztrátové funkce je velmi citlivé na volbu nákladové matice. Pro ilustraci uvažujme následující matice (kde matice C_2 bude námi doposud používaná matice definovaná v podkapitole 3.5):

$$C_1 = \begin{pmatrix} -1 & 15 \\ 1 & 0 \end{pmatrix}\ ,\ C_3 = \begin{pmatrix} -0,5 & 6 \\ 1 & 0 \end{pmatrix}\ ,\ C_4 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\ .$$

Výsledky jsou ilustrovány na obrázku 7.1. Na horizontální ose je postupně 5 modelů H.1, 5 modelů H.3, modely H.2, K.1–K.4, dále naše modely 4.1–4.4, 5.1–5.3, 6.1 a 6.2. Pro každou nákladovou matici je symboly označeno vždy pět nejlepších modelů, jejichž výsledky shrnuje tabulka 7.2. Vidíme, že díky velké apriorní pravděpodobnosti π_0 jsou hodnoty $c_{0|0}$ a $c_{1|0}$ mnohem důležitější než hodnota $c_{0|1}$. V případě standardizovaných nákladů (matice C_4), které bývají v odborné literatuře často používány, jsou jako nejlepší mo-

C_1		C_2		C_3		C_4	
Model	$-L$	Model	$-L$	Model	$-L$	Model	$-L$
H.1(0,1)	0,212	H.3(0,25)	0,596	4.2	0,136	H.1(0,75)	0,060
H.3(0,1)	0,208	H.1(0,25)	0,593	H.3(0,25)	0,136	H.3(0,75)	0,060
4.2	0,179	4.2	0,585	H.1(0,25)	0,135	6.1	0,060
4.4	0,153	H.1(0,75)	0,580	4.4	0,125	H.1(0,5)	0,061
4.1	0,147	H.3(0,75)	0,580	4.1	0,119	H.3(0,5)	0,061

Tabulka 7.2: Nejlepší modely pro jednotlivé nákladové matice s příslušnými hodnotami očekávaného zisku.

dely vybrány ty, které akceptují všechny klienty. V ostatních případech se do první trojice vždy dostal náš model 4.2., který se tedy jeví jako nejlepší z hlediska očekávané ztráty mezi všemi zbylými modely. Nedostatkem konkurujících modelů H.1 a H.3 je skutečnost, že z logiky věci je nesprávné volit prahovou hodnotu $\bar{a}$ až podle výsledků testovacího vzorku, aby takto získaná očekávaná ztráta byla minimální.

Kapitola 8

Závěr

V naší práci jsme provedli podrobnou empirickou studii využití klasifikačních metod při aplikaci na kreditní skórování. V porovnání s metodou nejbližších sousedů, klasifikačních stromů, vícevrstevného perceptronu a radiálních bázových funkcí si nejlépe vedla klasická logistická regrese. Nejenže v našem případě dává nejpřesnější klasifikační pravidlo, ale na rozdíl od ostatních jmenovaných metod umožňuje provádění celé řady statistických testů. Navíc jsme viděli, že metoda klasifikačních stromů může mít jisté problémy při zpracování značně nevyvážených datových souborů.

Všimli jsme si, že k používání měr podobnosti je třeba mít dostatečně spojitou skóringovou funkci. Tu v našem případě dává pouze logistická regrese, a proto jsme pro srovnání jednotlivých modelů využili míru chybovosti a očekávanou ztrátu.

Ukázali jsme, že používání standardizovaných nákladů, které je v odborné literatuře časté, je v případě kreditního skórování naprosto nevhodné a může vést až k akceptaci všech klientů. Úloha kreditního skórování vykazuje na rozdíl od obecného problému klasifikace jistá specifika, jedná se zejména o vyšší počet kategorických proměnných, přítomnost závisle proměnné o dvou třídách a hlavně skutečnost, že jedna z těchto tříd je tvořena pouze malým zlomkem celkového počtu pozorování.

Někteří autoři považují klasifikační úlohu dokonce za paradigmatický statistický problém (Hand 1997). Méně vzletnými slovy můžeme říci, že klasifikační techniky nacházejí širokého uplatnění v četných praktických aplikacích a bezesporu si zaslouží značnou pozornost, neboť nabízejí ještě velký prostor pro další studium a výzkum.

Dodatek A

Důkazy tvrzení

VĚTA A.1 *Nechť každý objekt charakterizovaný náhodným vektorem* $\mathbf{X}$ *patří do právě jedné ze dvou tříd. Nechť rozdělení vektorů* $\mathbf{X}$ *objektů třídy* 1 *má hustotu* f_1 *a rozdělení vektorů* $\mathbf{X}$ *objektů třídy* 2 *má hustotu* f_2 *vzhledem k nějaké* σ*-konečné míře* μ*. Nechť* π_1, π_2 *jsou nenulové apriorní pravděpodobnosti, že libovolný objekt patří do třídy* 1*, resp.* 2*. Předpokládejme, že pro náklady spojené s klasifikací platí:* $c_{1|1}, c_{2|2} \leq 0$ *a* $c_{1|2}, c_{2|1} > 0$*. Diskriminační pravidlo minimalizující ztrátovou funkci* L *je pak dáno rozkladem:*

$$A_1 = \left\{ \mathbf{x} \in \mathcal{X}; \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{c_{1|2} - c_{2|2}}{c_{2|1} - c_{1|1}} \frac{\pi_2}{\pi_1} \right\} , \tag{A.1}$$

$$A_2 = \left\{ \mathbf{x} \in \mathcal{X}; \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c_{1|2} - c_{2|2}}{c_{2|1} - c_{1|1}} \frac{\pi_2}{\pi_1} \right\} . \tag{A.2}$$

Body na hranici jsou náhodně rozřazeny do jedné z množin A_1, A_2*.*

DŮKAZ Celková očekávaná ztráta jako funkce diskriminačního pravi-

dla (tedy množin rozkladu) je:

$$\begin{aligned}
L(A_1) &= \pi_1 c_{1|1} \int_{A_1} f_1(\mathbf{x}) d\mu(\mathbf{x}) + \pi_1 c_{2|1} \int_{A_2} f_1(\mathbf{x}) d\mu(\mathbf{x}) \\
&\quad + \pi_2 c_{1|2} \int_{A_1} f_2(\mathbf{x}) d\mu(\mathbf{x}) + \pi_2 c_{2|2} \int_{A_2} f_2(\mathbf{x}) d\mu(\mathbf{x}) \\
&= \pi_1 c_{1|1} \int_{\mathcal{X}} \mathcal{I}(\mathbf{x} \in A_1) f_1(\mathbf{x}) d\mu(\mathbf{x}) \\
&\quad + \pi_1 c_{2|1} \int_{\mathcal{X}} (1 - \mathcal{I}(\mathbf{x} \in A_1)) f_1(\mathbf{x}) d\mu(\mathbf{x}) \\
&\quad + \pi_2 c_{1|2} \int_{\mathcal{X}} \mathcal{I}(\mathbf{x} \in A_1) f_2(\mathbf{x}) d\mu(\mathbf{x}) \\
&\quad + \pi_2 c_{2|2} \int_{\mathcal{X}} (1 - \mathcal{I}(\mathbf{x} \in A_1)) f_2(\mathbf{x}) d\mu(\mathbf{x}) \\
&= \pi_1 c_{2|1} + \pi_2 c_{2|2} \\
&\quad + \int_{\mathcal{X}} \mathcal{I}(\mathbf{x} \in A_1) \left[\pi_1 f_1(\mathbf{x})(c_{1|1} - c_{2|1}) + \pi_2 f_2(\mathbf{x})(c_{1|2} - c_{2|2}) \right] d\mu(\mathbf{x})
\end{aligned}$$

První dva sčítance posledního výrazu jsou konstantní, a tak vidíme, že celý výraz bude minimalizován na množině:

$$A_1 = \{\mathbf{x} \in \mathcal{X}; \pi_1 f_1(\mathbf{x})(c_{1|1} - c_{2|1}) + \pi_2 f_2(\mathbf{x})(c_{1|2} - c_{2|2}) < 0\} \ ,$$

což je ekvivalentní vztahu A.1, neboť $c_{1|1} - c_{2|1} < 0$, $\pi_1 > 0$ a $f_2(\mathbf{x}) > 0$ skoro všude na $\mathcal{X}$.

■

VĚTA A.2 *Nechť jsou splněny předpoklady věty A.1 a nechť dále $s(\mathbf{x})$ je nějaká skóringová funkce definovaná na $\mathcal{X}$, G_1, G_2 jsou distribuční funkce veličiny $s(\mathbf{X})$ podmíněné jevem $[Y = 1]$ resp. $[Y = 2]$ a g_1, g_2 jsou odpovídající podmíněné hustoty, o kterých předpokládáme, že mají v nějakém intervalu $[c, d]$ první derivaci (v krajních bodech uvažujeme jednostranné derivace). Nechť ve vnitřních bodech tohoto intervalu dále platí*

$$\pi_2(c_{1|2} - c_{2|2}) g_2'(x) > \pi_1(c_{2|1} - c_{1|1}) g_1'(x) \ . \tag{A.3}$$

Optimální prahový bod $\bar{a}^*$ minimalizující ztrátovou funkci L je pak implicitně určen vztahem:

$$\frac{g_1(\bar{a}^*)}{g_2(\bar{a}^*)} = \frac{c_{1|2} - c_{2|2}}{c_{2|1} - c_{1|1}} \frac{\pi_2}{\pi_1} \ . \tag{A.4}$$

DŮKAZ Celkovou očekávanou ztrátu přepíšeme:

$$\begin{aligned}
L &= \pi_1 \left(c_{1|1} \int_{A_1} f_1(\mathbf{x}) d\mu(\mathbf{x}) + c_{2|1} \int_{A_2} f_1(\mathbf{x}) d\mu(\mathbf{x}) \right) \\
&\quad + \pi_2 \left(c_{1|2} \int_{A_1} f_2(\mathbf{x}) d\mu(\mathbf{x}) + c_{2|2} \int_{A_2} f_2(\mathbf{x}) d\mu(\mathbf{x}) \right) \\
&= \pi_1 c_{1|1} \mathrm{P}[s(\mathbf{x}) \leq \bar{a} | Y = 1] + \pi_1 c_{2|1} \mathrm{P}[s(\mathbf{x}) > \bar{a} | Y = 1] \\
&\quad + \pi_2 c_{1|2} \mathrm{P}[s(\mathbf{x}) \leq \bar{a} | Y = 2] + \pi_2 c_{2|2} \mathrm{P}[s(\mathbf{x}) > \bar{a} | Y = 2] \\
&= \pi_1 c_{1|1} G_1(\bar{a}) + \pi_1 c_{2|1} (1 - G_1(\bar{a})) \\
&\quad + \pi_2 c_{1|2} G_2(\bar{a}) + \pi_2 c_{2|2} (1 - G_2(\bar{a})) \\
&= \pi_1 c_{2|1} + \pi_1 (c_{1|1} - c_{2|1}) G_1(\bar{a}) + \pi_2 c_{2|2} + \pi_2 (c_{1|2} - c_{2|2}) G_2(\bar{a})
\end{aligned}$$

Pro nalezení lokálního extrému funkci zderivujeme:

$$\frac{\partial L}{\partial \bar{a}} = \pi_1 (c_{1|1} - c_{2|1}) g_1(\bar{a}^*) + \pi_2 (c_{1|2} - c_{2|2}) g_2(\bar{a}^*) \stackrel{!}{=} 0$$

a dostáváme tak vztah A.4. Aby se jednalo o lokální minimum, musí být L konvexní, tj. druhá derivace ztrátové funkce kladná, to ale zajišťuje podmínka A.3:

$$L'' = \pi_1 (c_{1|1} - c_{2|1}) g_1'(\bar{a}^*) + \pi_2 (c_{1|2} - c_{2|2}) g_2'(\bar{a}^*) > 0$$

■

Dodatek B

Ke spojitým veličinám

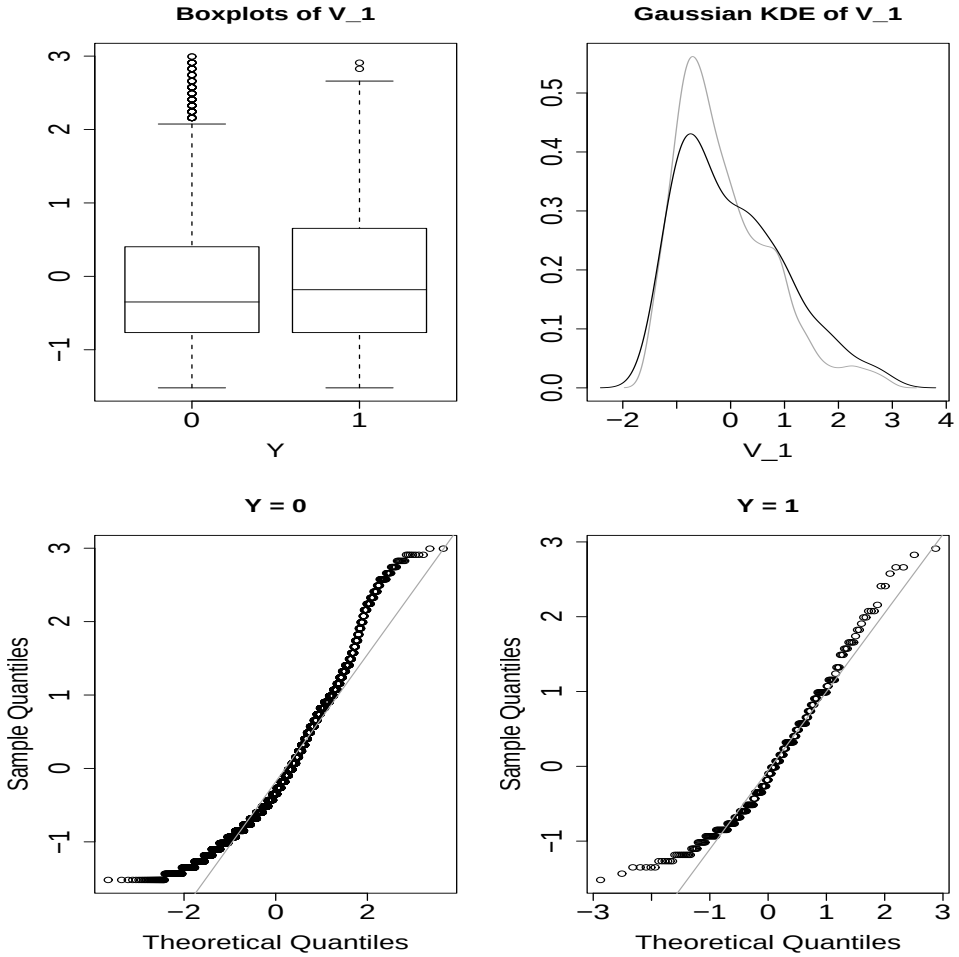
V tomto a následujícím dodatku přinášíme výsledky analýzy vstupních dat vývojového vzorku, které byly zmíněny v kapitole 3. Nejprve uvádíme hodnoty korelačních koeficientů (tabulka B.1), na dalších stranách je uvedeno několik grafů umožňujících explorativní analýzu jednotlivých spojitých veličin v závislosti na hodnotě Y. Krabicové diagramy (vlevo nahoře) srovnávají rozdělení spolehlivých klientů ($Y = 0$, diagram vlevo) a klientů nespolehlivých ($Y = 1$, diagram vpravo). Obrázek vpravo nahoře srovnává jádrové odhady hustot (normální jádro, šířka okna rovna směrodatné odchylce) dvou populací klientů. Hustota spolehlivých klientů je značena šedě. Dolní řada grafů ukazuje normální diagramy pro spolehlivé (vlevo) a nespolehlivé (vpravo) klienty. Ke každé veličině V_i rovněž uvádíme tabulku s příslušnými hodnotami extrémů, kvartilů, středními hodnotami $\bar{V}_i$ a směrodatnými odchylkami σ_i pro populaci dobrých a špatných klientů.

	V_1	V_2	V_3	V_4	V_6	V_7	V_8
V_1	1,00	0,43	0,36	0,27	0,33	0,09	0,02
V_2	0,43	1,00	0,35	0,24	0,19	0,05	0,05
V_3	0,36	0,35	1,00	0,17	0,27	0,04	−0,01
V_4	0,27	0,24	0,17	1,00	0,01	−0,02	0,03
V_6	0,33	0,19	0,27	0,01	1,00	0,21	−0,02
V_7	0,09	0,05	0,04	−0,02	0,21	1,00	0,07
V_8	0,02	0,05	−0,01	0,03	−0,02	0,07	1,00

Tabulka B.1: Korelační matice spojitých veličin.

Y	Min.	1. Kvart.	Medián	$\bar{V_1}$	3.Kvart.	Max.	σ_1
0	−1,520	−0,766	−0,349	−0,129	0,403	2,990	0,884
1	−1,520	−0,766	−0,182	0,034	0,654	2,910	0,996

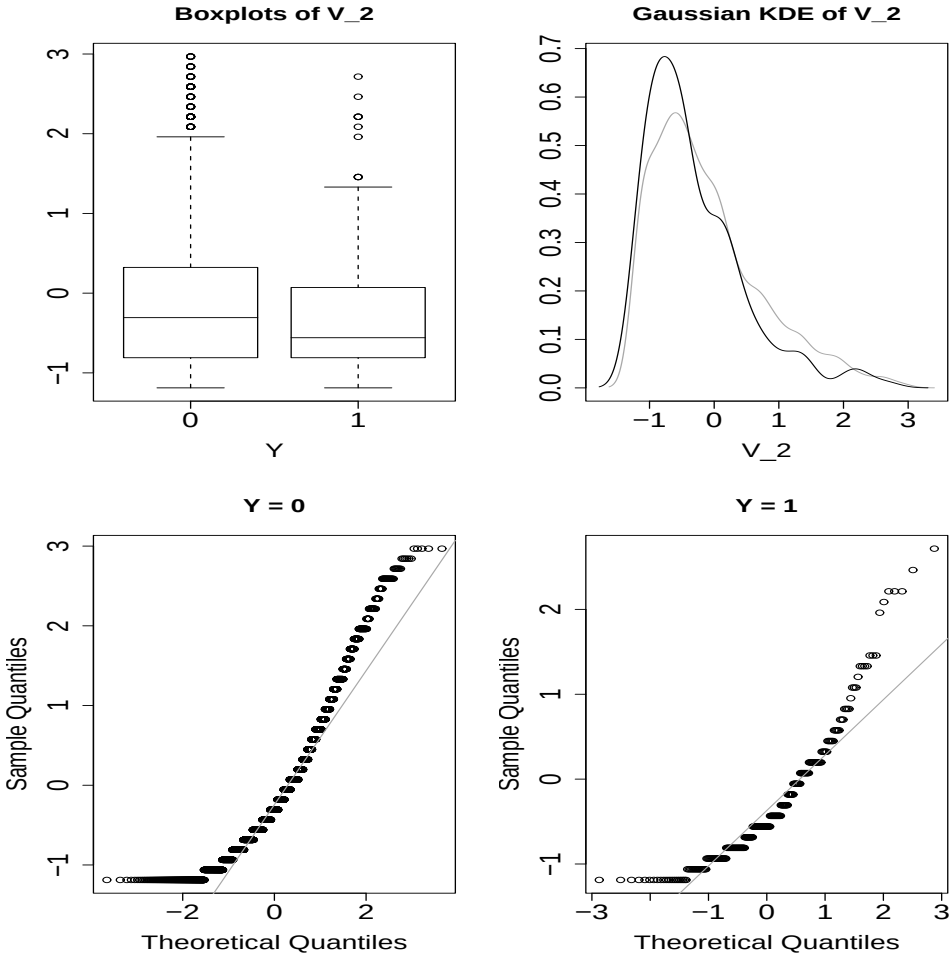
Tabulka B.2: Souhrn charakteristik proměnné V_1.



Obrázek B.1: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_1.

Y	Min.	1. Kvart.	Medián	$\bar{V}_2$	3.Kvart.	Max.	σ_2
0	−1,190	−0,810	−0,307	−0,111	0,323	2,970	0,850
1	−1,190	−0,810	−0,559	−0,308	0,071	2,720	0,770

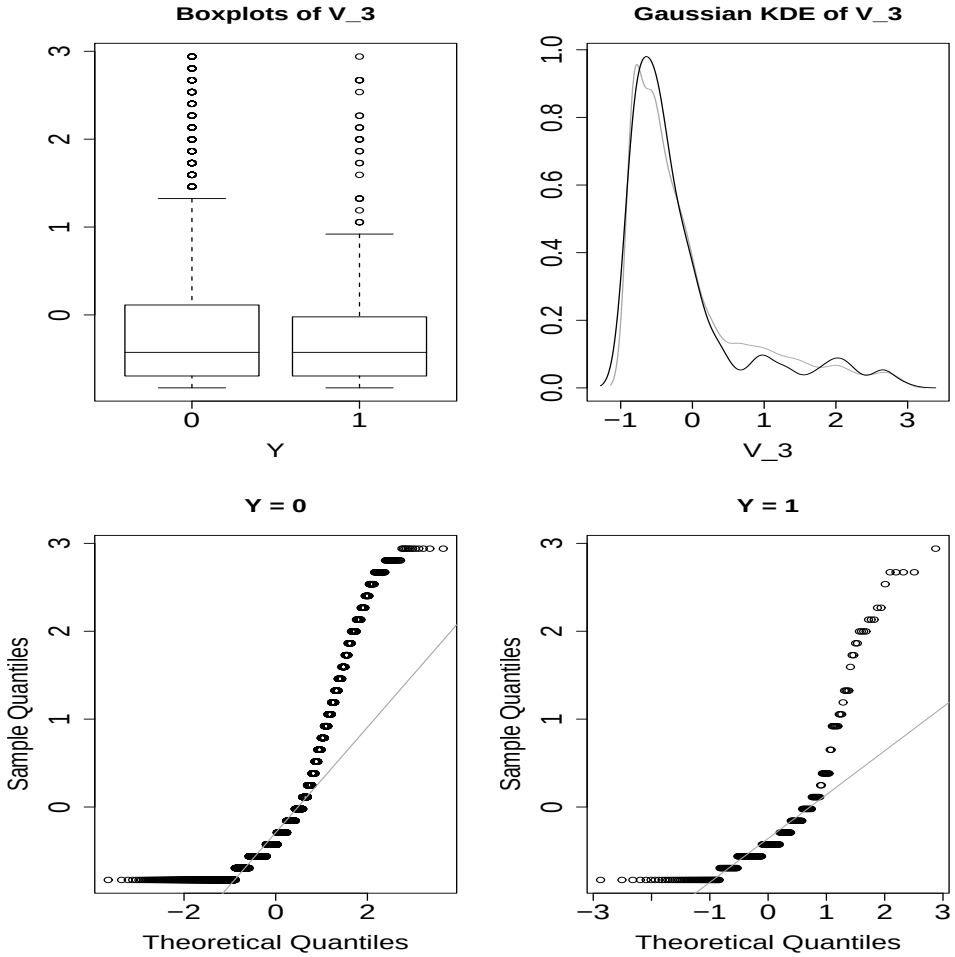
Tabulka B.3: Souhrn charakteristik proměnné V_2.



Obrázek B.2: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_2.

Y	Min.	1. Kvart.	Medián	$\bar{V}_3$	3.Kvart.	Max.	σ_3
0	−0,830	−0,695	−0,426	−0,079	0,113	2,940	0,851
1	−0,830	−0,695	−0,426	−0,146	−0,022	2,940	0,852

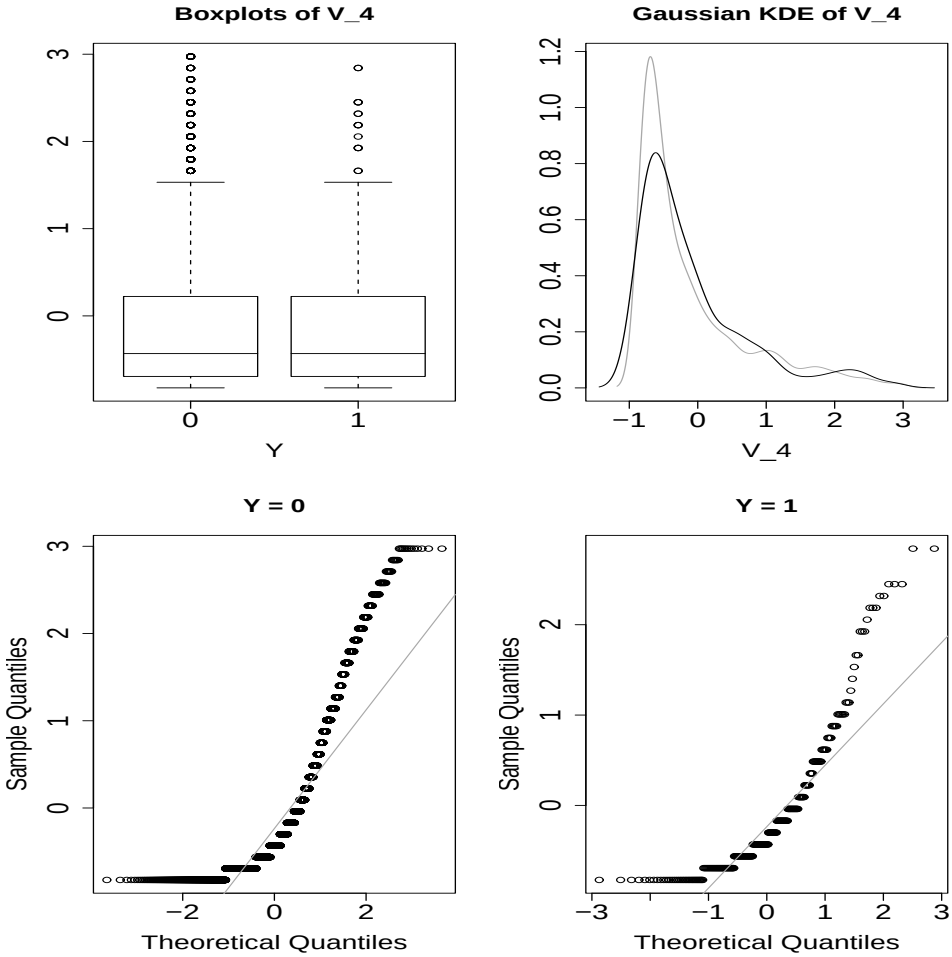
Tabulka B.4: Souhrn charakteristik proměnné V_3.



Obrázek B.3: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_3.

Y	Min.	1. Kvart.	Medián	$\bar{V}_4$	3.Kvart.	Max.	σ_4
0	−0,825	−0,694	−0,432	−0,115	0,223	2,970	0,817
1	−0,825	−0,694	−0,432	−0,083	0,223	2,840	0,814

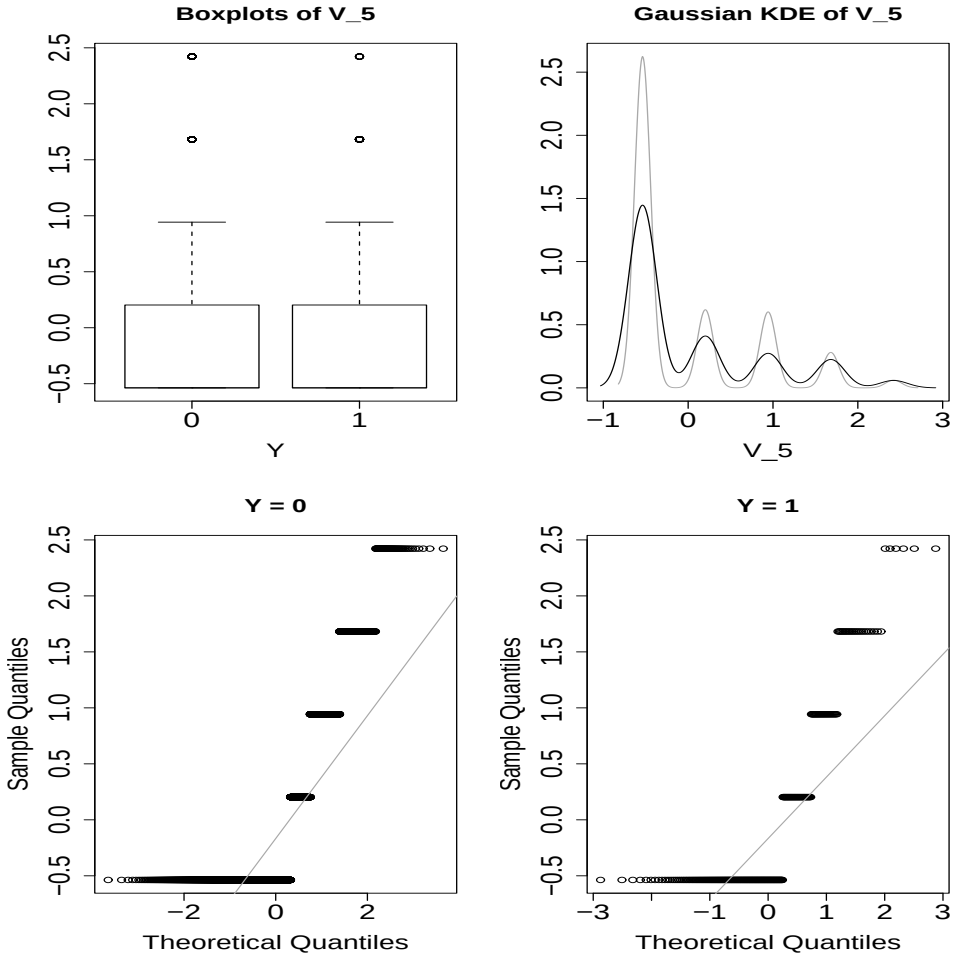
Tabulka B.5: Souhrn charakteristik proměnné V_4.



Obrázek B.4: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_4.

Y	Min.	1. Kvart.	Medián	$\bar{V_5}$	3.Kvart.	Max.	σ_5
0	−0,537	−0,537	−0,537	−0,022	0,203	2,420	0,768
1	−0,537	−0,537	−0,537	0,035	0,203	2,420	0,830

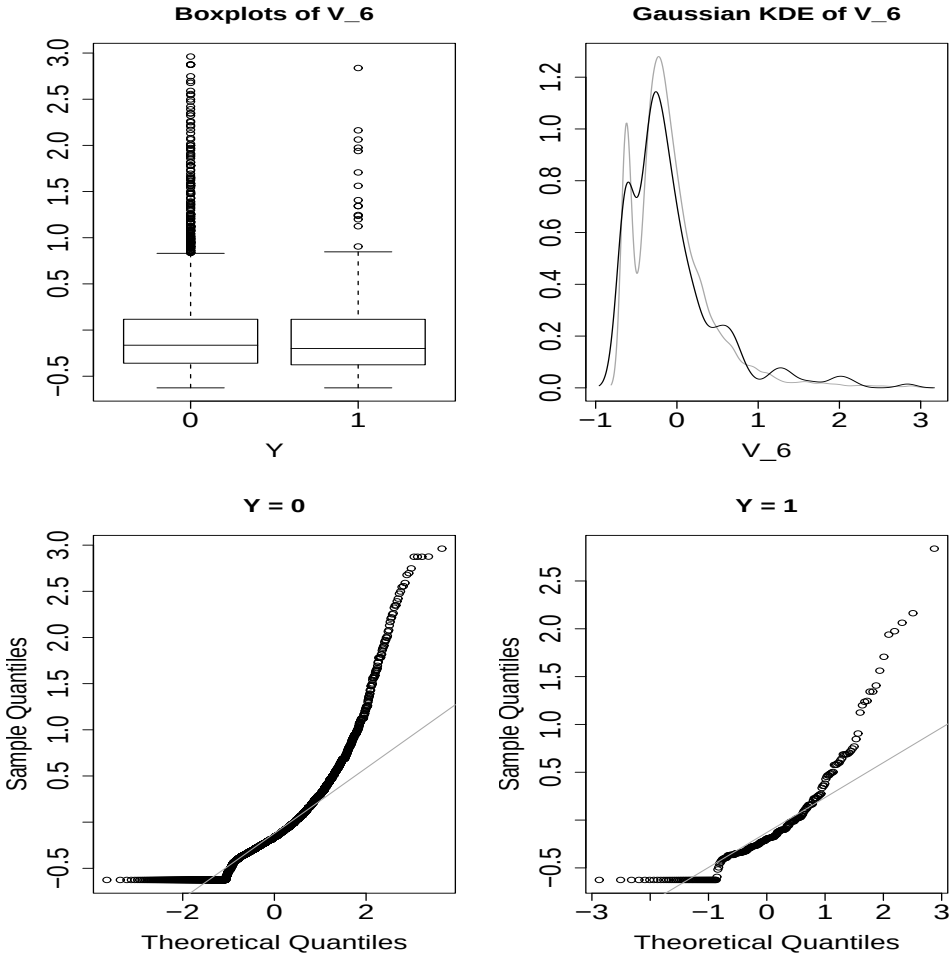
Tabulka B.6: Souhrn charakteristik proměnné V_5.



Obrázek B.5: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_5.

Y	Min.	1. Kvart.	Medián	$\bar{V}_6$	3.Kvart.	Max.	σ_6
0	−0,626	−0,359	−0,164	−0,070	0,117	2,960	0,487
1	−0,626	−0,377	−0,200	−0,061	0,117	2,840	0,569

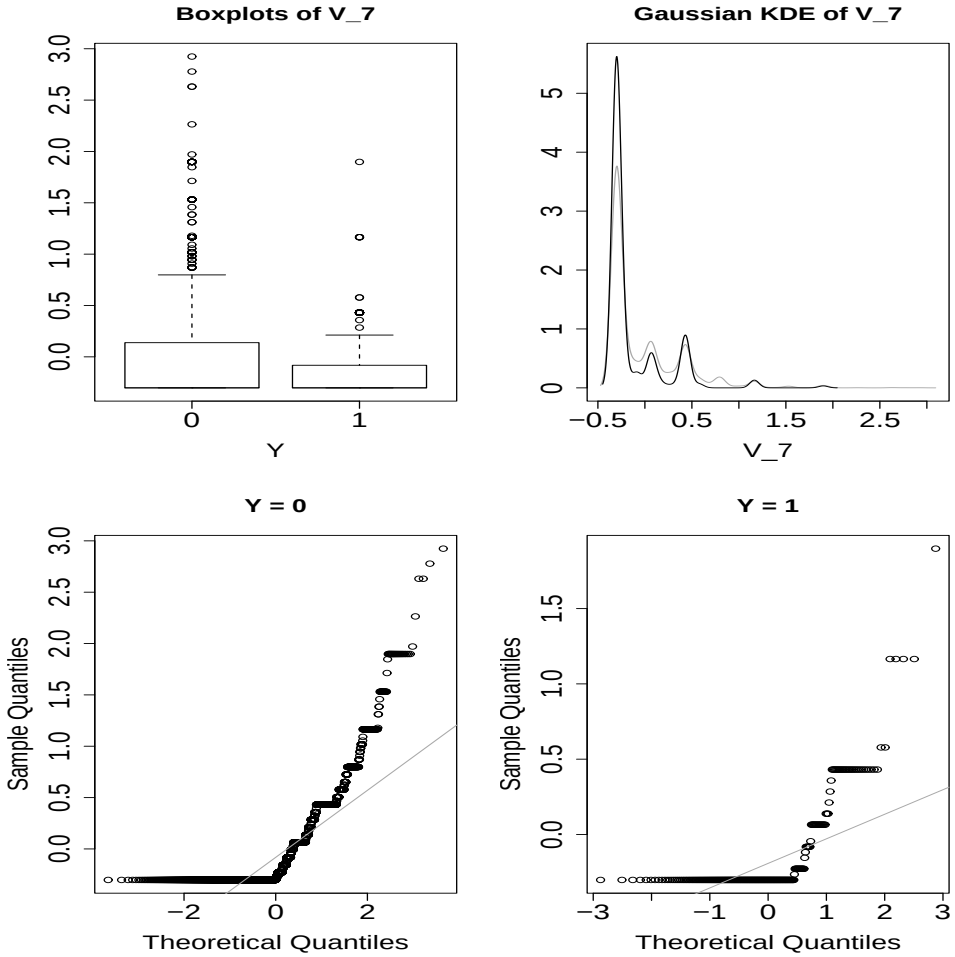
Tabulka B.7: Souhrn charakteristik proměnné V_6.



Obrázek B.6: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_6.

Y	Min.	1. Kvart.	Medián	$\bar{V_7}$	3.Kvart.	Max.	σ_7
0	−0,302	−0,302	−0,302	−0,024	0,138	2,920	0,412
1	−0,302	−0,302	−0,302	−0,132	−0,082	1,900	0,330

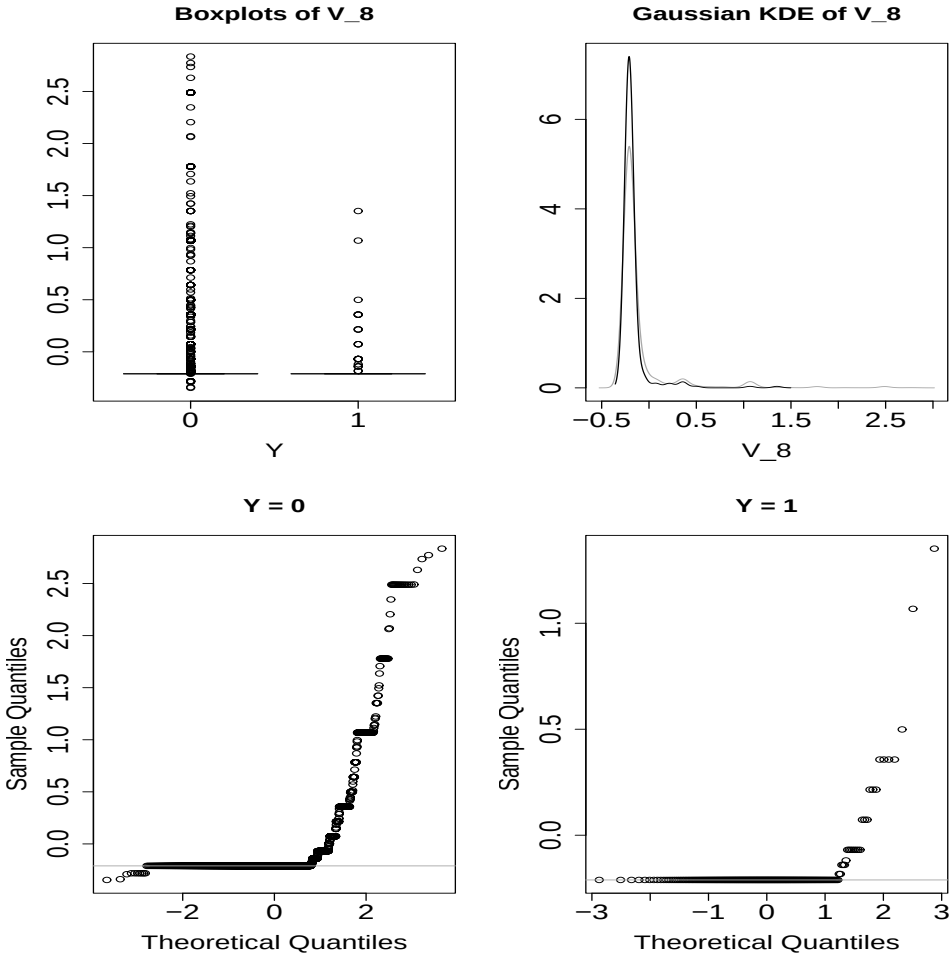
Tabulka B.8: Souhrn charakteristik proměnné V_7.



Obrázek B.7: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_7.

Y	Min.	1. Kvart.	Medián	$\bar{V}_8$	3.Kvart.	Max.	σ_8
0	−0,346	−0,211	−0,211	−0,101	−0,211	2,840	0,348
1	−0,211	−0,211	−0,211	−0,173	−0,211	1,350	0,162

Tabulka B.9: Souhrn charakteristik proměnné V_8.

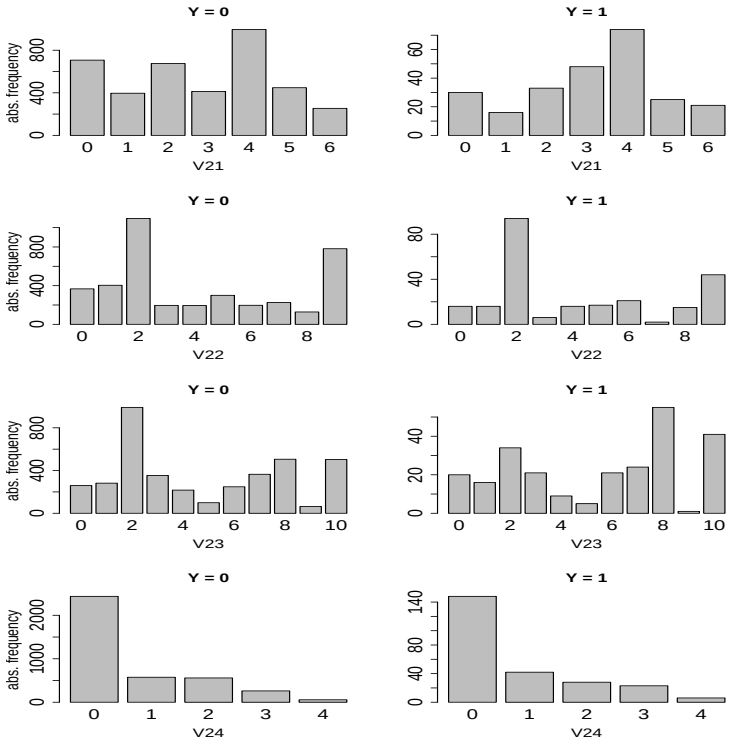


Obrázek B.8: Krabicové diagramy, odhady hustot a normální diagramy pro veličinu V_8.

Dodatek C

K diskrétním veličinám

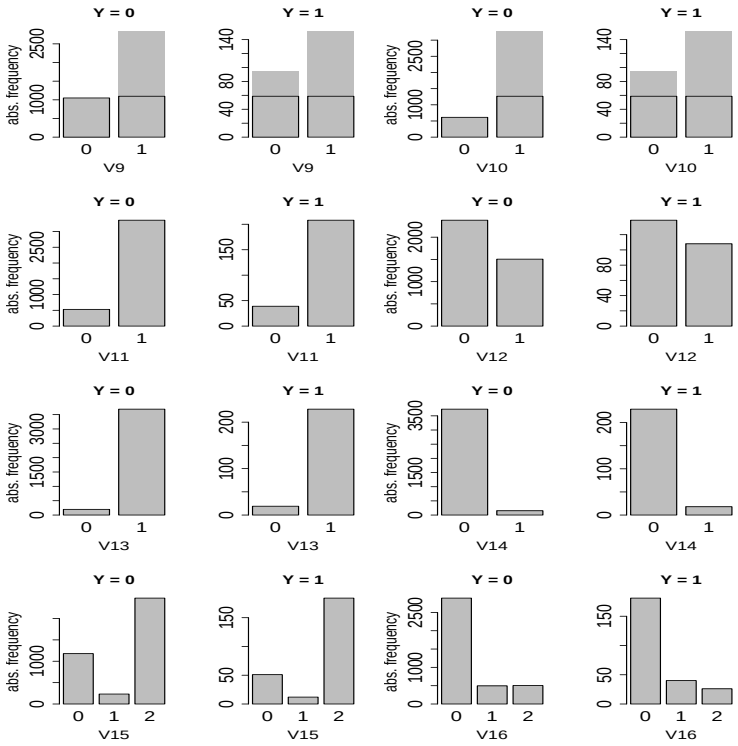
Na následujících stránkách uvádíme absolutní a relativní četnosti jednotlivých kategorií diskrétních veličin a jejich sloupcové diagramy.



Obrázek C.1: Sloupcové diagramy proměnných V_{21}–V_{24}.

V_{21}:	0	737	(17,8%)	V_{23}:	0	279	(6,7%)
	1	412	(10,0%)		1	298	(7,2%)
	2	708	(17,1%)		2	1025	(24,8%)
	3	461	(11,1%)		3	375	(9,1%)
	4	1069	(25,9%)		4	226	(5,5%)
	5	473	(11,4%)		5	104	(2,5%)
	6	275	(6,7%)		6	269	(6,5%)
V_{22}:	0	383	(9,3%)		7	389	(9,4%)
	1	420	(10,2%)		8	561	(13,6%)
	2	1189	(28,8%)		9	65	(1,6%)
	3	202	(4,9%)		10	544	(13,2%)
	4	210	(5,1%)	V_{24}:	0	2584	(62,5%)
	5	317	(7,7%)		1	617	(14,9%)
	6	218	(5,3%)		2	587	(14,2%)
	7	227	(5,5%)		3	284	(6,9%)
	8	143	(3,5%)		4	63	(1,5%)
	9	826	(20,0%)				

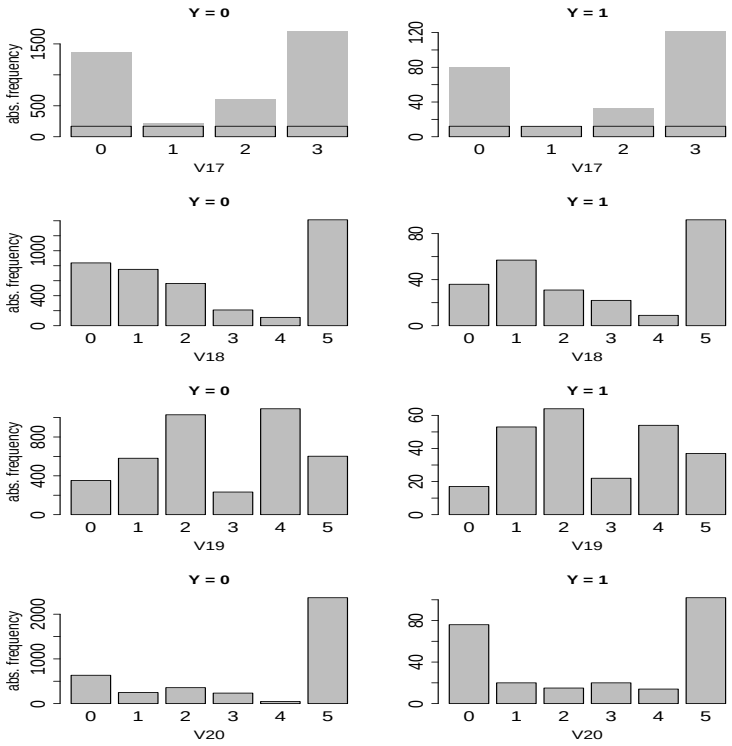
Tabulka C.1: Četnosti jednotlivých kategorií proměnných V_{21}–V_{24}.



Obrázek C.2: Sloupcové diagramy proměnných V_9–V_{16}.

V_9:	0	1145	$(27,7\%)$	V_{10}:	0	708	$(17,1\%)$
	1	2990	$(72,3\%)$		1	3427	$(82,9\%)$
V_{11}:	0	569	$(13,8\%)$	V_{12}:	0	2521	$(61,0\%)$
	1	3566	$(86,2\%)$		1	1614	$(39,0\%)$
V_{13}:	0	217	$(5,2\%)$	V_{14}:	0	3965	$(95,9\%)$
	1	3918	$(94,8\%)$		1	170	$(4,1\%)$
V_{15}:	0	1229	$(29,7\%)$	V_{16}:	0	3071	$(74,3\%)$
	1	245	$(5,9\%)$		1	535	$(12,9\%)$
	2	2661	$(64,4\%)$		2	529	$(12,8\%)$

Tabulka C.2: Četnosti jednotlivých kategorií proměnných V_9–V_{16}.



Obrázek C.3: Sloupcové diagramy proměnných V_{17}–V_{20}.

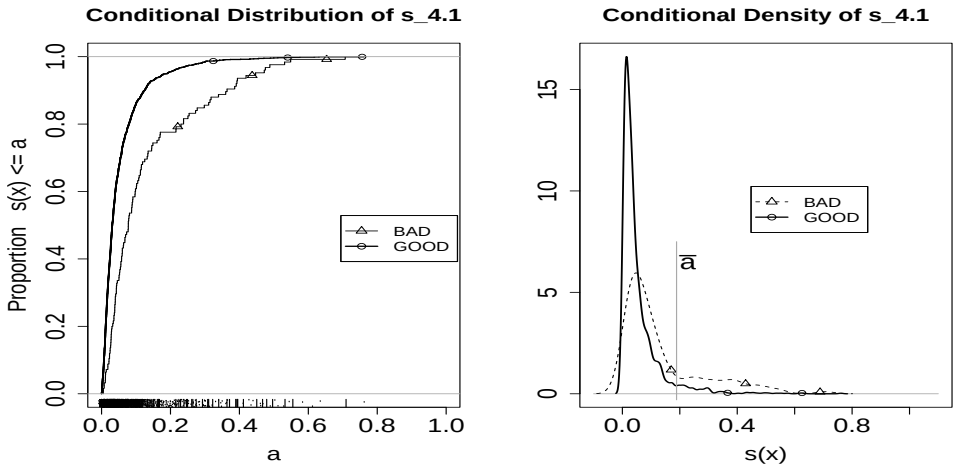
V_{17}:	0	1440	$(34,8\%)$	V_{19}:	0	369	$(8,9\%)$
	1	223	$(5,4\%)$		1	634	$(15,3\%)$
	2	639	$(15,5\%)$		2	1093	$(26,4\%)$
	3	1833	$(44,3\%)$		3	255	$(6,2\%)$
					4	1145	$(27,7\%)$
V_{18}:	0	874	$(21,1\%)$		5	639	$(15,5\%)$
	1	809	$(19,6\%)$	V_{20}:	0	710	$(17,2\%)$
	2	595	$(14,4\%)$		1	267	$(6,5\%)$
	3	232	$(5,6\%)$		2	372	$(9,0\%)$
	4	120	$(2,9\%)$		3	255	$(6,2\%)$
	5	1505	$(36,4\%)$		4	60	$(1,5\%)$
					5	2471	$(59,8\%)$

Tabulka C.3: Četnosti jednotlivých kategorií proměnných V_{17}–V_{20}.

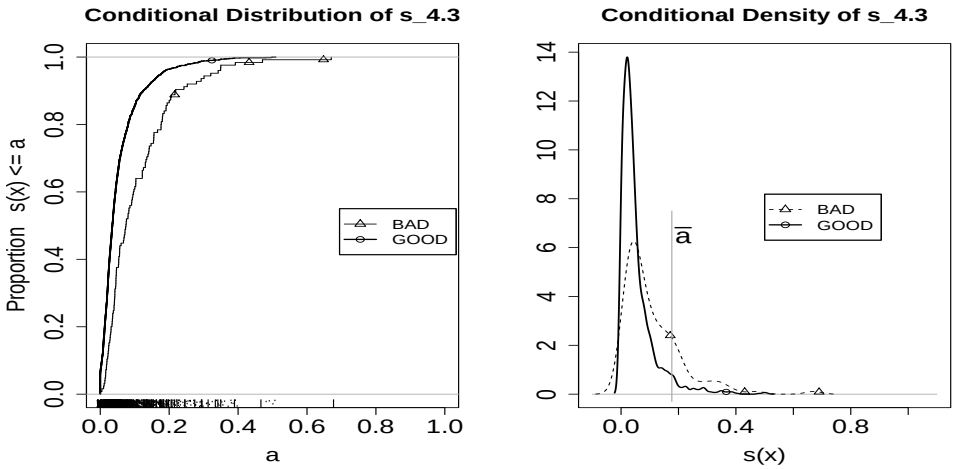
Dodatek D

K logistické regresi

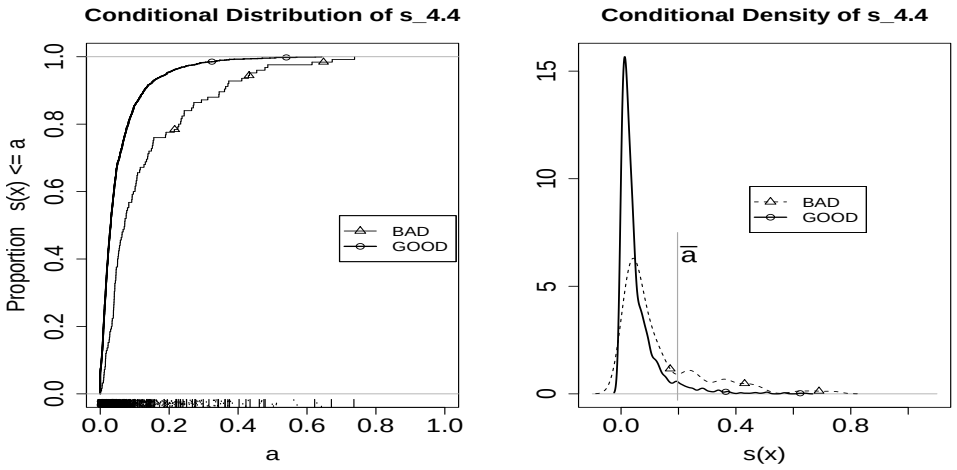
V tomto dodatku uvádíme grafy empirických distribučních funkcí $\hat{G}_0, \hat{G}_1$ náhodné veličiny $s(\mathbf{X}_K)$ za podmínky $[Y_K = 0]$ a $[Y_K = 1]$ včetně odhadů hustot $\hat{g}_0, \hat{g}_1$ (normální jádro, šířka okna rovna směrodatné odchylce) těchto veličin pro modely uvedené ve 4.kapitole.



Obrázek D.1: Empirické distribuční funkce $\hat{G}_0, \hat{G}_1$ a odhady hustot $\hat{g}_0, \hat{g}_1$ pro model 4.1 ($\bar{a} = 0,189$).



Obrázek D.2: Empirické distribuční funkce $\hat{G}_0, \hat{G}_1$ a odhady hustot $\hat{g}_0, \hat{g}_1$ pro model 4.3 ($\bar{a} = 0,177$).



Obrázek D.3: Empirické distribuční funkce $\hat{G}_0, \hat{G}_1$ a odhady hustot $\hat{g}_0, \hat{g}_1$ pro model 4.4 ($\bar{a} = 0,197$).

Dodatek E

Ke klasifikačním stromům

V tomto dodatku uvádíme podrobně rozepsanou strukturu stromů 5.2 a 5.3. Každý řádek obsahuje číslo uzlu, jméno proměnné, podle které proběhlo dělení včetně jejích hodnot (resp. seznamu kategorií v případě kategorické veličiny), počet objektů v daném uzlu, hodnotu jeho deviance, třídu přiřazenou tomuto uzlu a v závorce jsou uvedeny pravděpodobnosti náležení do jednotlivých tříd. Strom je číslován tak, aby na i-té hladině stromu byly uzly s indexem 2^i až $2^{i+1} - 1$, kde kořen je na nulté hladině. V případě zastavení dělení některého uzlu vynecháme na následující hladině čísla, která by příslušela dceřinným uzlům tohoto uzlu. Koncové listy jsou značeny hvězdičkou.

```
Variables actually used in tree construction:
 [1] "V10" "V8"  "V23" "V7"  "V19" "V18" "V4"  "V6"  "V20" "V1"  "V21" "V22"
Number of terminal nodes:  20
Residual mean deviance:  0.403 = 1660 / 4120
Misclassification error rate: 0.0534 = 221 / 4135
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

  1) root 4135 1900.0 0 ( 0.940 0.060 )
    2) V10: 0 708  560.0 0 ( 0.866 0.134 )
      4) V8 < -0.204 575  490.0 0 ( 0.849 0.151 )
        8) V23: 1,2,3,4,5,6,7,9 360  240.0 0 ( 0.897 0.103 )
         16) V7 < -0.1915 194  160.0 0 ( 0.851 0.149 )
           32) V7 < -0.2945 175  140.0 0 ( 0.869 0.131 ) *
           33) V7 > -0.2945 19   24.0 0 ( 0.684 0.316 )
             66) V19: 2,4,5 12    6.9 0 ( 0.917 0.083 ) *
             67) V19: 0,1,3 7    8.4 1 ( 0.286 0.714 ) *
```

```
      17) V7 > -0.1915 166   64.0 0 ( 0.952 0.048 ) *
     9) V23: 0,8,10 215  230.0 0 ( 0.767 0.233 )
      18) V7 < -0.2835 129  150.0 0 ( 0.736 0.264 )
        36) V18: 0,2 43   27.0 0 ( 0.907 0.093 ) *
        37) V18: 1,3,4,5 86  110.0 0 ( 0.651 0.349 )
          74) V4 < -0.7595 18   21.0 0 ( 0.722 0.278 )
           148) V6 < -0.2115 6    7.6 1 ( 0.333 0.667 ) *
           149) V6 > -0.2115 12    6.9 0 ( 0.917 0.083 ) *
          75) V4 > -0.7595 68   89.0 0 ( 0.632 0.368 )
           150) V20: 2,5 41   46.0 0 ( 0.756 0.244 )
             300) V1 < 1.1555 33   24.0 0 ( 0.879 0.121 ) *
             301) V1 > 1.1555 8    9.0 1 ( 0.250 0.750 ) *
           151) V20: 0,1,3,4 27   37.0 1 ( 0.444 0.556 )
             302) V21: 0,1,2,4 14   17.0 0 ( 0.714 0.286 )
               604) V23: 0,8 9    6.3 0 ( 0.889 0.111 ) *
               605) V23: 10 5    6.7 1 ( 0.400 0.600 ) *
             303) V21: 3,5,6 13   11.0 1 ( 0.154 0.846 ) *
      19) V7 > -0.2835 86   83.0 0 ( 0.814 0.186 )
        38) V22: 0,1,3,5,7 29    8.7 0 ( 0.966 0.034 ) *
        39) V22: 2,4,6,8,9 57   66.0 0 ( 0.737 0.263 )
          78) V6 < 0.348 48   46.0 0 ( 0.813 0.188 )
           156) V20: 0,2,5 42   31.0 0 ( 0.881 0.119 ) *
           157) V20: 1,3,4 6    7.6 1 ( 0.333 0.667 ) *
          79) V6 > 0.348 9   11.0 1 ( 0.333 0.667 ) *
   5) V8 > -0.204 133   60.0 0 ( 0.940 0.060 )
    10) V6 < 0.564 121   35.0 0 ( 0.967 0.033 ) *
    11) V6 > 0.564 12   15.0 0 ( 0.667 0.333 )
      22) V1 < 0.445 6    0.0 0 ( 1.000 0.000 ) *
      23) V1 > 0.445 6    7.6 1 ( 0.333 0.667 ) *
 3) V10: 1 3427 1200.0 0 ( 0.956 0.044 ) *
```

```
Variables actually used in tree construction:
 [1] "V10" "V8"  "V23" "V19" "V20" "V22" "V7"  "V21" "V2"  "V18" "V6"
Number of terminal nodes:  14
Residual mean deviance:  0.392 = 800 / 2040
Misclassification error rate: 0.0531 = 109 / 2053
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

  1) root 2053 940.0 0 ( 0.940 0.060 )
    2) V10: 0 346 290.0 0 ( 0.853 0.147 )
      4) V8 < -0.197 278 250.0 0 ( 0.838 0.162 )
        8) V23: 1,2,3,4,5,6,7,9 173 100.0 0 ( 0.913 0.087 )
         16) V19: 2,4,5 126  48.0 0 ( 0.952 0.048 ) *
         17) V19: 0,1,3 47  46.0 0 ( 0.809 0.191 )
           34) V20: 1,2,5 26  14.0 0 ( 0.923 0.077 ) *
           35) V20: 0,3,4 21  27.0 0 ( 0.667 0.333 )
             70) V22: 2,3,4,5,6,7 14  11.0 0 ( 0.857 0.143 ) *
             71) V22: 1,8,9 7   8.4 1 ( 0.286 0.714 ) *
        9) V23: 0,8,10 105 130.0 0 ( 0.714 0.286 )
         18) V7 < -0.2835 63  80.0 0 ( 0.667 0.333 )
           36) V21: 0,1,2,3,4 52  58.0 0 ( 0.750 0.250 )
             72) V22: 0,1,2,5,7,9 34  28.0 0 ( 0.853 0.147 ) *
             73) V22: 3,4,6,8 18  25.0 0 ( 0.556 0.444 )
              146) V2 < 0.512 13  17.0 1 ( 0.385 0.615 ) *
              147) V2 > 0.512 5   0.0 0 ( 1.000 0.000 ) *
           37) V21: 5,6 11  13.0 1 ( 0.273 0.727 )
             74) V18: 2,5 5   6.7 0 ( 0.600 0.400 ) *
             75) V18: 0,1 6   0.0 1 ( 0.000 1.000 ) *
         19) V7 > -0.2835 42  44.0 0 ( 0.786 0.214 ) *
      5) V8 > -0.197 68  41.0 0 ( 0.912 0.088 )
       10) V6 < 0.427 55  10.0 0 ( 0.982 0.018 ) *
       11) V6 > 0.427 13  17.0 0 ( 0.615 0.385 )
         22) V22: 6,7,9 6   0.0 0 ( 1.000 0.000 ) *
         23) V22: 2,5 7   8.4 1 ( 0.286 0.714 ) *
    3) V10: 1 1707 600.0 0 ( 0.957 0.043 ) *
```

Literatura

Anděl, J. (1985): *Matematická statistika*. SNTL, Praha.

Arminger, G., Enache, D. & Bonne, T. (1997): Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks. *Computational Statistics* **12**, 293–310.

Breimann, L., Friedmann, J.H., Olshen, R.A. & Stone, C.J. (1984): *Classification and Regression Trees*. Wadsworth Publishers.

Fahrmeir, L. & Hamerle, A. (1984): *Multivariate statistische Verfahren*. Walter de Gruyter, Berlin.

Hand, D.J. (1997): *Construction and Assessment of Classification Rules*. Wiley Publishers.

Hand, D.J. & Henley, W.E. (1996): A k-nearest neighbour classifier for assessing consumer credit risk. *The Statistician* **45**, No.1, 77–95.

Härdle, W., Müller, M. & Rönz, B. (2001): *Credit Scoring*. Humboldtova Univerzita, Berlín.

Härdle, W., Müller, M., Sperlich, S. & Werwatz, A. (2000): *A Course on Non- and Semiparametric Modelling*. Skripta k semestrální přednášce, Humboldtova Univerzita, Berlín.

Härdle, W. & Simar, L. (2002): *Applied Multivariate Statistical Analysis*. Skripta k semestrální přednášce, Humboldtova Univerzita, Berlín.

Harrell Jr., F. E. (2002): *Regression Modeling Strategies (With Applications to Linear Models, Logistic Regression, and Survival Analysis)*. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2001): *The Elements of Statistical Learning*. Springer.

Ihaka, R. & Gentleman, R. (1996): R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **Vol.5**, No.3, Pg. 299–314.

Kaiser, U. & Szczesny, A. (2000): Einfache ökonometrische Verfahren für die Kreditrisikomessung: Logit- und Probit-Modelle. *Working Paper Series: Finance and Accounting* No.61, Goethe Universität Frankfurt am Main.

Kolektiv autorů (2001): *S-PLUS 6 for Windows Guide to Statistics*. Insightful Corporation, Seattle.

Komorád, K. (2002): On Credit Scoring Estimation. *Master's Thesis*, Humboldt-University Berlin.

Lewis, E.M. (1994): *An Introduction to Credit Scoring*. San Raphael, California, The Athena.

Müller, M. & Rönz, B. (1999): Credit Scoring using Semiparametric Methods. *Proceedings of Measuring Risk in Complex Statistical Systems*, Humboldt-University Berlin.

Ripley, B.D. (1996): *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rönz, B. (1998): *Computergestützte Statistik*. Humboldt-Universität, Berlin.

Thomas, L.C. (2000): A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16**, Pg. 149–172.

Venables, W.N. & Ripley, B.D. (1994): *Modern Applied Statistics with S-Plus*. Springer.

Zvára, K. (2003): *R & Regrese*. Poznámky k semestrální přednášce 2003/2004, MFF UK Praha.