Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Jaromír Malec

Vybrané rizikové parametry v IRB přístupu a jejich modelování

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Ing. Miloš Kopa, Ph. D.

Studijní program: Matematika

Studijní obor: Finanční a pojistná matematika

Praha 2013

Na tomto místě bych rád poděkoval panu RNDr. Ing. Miloši Kopovi, Ph. D., vedoucímu diplomové práce, za jeho cenné rady a nápady, které vedly k úspěšnému napsání diplomové práce. Zároveň děkuji bance, která působí na bankovním trhu v České republice, za poskytnutí dat a zapůjčení literatury.

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

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

V Praskolesích dne 27.7.2013 Jaromír Malec

Název práce: Vybrané rizikové parametry v IRB přístupu a jejich modelování

Autor: Jaromír Malec

Katedra / Ústav: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Ing. Miloš Kopa, Ph. D.
Email vedoucího: kopa@karlin.mff.cuni.cz

Abstrakt: Mezi nejdůležitější oblasti činnosti bank patří stanovení úvěrového (kreditního) rizika. Tato práce pojednává o IRB přístupu na základě BASEL II. Mezi nejdůležitější parametry v rámci tohoto přístupu patří parametry LGD, EAD a PD. Všechny parametry banka individuálně modeluje pomocí regulátorem povolených modelů. Práce se nejvíce zaměřuje na parametr PD. Teorie v případě tohoto parametru je do značné míry rozvinuta. Nicméně v poslední době se začíná projevovat potřeba modelovat parametr PD pro více let. Parametru LGD se tato práce též věnuje. Poměrně okrajově se dotýká parametru EAD. Práce nejprve pojednává o IRB přístupu, regresních modelech a hodnotících ukazatelích a pak se věnuje výše uvedeným parametrům.

Klíčová slova: kreditní riziko, LGD modely, PD modely, EAD modely, logistická regrese

Title: Selected risk parameters in IRB approach and their modeling

Author: Jaromír Malec

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Ing. Miloš Kopa, Ph. D.

Supervisor's email address: kopa@karlin.mff.cuni.cz

Abstract: The determination of lending (credit) risk is one of the most important fields of bank activities. This thesis discusses the IRB approach under Basel II. This approach includes the LGD, EAD and PD parameters. All parameters are individually modelled by the bank using regulator approved models. Parameter PD is the most focused one in this thesis. Theory for this parameter is of interest in many papers. However, at present the need for modelling of PD parameter over more years is appearing. Parameter LGD is also discussed in this thesis. The parameter EAD is only briefly presented. The thesis begins with the IRB approach, regression models and evaluation indicators, and then it focuses on the above parameters.

Keywords: credit risk, LGD models, PD models, EAD models, logistic regression

Obsah

Předmluva 1

1. Úvod 2

1.1. Basilejské předpisy a jejich pilíře 2

1.2. IRB přístup 3

1.3. IRB přístup versus standardní přístup 4

2. Modelování parametru M 8

3. Binární logistická regrese 13

3.1. Model binární logistické regrese a jeho odhad 13

3.2. Testy významnosti modelu 18

3.3. Algoritmus Stepwise pro výběr proměnných 20

3.4. Verifikace modelu 24

4. Diverzifikační schopnosti modelu 27

5. Parametr PD 33

5.1. Požadavky kladené na parametr PD 33

5.2. Modelování parametru PD 34

5.2.1. Používané metody 34

5.2.2. Postup stavby modelu 36

5.3. Výstavba modelu binární logistické regrese a jeho alternativy 38

5.3.1. Data 38

5.3.2. Výběr proměnných pro odhad modelu 40

5.3.3. Odhad modelu, jeho vlastnosti a ověřování 43

5.3.4. Alternativní možnosti k modelu 46

5.3.5. Srovnání s modelem lineární regrese 52

5.3.6. Kategorizace a faktorizace dat 53

5.3.7. Odhad modelu při použití dummy proměnných 55

5.3.8. Kalibrace modelu 61

5.4. Další příklady modelů 62

5.4.1. Altmanův model 62

5.4.2. Srovnání Altmanova modelu s modelem binární logistické regrese 65

5.4.3. KMV model 67

5.5. Problém víceletého PD 69

6. LGD: ztráta banky při selhání 74

6.1. Úvod do modelování parametru LGD 74

6.2. Praktická metoda používaná k modelování 77

6.2.1. Definice pojmů 77

6.2.2. Modelování PC 79

6.2.3. Modelování $\mathrm{LGD}_{\mathrm{cured}}$ 79

6.2.4. Modelování $\mathrm{LGD}_{\mathrm{irr}}$ 80

6.2.5. Smíšený model, jeho kalibrace a testování 86

6.2.6. Volba pozorovacího dne 91

6.3. Ilustrace používané metody 91

6.3.1. Meziroční srovnání výsledků modelu 99

7. Parametr EAD: expozice v selhání 101

Závěr 107

Seznam použité literatury 108

Předmluva

Tento text se zabývá rizikovými parametry v IRB přístupu, tedy přístupu, který používají banky při modelování kreditního rizika. Souvisí nejen s bankami, ale též s pravděpodobností a statistikou. Autor předpokládá základní znalost pojmů z teorie pravděpodobnosti a matematické statistiky (například náhodná veličina a její rozdělení, střední hodnota a průměr jako její odhad, rozptyl a jeho odhad a další pojmy). Speciálně je nutné znát model lineární regrese, který je na několika místech zmíněn v rámci numerických ilustrací. Dále je vhodné znát obecnou teorii zobecněných lineárních modelů, nicméně autor nepředpokládá její znalost a speciálně ji nevyužívá. Autor dále předpokládá znalost složeného úročení a diskontování a pojmů z této oblasti (například úroková míra a úrok) a též znalost základních typů úvěrů. Informace o pojmech z oblasti pravděpodobnosti a statistiky poskytuje jakákoli základní literatura o pravděpodobnosti a statistice, mírně pokročilejší literatura poskytuje informace o modelu lineární regrese, základní literatura z oblasti finanční matematiky poskytuje informace o úročení a diskontování, informace o zobecněných lineárních modelech lze najít v [8].

První kapitola krátce pojednává o smlouvě BASEL II, o standardním a IRB přístupu a oba přístupy srovnává. Druhá kapitola pojednává o parametru M (splatnost), který se v rámci IRB přístupu odhaduje. Třetí kapitola pojednává o modelu logistické regrese, jeho odhadu, vlastnostech, testech významnosti proměnných a správnosti modelu a algoritmu pro výběr proměnných. Čtvrtá kapitola pojednává o diverzifikačních schopnostech modelu, který předpovídá selhání klienta. Pátá kapitola pojednává o parametru PD, zavádí pojem selhání, numericky ilustruje průběh algoritmu pro výběr proměnných v modelu logistické regrese, ilustruje odhad modelu logistické regrese, hodnotí jeho vlastnosti a navrhuje jeho alternativy, srovnává ho s modelem lineární regrese a srovnává též použití spojitých proměnných a dummy proměnných, přičemž oba postupy též představuje. Zároveň krátce pojednává o často citovaném Altmanově modelu, který srovnává s modelem logistické regrese, a o Kealhoferově – McQuownově – Vašíčkově modelu. Součástí této kapitoly je též jednoduchý návrh řešení problému víceletého PD. Víceletý PD se v současné době z legislativních důvodů neodhaduje, ale parametr PD se odhaduje každý rok zvlášť (jiný postup bankovní legislativa nedovoluje). Na okraj tohoto pojednání je též navrhnut model „useknutého“ geometrického rozdělení. Šestá kapitola pojednává o parametru LGD, představuje jednu z metod používaných k jeho modelování a na konkrétních datech ji ilustruje. Sedmá kapitola krátce pojednává o parametru EAD a ukazuje jeho typy.

Při numerických ilustracích byl použit program Mathematica, verze 8. Znak Q. E. D. na konci řádku značí konec důkazu, znak $\square$ na konci řádku značí konec příkladu. Přiložené CD obsahuje data použitá při numerických ilustracích a veškeré příkazy použité při programování včetně jednotlivých výstupů.

1. Úvod

V této kapitole postupujeme podle [1].

1.1. Basilejské předpisy a jejich pilíře

Po krachu banky Bankhaus Herstatt v Západním Německu v roce 1974 byla skupinou deseti centrálních bank ustanovena komise The Basel Committee on Banking Supervision (nadále BCBS, viz též [13]), jejímž hlavním cílem byla podpora stability v mezinárodním bankovním systému ve snaze o provádění dvou principů. První princip spočívá v tom, že zahraniční banky se nemají vyhýbat kontrole. Dle druhého principu musí být tato kontrola dostatečná pro všechny banky, které operují na mezinárodním trhu. Primárním záměrem této komise je tvorba standardů, směrnic a nejlepších postupů, které každý představitel zavede.

V současnosti se vzájemně prosazují tři základní pilíře. První pilíř spočívá ve vývojových přístupech k řízení kreditního, tržního a operačního rizika a stanovuje minimální kapitálové požadavky. Druhý pilíř se týká procesu dohledu. K tomu se používají čtyři principy: vnitřní posudky, kontrolní zásahy, minimální kapitálový požadavek, kontrolní přezkoumání. V rámci tohoto pilíře se hodnotí riziko a kapitálová přiměřenost jednotlivých bank. Třetí pilíř se týká transparentnosti a tržní disciplíny. V rámci tohoto pilíře se požaduje zvýšená informační otevřenost o kapitálových požadavcích a metodách pro hodnocení rizika.

Komise BCBS v minulosti připravila dvě smlouvy týkající se bankovního práva a regulace. V roce 2006 začala platit druhá z Basilejských smluv (BASEL II), která nahradila původní koncept BASEL I. BASEL II v sobě zahrnuje všechny tři výše popsané pilíře. Pravidla BASEL II používají holdingy i mezinárodní banky, týkají se všech finančních institucí a pojišťoven, které však mají své smlouvy (SOLVENCY I a II). Kontrolu domácích i zahraničních bank provádí národní dohled. Druhý a třetí pilíř přesahují rámec tohoto textu, následující řádky a kapitoly se věnují prvnímu pilíři, jehož součástí je IRB přístup ke kreditnímu riziku.

Jak již bylo zmíněno výše, první pilíř se týká tržního, operačního a kreditního rizika. Tržní riziko je riziko ztráty banky, která plyne ze změn cen, kurzů a sazeb na finančních trzích. Je to též možná ztráta portfolia v důsledku změny tržních podmínek nebo potenciální změna v hodnotě aktiva nebo derivátů v důsledku změny v základních zdrojích tržní nejistoty. Je s ním spojena též nejistota budoucích zisků vyplývající ze změny tržních podmínek. Tržní rizika představují nejvýše 20% všech bankovních rizik. Operační riziko je riziko ztráty vlivem vnitřního selhání nebo vlivem vnějších skutečností. Toto riziko představuje zhruba 30% všech bankovních rizik. Tržní a operační riziko přesahují rámec textu, navíc nejsou tak významná jako kreditní riziko.

Kreditní riziko je riziko ztráty plynoucí ze selhání smluvní strany tím, že nedostojí svým závazkům vůči bance. Zahrnuje v sobě též potenciální ztrátu při (částečném) selhání protistrany. Představuje 50 – 70% všech bankovních rizik. Metodika BASEL I zavádí pouze standardní přístup. Tento přístup se v průběhu let stal nevýhodným pro banky. Metodika BASEL II přináší IRB přístup, který používá ekono-

mické modely. Oba přístupy modelují hodnoty rizikově vážených aktiv RWA (z angl. Risk – weighted asset), které slouží k určení kapitálové přiměřenosti CAD (z angl. Capital adequacy) a následně kapitálového požadavku dle:

$$\mathrm{CAD} = \frac{\mathrm{kapitál}}{\mathrm{RWA}},$$

přičemž CAD musí dle BASEL II splňovat:

$$\mathrm{CAD} \geq 8\%.$$

Standardní přístup podle [9] modeluje RWA pomocí:

$$\mathrm{RWA} = \sum_{i} w_i \, A_i \, ,$$

kde:

w_i – i – tá riziková váha,
A_i – i – té aktivum, respektive jeho hodnota.

Váhy jsou určeny deterministicky například na základě vnějšího ratingu.

V následující části pojednáme o tom, jak tyto hodnoty modeluje IRB přístup. O výhodách IRB přístupu proti standardnímu přístupu pojednáme později.

1.2. IRB přístup

IRB přístup (z angl. Internal rating based) modeluje RWA na základě vnitřního ratingu pomocí funkce šesti parametrů f:

$$\mathrm{RWA} = \mathrm{f}(\mathrm{PD}, \mathrm{LGD}, \mathrm{EAD}, \mathrm{M}, \mathrm{CI}, \rho),$$

kde:

PD – Pravděpodobnost selhání (z angl. Probability of default),
LGD – Ztráta při selhání (z angl. Loss given default),
EAD – Expozice v selhání (z angl. Exposure at default), například dlužná částka při selhání,
M – Splatnost (z angl. Maturity),
CI – Interval spolehlivosti (z angl. Confidence interval),
ρ – Korelace mezi aktivy.

Funkce f se nazývá RWA – funkce. Tuto funkci určují tzv. expoziční třídy. První třídou jsou firemní a vedoucí banky, dalšími třídami pak jsou maloobchod, vlastní kapitál, uznatelné nakoupené pohledávky a ochranné expozice.

Banka všechny parametry nemodeluje. Parametry CI a ρ zpravidla určuje regulátor, v případě ostatních čtyř parametrů záleží na tom, který přístup banka zvolí. Existují dva IRB přístupy: FIRB (základní IRB, z angl. Foundation IRB) a AIRB

(pokročilý IRB, z angl. Advanced IRB). Použije – li banka AIRB přístup, pak na základě svých dat a regulátorem schválených modelů odhaduje všechny čtyři parametry. V případě použití FIRB přístupu modeluje pouze parametr PD, ostatní tři parametry buď určí regulátor, nebo je odhadne banka na základě dat od regulátora.

IRB přístup má tři složky, které se musí určit pro všechny expoziční třídy. Nejprve se musí určit první složka, která zahrnuje všechny rizikové komponenty. S pomocí těchto rizikových komponent se spočítá druhá složka, RWA – funkce. Z druhé složky se určí kapitálové požadavky. Z obou složek a z kapitálových požadavků se určí třetí složka, minimální požadavky, která je, společně s kapitálovými požadavky, hlavním výstupem IRB přístupu.

Pro použití IRB přístupu je zavedeno jedno pravidlo: pokud se banka rozhodne použít IRB přístup, musí ho použít na všechna aktiva. I z tohoto pravidla existují výjimky. IRB přístup se nemusí použít, pokud jsou daným aktivem specializované půjčky, ochrana banky nebo nemateriální expozice. Pro ostatní aktiva se při použití IRB přístupu musí též vytvořit tzv. roll out plán, který se týká poskytování zpráv veřejnosti. Je též potřeba mít pro toto zveřejnění správné načasování a zdroje. Vše musí být schváleno regulátorem.

IRB přístup je zaveden ve všech národních legislativách států Evropské unie. Pro jeho používání bylo stanoveno tříleté přechodné období. Důvodem pro ustanovení tohoto období je potřeba uvolnění dat z minulosti, která jsou nutná pro výpočet parametrů RWA funkce. Například pro výpočet parametru LGD je třeba mít data za posledních sedm let apod.

1.3. IRB přístup versus standardní přístup

Zodpovězme nyní otázku, proč je výhodné, aby banky používaly IRB přístup. Standardní přístup má oproti IRB přístupu několik výhod. Mezi základní výhody patří jeho jednoduchost. Navíc má méně předpokladů než IRB přístup, jehož předpoklady plynou z metod používaných pro modelování jednotlivých parametrů, protože k tomuto modelování zpravidla slouží regrese (ať už lineární nebo nějaký zobecněný lineární model), která má několik předpokladů, které často nejsou splněny. K tomu všemu standardní přístup ne vždy přinese vyšší kapitálový požadavek než IRB přístup.

Proto je namístě otázka, v čem spočívá výhoda IRB přístupu proti standardnímu přístupu. Pokud bychom měli říci jen jeden důvod pro použití IRB přístupu, pak je to fakt, že banky by měli být řízeny ekonomicky, což umožňuje IRB přístup, zatímco při použití standardního přístupu jsou banky z části řízeny zvenčí. Tabulka 1 srovnává standardní přístup a IRB přístup. Z tabulky 1 plyne hlavní výhoda IRB přístupu: banka provádí veškeré hodnocení sama na základě ekonomických modelů, které jsou sice složitější, ale banka na jejich základě získá více informací než při použití standardního přístupu.

	Standardní přístup	IRB přístup
Posouzení individuálního rizika	• Pouze vnější rating • Implikovaný PD daný regulátorem • Velká část portfolia není zhodnocena • Nelze dopočítat LGD • Omezené uznání záruk	• Vnitřní rating a PD • Lze pokrýt celé portfolio • Vnitřní modelování parametru LGD • Rozšířené uznání záruk
Posouzení rizika portfolia	• Jednoduché váhy • Bez ekonomických základů	• Složitější funkce sjednocující rizikové parametry • Na základě nezávislých ekonomických modelů

Tabulka 1 – srovnání standardního a IRB přístupu

Numerická ilustrace (viz též [14])

Ilustrace vychází z dat získaných studií QIS3 v roce 2003, která sledovala kapitálové požadavky. Tato studie rozděluje banky do dvou skupin: skupinu 1 tvoří velké, diverzifikované a mezinárodně aktivní banky, skupinu 2 tvoří malé a často více specializované banky. Pro skupinu 2 studie neuvádí výsledky při použití AIRB přístupu. Jedním z výsledků této studie bylo sledování průměrného přínosu daného aktiva ke změně kapitálového požadavku při použití daného přístupu. Výsledky jsou v tabulce 2, ve které jsou příspěvky daného portfolia k celkové změně kapitálu v %. Z této tabulky plyne průměrný zisk při použití AIRB přístupu.

Dalším výsledkem ze studie byly změny v kapitálu podle jednotlivých portfolií. Jejich vyjádření v % je v tabulce 3. Z této tabulky vyplývá, že nejlepšího výsledku dosáhl maloobchod, zatímco čisté jmění dopadlo nejhůře.

Nejdůležitějším výsledkem studie byly celkové změny v kapitálových požadavcích. Jejich maximální, minimální a průměrné hodnoty uvádí tabulka 4. Z této tabulky vyplývá, že průměrná odchylka je poměrně nízká. Tato skutečnost může být pro banku velmi významná, protože nízká odchylka znamená přesné výsledky.

Přístup:	Standardní		FIRB		AIRB
Portfolio	*Skupina 1*	*Skupina 2*	*Skupina 1*	*Skupina 2*	*Skupina 1*
Firmy	1%	−1%	−2%	−4%	−4%
Vedoucí banky	0%	0%	2%	0%	1%
Banky	2%	0%	2%	−1%	0%
Maloobchod	−5%	−10%	−9%	−17%	−9%
Malé a střední podniky	−1%	−2%	−2%	−4%	−3%
Chráněná aktiva	1%	0%	0%	−1%	0%
Ostatní	2%	1%	4%	3%	2%
Celkové kreditní riziko	0%	−11%	−7%	−27%	−13%
Operační riziko	10%	15%	10%	7%	11%
Celková změna	11%	3%	3%	−19%	−2%

Tabulka 2 – příspěvky daného portfolia k celkové změně kapitálu

Přístup:	Standardní		FIRB		AIRB
Portfolio	*Skupina 1*	*Skupina 2*	*Skupina 1*	*Skupina 2*	*Skupina 1*
Firmy	1%	−10%	−9%	−27%	−14%
Vedoucí banky	19%	1%	47%	51%	28%
Banky	43%	15%	45%	−5%	16%
Maloobchod (celkem)	−21%	−19%	−47%	−54%	−50%
− hypotéky	−20%	−14%	−56%	−55%	−60%
− nehypoteční úvěry	−22%	−19%	−34%	−27%	−41%
− revolving	−14%	−8%	−3%	−33%	14%
Malé a střední podniky	−3%	−5%	−14%	−17%	−13%
Specializované půjčky	2%	2%	neuvedeno	neuvedeno	neuvedeno
Čisté jmění	6%	8%	115%	81%	114%
Obchodní knihy	12%	4%	5%	4%	2%
Chráněná aktiva	86%	61%	103%	62%	129%

Tabulka 3 – změny v kapitálu v samostatných portfoliích

Přístup:	Standardní		FIRB		AIRB
	Skupina 1	*Skupina 2*	*Skupina 1*	*Skupina 2*	*Skupina 1*
Minimum	−15%	−23%	−32%	−58%	−36%
Maximum	84%	81%	55%	41%	46%
Průměr	11%	3%	3%	−19%	−2%

Tabulka 4 – celkové změny v kapitálových požadavcích

2. Modelování parametru M

V této kapitole pojednáme o modelování parametru M. Pojednání o tomto parametru nezapadá do kontextu tohoto textu, ale parametr M je bankou modelovaným parametrem. Parametr M se modeluje deterministicky. Záleží na tom, zda banka používá FIRB přístup nebo AIRB přístup. Dle [1], pokud banka použije FIRB přístup, pak:

$$\mathrm{M} = 2{,}5 \text{ roku},$$

přičemž pro repo – operace je:

$$\mathrm{M} = 6 \text{ měsíců}.$$

Nyní předpokládejme, že aktivum, pro které počítáme parametr M, přináší n plateb stejné výše CF. Pak můžeme použít AIRB přístup a dle [1] modelovat M podle:

$$\mathrm{M} = \frac{\sum_{t=1}^{n} t\,\mathrm{CF}}{\sum_{t=1}^{n} \mathrm{CF}} = \frac{\sum_{t=1}^{n} t}{n} = \frac{\frac{n\,(n+1)}{2}}{n} = \frac{n+1}{2}\,. \tag{2.1}$$

M se v tomto případě interpretuje jako střední doba splatnosti aktiva. Má platit:

$$1 \text{ rok} \leq \mathrm{M} \leq 5 \text{ let}.$$

Tento požadavek do značné míry vymezuje aktiva, pro která lze vzorec (2.1) použít. Pokud banka tento vzorec nemůže použít, dosadí za M finální splatnost daného aktiva, tedy smluvně danou dobu trvání aktiva. Horní a dolní hranice pro M bývá porušena na základě národních legislativ, které umožňují provádět operace, které jsou velmi krátké, např. v řádu několika týdnů (např. repo – operace, FX operace apod.).

Další omezující požadavek na vzorec (2.1) je předpoklad stejné výše plateb v ekvidistantních časových okamžicích (stejně časově vzdálených). Vzorec (2.1) tedy nelze použít například pro obligace, u kterých je poslední platba jiná než všechny předchozí (kromě kuponové platby dostane banka i nominální hodnotu). Nelze použít u produktů s pohyblivým výnosem, u kterých je velikost platby nebo splatné částky vázána na pohyb vnějšího indexu. Použití vzorce (2.1) je vyloučeno i v případě produktů, u kterých dochází k platbám pouze za splnění určitých podmínek. Banka tak dostává platby nepravidelně. Tyto platby mohou být navíc různě vysoké. V takových případech se za M dosazuje finální splatnost, ale v tom případě nezískáme průměrnou splatnost aktiva. Proto zobecníme vzorec (2.1).

Uvažujme n různých časových okamžiků:

$$\mathbf{t} = (t_1, \dots, t_n),$$

v nich uvažujme platby peněžního toku **CF**:

$$\mathbf{CF} = \left(\mathrm{CF}_{t_1}, \dots, \mathrm{CF}_{t_n}\right).$$

Pak

$$\mathrm{M} = \frac{\mathbf{t} \cdot \mathbf{CF}}{\mathbf{1} \cdot \mathbf{CF}} = \frac{\sum_{j=1}^{n} t_j \, \mathrm{CF}_{t_j}}{\sum_{j=1}^{n} \mathrm{CF}_{t_j}}, \tag{2.2}$$

kde

$$\mathbf{1} = (1, \dots, 1),$$

neboli jednotkový vektor délky n, a zápis

$$\mathbf{u} \cdot \mathbf{v}$$

znamená skalární součin vektorů **u** a **v**. Interpretace M podle (2.2) je opět střední doba splatnosti, rozdíl je v tom, že v tomto případě se výrazným způsobem projevují váhy, které jsou dány výší jednotlivých plateb.

V případě, že máme aktivum s finální splatností menší než 5 let, ale nelze aplikovat vzorec (2.1), můžeme použít vzorec (2.2). Ani jeden ze vzorců však nelze použít pro aktivum s finální splatností větší než 5 let. U těchto aktiv se výrazným způsobem projevuje úročení. Nicméně teorie finanční matematiky má instrument, který má stejnou interpretaci jako M ze vzorce (2.1) a započítává úročení. Tento instrument se nazývá durace nebo též střední doba splatnosti případně průměrná doba splatnosti. Buď i úroková míra. Označme D jako střední dobu splatnosti, která závisí na peněžním toku **CF** a úrokové míře i. Při stejném označení jako ve vzorci (2.2) máme:

$$\mathrm{D}(\mathbf{CF}, \mathrm{i}) = \frac{\sum_{j=1}^{n} t_j \; \mathrm{CF}_{t_j} \, v^{t_j}}{\sum_{j=1}^{n} \mathrm{CF}_{t_j} \, v^{t_j}}, \tag{2.3}$$

kde

$$\mathrm{v} = \frac{1}{1 + \mathrm{i}}$$

se nazývá diskontní faktor. Vzorec (2.3) je zobecněním vzorce (2.2) a lze použít v případě, kdy chceme do splatnosti započítat úrokovou míru. Použití vzorce (2.3) na stejně vysoké platby v ekvidistantních časových okamžicích je zřejmé.

Další zobecnění vzorce (2.1) plyne ze zobecnění vzorce (2.3). Zobecnění spočívá v předpokladu proměnlivé úrokové míry (je například vázaná na pohyb nějakého indexu). Buď tedy i_k úroková míra v čase k, $k \geq 0$. Při stejném značení jako výše označme

$$\mathbf{i} = \left(\mathrm{i}_{t_1}, \dots, \mathrm{i}_{t_n}\right)$$

vektor úrokových měr v daných časech. Pak:

$$v_{t_j} = \frac{1}{1 + i_{t_j}}$$

a

$$D(\mathbf{CF}, \mathbf{i}) = \frac{\sum_{j=1}^{n} t_j \; CF_{t_j} \left(v_{t_j}\right)^{t_j}}{\sum_{j=1}^{n} CF_{t_j} \left(v_{t_j}\right)^{t_j}}. \tag{2.4}$$

Použití vzorce (2.4) na stejně vysoké platby v ekvidistantních časových okamžicích je zřejmé. Následující tvrzení ukáže, že vzorec (2.4) je opravdu zobecněním vzorce (2.1), respektive že vzorec (2.1) je speciálním případem vzorce (2.4).

Tvrzení: Vzorce (2.1), (2.2) a (2.3) jsou speciálním případem vzorce (2.4). Navíc každý vzorec je speciálním případem vzorce s vyšším číslem.

Důkaz: Je zřejmé, že první tvrzení plyne z druhého. Při důkazu druhého tvrzení stačí ukázat, že (2.1) je speciální případ (2.2), (2.2) je speciální případ (2.3) a (2.3) je speciální případ (2.4). Nejprve položme:

$$i_{t_j} = i = \text{konst.}$$

Pak

$$v_{t_j} = \frac{1}{1 + i} = v = \text{konst.}$$

Nyní dosadíme do (2.4):

$$\frac{\sum_{j=1}^{n} t_j \; CF_{t_j} \left(v_{t_j}\right)^{t_j}}{\sum_{j=1}^{n} CF_{t_j} \left(v_{t_j}\right)^{t_j}} = \frac{\sum_{j=1}^{n} t_j \; CF_{t_j} \, v^{t_j}}{\sum_{j=1}^{n} CF_{t_j} \, v^{t_j}}.$$

Tedy (2.3) je speciální případ (2.4). Nyní položme:

$$t_j = j; \; CF_{t_j} = CF = \text{konst.}$$

Pak:

$$\frac{\sum_{j=1}^{n} t_j \, CF_{t_j}}{\sum_{j=1}^{n} CF_{t_j}} = \frac{\sum_{t=1}^{n} t \, CF}{\sum_{t=1}^{n} CF}.$$

Odsud (2.1) je speciálním případem (2.2). Nakonec položme:

$$i = 0 \Rightarrow v = \frac{1}{1 + 0} = 1 \Rightarrow v^k = 1, k \geq 0.$$

Pak:

$$\frac{\sum_{j=1}^{n} t_j \; CF_{t_j} \, v^{t_j}}{\sum_{j=1}^{n} CF_{t_j} \, v^{t_j}} = \frac{\sum_{j=1}^{n} t_j \, CF_{t_j}}{\sum_{j=1}^{n} CF_{t_j}}.$$

Odsud (2.2) je speciálním případem (2.3) a tím je tvrzení dokázáno.

Q. E. D.

Přímým důsledkem tvrzení je fakt, že (2.4) je zobecněním (2.1). Chce – li banka do splatnosti započítat úrokovou míru, lze použít vzorec (2.3) nebo (2.4), neboť interpretace u těchto dvou vzorců je stejná jako u vzorce (2.1), tedy se jedná o průměrnou splatnost. Jediným rozdílem jsou složitější váhy.

Numerická ilustrace

Ilustrace ukazuje vliv úroku na splatnost aktiva. Předpokládejme aktivum, které přináší roční příjmy velikosti 1000. Uvažujme jeho dobu trvání aktiva postupně 5, 10, 15, 20, 25 let a různé v čase konstantní úrokové míry. Tabulka 5 uvádí závislost splatnosti na délce trvání aktiva a úrokové míře.

úrok / trvání	5 let	10 let	15 let	20 let	25 let
0%	3	5,5	8	10,5	13
5%	2,9	5	7	8,9	10,5
10%	2,8	4,7	6,2	7,5	8,4
15%	2,7	4,3	5,5	6,3	6,8

Tabulka 5 – závislost splatnosti na délce trvání a fixní úrokové míře

První řádek je splatnost počítaná podle vzorce (2.1), ostatní řádky jsou počítané podle vzorce (2.3). Lze si povšimnout, že střední doba splatnosti při rostoucím úroku klesá a navíc velikost poklesu roste při rostoucí délce trvání aktiva.

Druhá ilustrace ukazuje totéž při nestejných úrokových měrách generovaných pomocí CIR procesu podle vzorce (dle [6]):

$$i_t = i_{t-1} + \alpha \, (\mu - i_{t-1}) \, \Delta + \sigma \sqrt{\Delta \, i_{t-1}} \, \varepsilon_t,$$

při volbě:

$$\alpha = 0{,}05; \; \mu = 0{,}03; \; \sigma = 0{,}025; \; \Delta = 1$$

a ε_t jsou nezávislé stejně rozdělené náhodné veličiny s normovaným normálním rozdělením (respektive jejich konkrétní realizace). Startovací úrovně jsou postupně 5%, 10%, 15%. Výsledky uvádí tabulka 6.

start / trvání	5 let	10 let	15 let	20 let	25 let
5%	2,9	5,2	7,4	9,6	11,8
10%	2,8	5	7,1	9,2	11,4
15%	2,7	4,8	6,8	8,7	10,8

Tabulka 6 – závislost splatnosti na délce trvání a proměnné úrokové míře

Vše je spočítáno podle vzorce (2.4). Stejně jako výše platí, že střední doba splatnosti klesá při rostoucím úroku i velikost skoků roste při rostoucí délce trvání aktiva.

3. Binární logistická regrese

V této části postupujeme podle [3].

3.1. Model binární logistické regrese a jeho odhad

V této kapitole popisujeme model binární logistické regrese (v této kapitole nadále jen logistická regrese), který slouží k modelování nejen parametru PD (pravděpodobnost selhání). Uvažujme p nezávislých proměnných. Označme:

$$\mathbf{x} = \left(1, x_1, x_2, \dots, x_p\right)^T$$

jako vektor nezávislých proměnných. Označme:

$$\pi(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}), \tag{3.1}$$

přičemž Y je náhodná veličina, která nabývá hodnot 0 nebo 1. Chceme modelovat

$$E[Y|\mathbf{x}] = \pi(\mathbf{x}),$$

neboli

$$Y = \pi(\mathbf{x}) + \varepsilon,$$

kde ε je náhodná veličina. Záleží na volbě $\pi(\mathbf{x})$. Lze použít přímo lineární regresi, ale to není nejvhodnější model, protože náhodná veličina Y nabývá pouze hodnot 0 a 1. Pokud by výsledek byl v intervalu $\langle 0; 1\rangle$, pak by tento výsledek bylo možné interpretovat jako pravděpodobnost daného jevu (například selhání, splacení úvěru), ale při výsledku mimo tento interval již není možná smysluplná interpretace. Označme:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

Funkce $g(\mathbf{x})$ se v obecné teorii zobecněných lineárních modelů (viz [8]) nazývá lineární prediktor. Dále označme:

$$\mathrm{odds}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}$$

jako šanci. Označme logitovou transformaci:

$$\mathrm{logit}(\mathbf{x}) = \log[\mathrm{odds}(\mathbf{x})] = \log\left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right],$$

kde log(x) značí přirozený logaritmus, a položme

$$\mathrm{logit}(\mathbf{x}) = g(\mathbf{x}).$$

Řešením vzhledem k $\pi(\mathbf{x})$ dostaneme:

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}, \tag{3.2}$$

což je vztah používaný v modelu logistické regrese. Logit je linková funkce používaná v teorii zobecněných lineárních modelů (viz [8]).

Nejdůležitější je odhad modelu. Uvažujme výběr n nezávislých pozorování

$$(\mathbf{x}_i, y_i), i = 1; 2; \dots; n, n \gg p,$$

přičemž hodnoty $y_1, y_2, \dots, y_n$ se nazývají odpovědi nebo odezvy a

$$\mathbf{x}_i = \left(1, x_{i1}, x_{i2}, \dots, x_{ip}\right)^T$$

Chceme odhadnout vektor koeficientů

$$\boldsymbol{\beta} = \left(\beta_0, \beta_1, \dots \beta_p\right)^T.$$

K tomuto odhadu použijeme metodu maximální věrohodnosti. Ze vzorce (3.1) zřejmě plyne

$$1 - \pi(\mathbf{x}) = P(Y = 0 \mid \mathbf{x}).$$

Tedy, pro ta pozorování, kde $y_i = 1$, je příspěvek k věrohodnostní funkci $\pi(\mathbf{x}_i)$ a pro pozorování, kde $y_i = 0$, je příspěvek k věrohodnostní funkci $1 - \pi(\mathbf{x}_i)$, kde $\pi(\mathbf{x}_i)$ je hodnota $\pi(\mathbf{x})$ spočítaná v bodě $\mathbf{x}_i$. Tedy příspěvek i – tého pozorování do věrohodnostní funkce je

$$\pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i}.$$

Z nezávislosti pozorování dostáváme věrohodnostní funkci

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i}.$$

Princip metody maximální věrohodnosti spočívá v nalezení maxima věrohodnostní funkce. K tomu je nutné věrohodnostní funkci derivovat podle všech prvků vektoru $\boldsymbol{\beta}$. Matematicky jednodušší je hledání maxima logaritmické věrohodnostní funkce:

$$L(\boldsymbol{\beta}) = \log[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \log[\pi(\mathbf{x}_i)] + (1 - y_i) \log[1 - \pi(\mathbf{x}_i)]\}.$$

Z monotonie logaritmu plyne, že nalezené maximum logaritmické věrohodnostní funkce je stejné jako hledané maximum věrohodnostní funkce. Pro jednoduchost zápisu uvažujeme:

$$g(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} = \boldsymbol{\beta}^T \mathbf{x}_i.$$

Před derivováním zjednodušíme logaritmickou věrohodnostní funkci. Zjevně platí:

$$1 - \pi(\mathbf{x}_i) = 1 - \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}} = \frac{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i} - e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}} = \frac{1}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}}.$$

Z vlastností logaritmu plyne:

$$\log[\pi(\mathbf{x}_i)] = \log\left[\frac{e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}}\right] = \log e^{\boldsymbol{\beta}^T \mathbf{x}_i} - \log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right) = \boldsymbol{\beta}^T \mathbf{x}_i - \log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right)$$

a

$$\log[1 - \pi(\mathbf{x}_i)] = \log\left[\frac{1}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}}\right] = \log 1 - \log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right) = -\log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right).$$

Odsud plyne:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \left[\boldsymbol{\beta}^T \mathbf{x}_i - \log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right)\right] + (1 - y_i)\left[-\log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right)\right] \right\} =$$

$$= \sum_{i=1}^{n} \left\{ y_i \boldsymbol{\beta}^T \mathbf{x}_i - y_i \log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right) - \log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right) + y_i \log\left(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}\right) \right\} =$$

$$= \sum_{i=1}^{n} \left\{ y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) - \log\left(1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}\right) \right\}.$$

Derivováním posledního výrazu podle všech prvků vektoru $\boldsymbol{\beta}$ dostaneme:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}} \right] = \sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)],$$

a

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}} x_{ij} \right] = \sum_{i=1}^{n} x_{ij} [y_i - \pi(\mathbf{x}_i)], \; j \neq 0.$$

Položíme derivace rovny nule a dostaneme věrohodnostní rovnice:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0$$

a

$$\sum_{i=1}^{n} x_{ij}\,[y_i - \pi(\mathbf{x}_i)] = 0, j = 1;\ 2; \dots;\ p.$$

Řešením této soustavy dostaneme hledaný odhad. Řešení této soustavy dostáváme pomocí statistických softwarů. Označme $\widehat{\boldsymbol{\beta}}$ jako řešení věrohodnostních rovnic. Potom vyrovnané hodnoty, nebo též odpovědi, popřípadě předpovědi, pro model logistické regrese jsou hodnoty spočítané dle vzorce (3.2) při použití $\widehat{\boldsymbol{\beta}}$ a $\mathbf{x}_i$, označme tento odhad jako $\widehat{\pi(\mathbf{x}_i)}$.

Další částí odhadu je odhad standardní chyby odhadu, která se posléze používá ke stanovení intervalu spolehlivosti, označme ji jako SE. Odhad SE vychází z teorie maximálně věrohodných odhadů. Tato teorie je založena na výpočtu druhých parciálních derivací logaritmické věrohodnostní funkce:

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} = \sum_{i=1}^{n} x_{ij}\left[-\frac{\left(1+e^{\boldsymbol{\beta}^T\mathbf{x}_i}\right)e^{\boldsymbol{\beta}^T\mathbf{x}_i}\,x_{ij} - e^{\boldsymbol{\beta}^T\mathbf{x}_i}\,e^{\boldsymbol{\beta}^T\mathbf{x}_i}\,x_{ij}}{\left(1+e^{\boldsymbol{\beta}^T\mathbf{x}_i}\right)^2}\right] = \sum_{i=1}^{n} x_{ij}^2\,\frac{(-1)e^{\boldsymbol{\beta}^T\mathbf{x}_i}}{\left(1+e^{\boldsymbol{\beta}^T\mathbf{x}_i}\right)^2} =$$

$$= -\sum_{i=1}^{n} x_{ij}^2\,\pi(\mathbf{x}_i)\frac{1}{1+e^{\boldsymbol{\beta}^T\mathbf{x}_i}} = -\sum_{i=1}^{n} x_{ij}^2\,\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)], j = 1;\ 2; \dots;\ p$$

a

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j\,\partial \beta_l} = \sum_{i=1}^{n} x_{ij}\left[-\frac{\left(1+e^{\boldsymbol{\beta}^T\mathbf{x}_i}\right)e^{\boldsymbol{\beta}^T\mathbf{x}_i}\,x_{il} - e^{\boldsymbol{\beta}^T\mathbf{x}_i}\,e^{\boldsymbol{\beta}^T\mathbf{x}_i}\,x_{il}}{\left(1+e^{\boldsymbol{\beta}^T\mathbf{x}_i}\right)^2}\right] =$$

$$= \sum_{i=1}^{n} x_{ij}x_{il}\,\frac{(-1)e^{\boldsymbol{\beta}^T\mathbf{x}_i}}{\left(1+e^{\boldsymbol{\beta}^T\mathbf{x}_i}\right)^2} == -\sum_{i=1}^{n} x_{ij}x_{il}\,\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)],\ j,l = 1;\ 2; \dots;\ p, j \neq l.$$

přičemž

$$x_{i0} = 1, \forall i.$$

Označme $\mathbf{I}(\boldsymbol{\beta})$ matici o rozměrech (p+1) × (p+1), která obsahuje záporně vzaté druhé parciální derivace logaritmické věrohodnostní funkce. Tato matice se nazývá sledovací informační matice. Z matice $\mathbf{I}(\boldsymbol{\beta})$ dostaneme rozptylovou matici $\boldsymbol{\beta}$:

$$\mathrm{Var}(\boldsymbol{\beta}) = \mathbf{I}(\boldsymbol{\beta})^{-1}.$$

V odhadu rozptylové matice použijeme pouze diagonální prvky, tedy odhady rozptylu jednotlivých prvků $\boldsymbol{\beta}$, a získáme odhad standardní chyby odhadu:

$$\widehat{\mathrm{SE}}(\hat{\beta}_j) = \left[\widehat{\mathrm{var}}(\hat{\beta}_j)\right]^{1/2}, \forall j. \tag{3.3}$$

Pro odhad rozptylové matice se použije odhad $\mathbf{I}(\boldsymbol{\beta})$:

$$\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}}) = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X},$$

kde

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ & \vdots & & \ddots & \vdots \\ 1 & x_{n1} & x_{x2} & \cdots & x_{np} \end{bmatrix}$$

a

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1\,(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2\,(1-\hat{\pi}_2) & & \\ \vdots & & \ddots & \vdots \\ & & \hat{\pi}_{n-1}\,(1-\hat{\pi}_{n-1}) & 0 \\ 0 & \cdots & 0 & \hat{\pi}_n\,(1-\hat{\pi}_n) \end{bmatrix},$$

kde

$$\hat{\pi}_i = \widehat{\pi(\mathbf{x}_i)}.$$

Součástí odhadu je intervalový odhad (neboli interval spolehlivosti) pro dané $\mathbf{x}$. Máme – li dané $\mathbf{x}$, pak intervalový odhad pro danou pravděpodobnost $\pi(\mathbf{x})$ podá lepší informaci než jedno číslo, které nemusí být zcela přesné. Intervalový odhad probíhá v několika krocích. Nejprve se odhadne standardní chyba odhadu β_j podle (3.3). Odsud dostáváme intervalový odhad pro koeficienty

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\,\widehat{\mathrm{SE}}(\hat{\beta}_j), \forall j,$$

kde z_α je α – kvantil normovaného normálního rozdělení. Dále je nutné sestavit intervalový odhad pro funkci $g(\mathbf{x})$. Ten sestavíme dle:

$$\hat{g}(\mathbf{x}) \pm z_{1-\frac{\alpha}{2}}\,\widehat{\mathrm{SE}}[\hat{g}(\mathbf{x})], \tag{3.4}$$

přičemž $\widehat{\mathrm{SE}}[\hat{g}(\mathbf{x})]$ je odmocnina z rozptylu:

$$\widehat{\mathrm{var}}[\hat{g}(\mathbf{x})] = \sum_{j=0}^{p} x_j^2\,\widehat{\mathrm{var}}(\hat{\beta}_j) + \sum_{j=0}^{p}\sum_{k=j+1}^{p} 2\,x_j\,x_k\,\widehat{\mathrm{cov}}(\hat{\beta}_j, \hat{\beta}_k),$$

kde je třeba dodefinovat

$$\sum_{k=p+1}^{p} \ldots = 0.$$

Intervalový odhad $\pi(\mathbf{x})$ pak dostaneme dosazením (3.4) do (3.2):

$$\frac{e^{\hat{g}(\mathbf{x})\pm z_{1-\frac{\alpha}{2}}\widehat{SE}[\hat{g}(\mathbf{x})]}}{1+e^{\hat{g}(\mathbf{x})\pm z_{1-\frac{\alpha}{2}}\widehat{SE}[\hat{g}(\mathbf{x})]}}.$$

Důležitá je též interpretace odhadnutých parametrů. Základní interpretace je změna odhadnutých předpovědí při změně dané proměnné. Jiná možná interpretace vychází z modelu pouze s konstantním členem a jednou proměnnou. Spočítáme $\pi(x)$ ve dvou bodech $x_1 \neq x_2$. Dále počítáme v obou bodech šance:

$$\mathrm{odds}(x_1) = \frac{\pi(x_1)}{1-\pi(x_1)} = e^{\beta_0+\beta_1 x_1}$$

a

$$\mathrm{odds}(x_2) = \frac{\pi(x_2)}{1-\pi(x_2)} = e^{\beta_0+\beta_1 x_2}.$$

Spočítáme poměr obou šancí:

$$\frac{\mathrm{odds}(x_1)}{\mathrm{odds}(x_2)} = \frac{e^{\beta_0+\beta_1 x_1}}{e^{\beta_0+\beta_1 x_2}} = e^{\beta_1(x_1-x_2)}.$$

Zlogaritmujeme pravou i levou stranu a dostaneme

$$\log\left(\frac{\mathrm{odds}(x_1)}{\mathrm{odds}(x_2)}\right) = \beta_1(x_1-x_2).$$

Odsud lze výsledný odhad též interpretovat.

3.2. Testy významnosti modelu

V této kapitole krátce pojednáme o testech významnosti dané proměnné. O testování významnosti samotného modelu pojednáme v kapitole 3.4. Zde chceme zjistit, zda je i – tá proměnná významná či nikoli, tedy testujeme nulovou hypotézu $\beta_i = 0$ proti alternativě, která tvrdí, že tomu tak není. Při testování významnosti dané proměnné nejprve sestavíme saturovaný model. Tento model sestavíme použitím takového počtu proměnných jako je počet pozorování. Pak spočítáme věrohodnostní funkce našeho modelu a saturovaného modelu a dosadíme do následujícího vzorce:

$$D = -2\ \log\left[\frac{(\text{věrohodnost odhadnutého modelu})}{(\text{věrohodnost saturovaného modelu})}\right].$$

Proměnnou D spočítáme se zkoumanou proměnnou a bez ní a spočítáme:

$$G = D(\text{model bez proměnné}) - D(\text{model s proměnnou}).$$

V případě logistické regrese platí:

$$G = -2\ \log\left[\frac{(\text{věrohodnost modelu bez proměnné})}{(\text{věrohodnost modelu s proměnnou})}\right] =$$
$$= -2\ [\log(\text{věrohodnost modelu bez proměnné})$$
$$- \log(\text{věrohodnost modelu s proměnnou})] =$$
$$= -2\ [(\text{logaritmická věrohodnost modelu bez proměnné})$$
$$- (\text{logaritmická věrohodnost modelu s proměnnou})].$$

O veličině G lze ukázat, že má chí – kvadrát rozdělení o jednom stupni volnosti. Odsud spočítáme p – hodnotu testu

$$p - \text{hodnota} = P[Y > G], \text{kde } Y \sim \chi^2(1).$$

Tento test se nazývá test poměrem věrohodností. V případě mnohorozměrného testování, kde testujeme, zda jsou všechny koeficienty nulové, probíhá test stejně s výjimkou počtu stupňů volnosti, který je v tomto případě roven p.

Další možností je provedení Waldova testu. V případě Waldova testu spočítáme pro j – tou proměnnou Waldovu statistiku

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}.$$

Pak spočítáme p – hodnotu pomocí

$$p - \text{hodnota} = P\left[|Y| > W_j\right], \text{kde } Y \sim N(0; 1).$$

V případě vícerozměrného testování spočítáme statistiku

$$W = \hat{\boldsymbol{\beta}}^T(\mathbf{X}^T\mathbf{V}\mathbf{X})\hat{\boldsymbol{\beta}},$$

která má v tomto případě chí – kvadrát rozdělení o p+1 stupních volnosti. Ve vícerozměrném případě opět testujeme, zda je všech p+1 koeficientů nulových.

Déle je možné provést skórový test. V rámci tohoto testu spočítáme skórovou statistiku pro j – tou proměnnou

$$ST = \frac{\sum_{i=1}^{n} x_{ij}(y_i - \bar{y})}{\sqrt{\bar{y}\ (1 - \bar{y})\ \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}.$$

P – hodnota pak je

$$p - \text{hodnota} = P[|Y| > ST], \text{kde } Y \sim N(0; 1).$$

Vícerozměrný skórový test závisí na derivacích $L(\boldsymbol{\beta})$ podle $\boldsymbol{\beta}$. Výpočet tohoto testu pak probíhá stejně jako v případě vícerozměrného Waldova testu.

3.3. Algoritmus Stepwise pro výběr proměnných

Častou situací v praxi je velký počet proměnných. V mnoha případech se stává:

$$\hat{\beta}_j < 10^{-10}, \text{pro nějaké } j \in \{1, \dots, p\}.$$

To však znamená, že tento koeficient a příslušná proměnná mohou být z modelu vyřazeny. Tato situace však otvírá otázku, zda by nebylo lepší takové proměnné předem z modelu vyřadit a při odhadování modelu je vůbec neuvažovat. Je tedy vhodné mít postup pro výběr proměnných, které budou zařazeny do modelu. K tomu existuje mnoho algoritmů, návodů a postupů. Pojednáme pouze o rekurzivním algoritmu stepwise, ve kterém se střídá dopředný výběr a zpětná eliminace proměnných. Tento algoritmus je často používán pro lineární regresi a většina statistických programů má tento algoritmus pro lineární regresi zabudován. Mnoho programů má tento algoritmus zabudovaný i pro logistickou regresi.

Algoritmus Stepwise lze shrnout do několika kroků:

<u>KROK (0)</u>: V tomto kroku se vybere první proměnná. Uvažujme p nezávislých proměnných. Pro začátek předpokládejme, že všechny mají svůj vliv. Krok (0) začíná odhadem modelu, který obsahuje pouze konstantní člen β_0, a jeho logaritmické věrohodnosti spočítané dosazením odhadu do logaritmické věrohodnostní funkce. Označme tuto věrohodnost jako L_0. Dále se pro každou proměnnou zvlášť odhadne model a jeho logaritmická věrohodnost. Pro proměnnou x_j označme danou logaritmickou věrohodnost jako $L_j^{(0)}$, kde j znamená, že daná proměnná byla zařazena do modelu a (0) značí, o jaký krok se jedná. Tato notace bude v popisu algoritmu používána i nadále.

Po odhadnutí modelů a jejich logaritmických věrohodností se provádí modifikace testu poměrem věrohodností. Pro potřeby testu uvažujme náhodnou veličinu A, pro kterou platí:

$$A \sim \chi^2(\nu).$$

O hodnotách pro ν pojednáme dále. Označme

$$G_j^{(0)} = -2\left(L_0 - L_j^{(0)}\right).$$

Tuto hodnotu tedy spočítáme pro každou proměnnou a pro každou proměnnou spočítáme její p – hodnotu:

$$p_j^{(0)} = P\left(A > G_j^{(0)}\right), \tag{3.5}$$

přičemž $\nu = k - 1$, pokud x_j je kategoriální proměnná s k kategoriemi, a $\nu = 1$, pokud je daná proměnná spojitá. Na základě tohoto testu vybereme proměnnou, kterou fi-

nálně zařadíme do modelu. Označíme – li tuto proměnnou jako x_{e_1}, pak pro její p – hodnotu musí platit

$$p_{e_1}^{(0)} = \min_j \left(p_j^{(0)}\right).$$

Zvolme

$$p_E \in (0; 1).$$

O doporučeních pro tuto volbu pojednáme později. Pak námi vybranou proměnnou zařadíme do modelu, pokud

$$p_{e_1}^{(0)} < p_E.$$

Algoritmus tak pokračuje k dalšímu kroku. V opačném případě žádnou proměnnou nezařadíme a algoritmus končí. Zde je nutné upozornit, že p – hodnota v tomto algoritmu není p – hodnota známá z teorie testování hypotéz.

<u>KROK (1)</u>: V tomto kroku odhadneme model pro proměnnou x_{e_1} a spočítáme jeho logaritmickou věrohodnost $L_{e_1}^{(1)}$. Dále zkusíme přidat každou ze zbývajících proměnných do modelu. Tedy model bude obsahovat proměnné x_{e_1} a $x_j, j \neq e_1$. Pro všechny modely spočítáme logaritmickou věrohodnost $L_{e_1 j}$. Provedeme modifikaci testu poměrem věrohodností, ve kterém

$$G_j^{(1)} = -2\left(L_{e_1}^{(1)} - L_{e_1 j}\right)$$

a $p_j^{(1)}$ je příslušná p – hodnota spočítaná dle (3.5) při současném označení. Označme x_{e_2} jako proměnnou, pro kterou platí

$$p_{e_2}^{(1)} = \min_j \left(p_j^{(1)}\right).$$

Pokud

$$p_{e_2}^{(1)} < p_E,$$

pak proměnnou x_{e_2} přidáme do modelu a pokračujeme k dalšímu kroku. Jinak algoritmus končí.

<u>KROK (2)</u>: Krok (2) začíná odhadem modelu pro proměnné x_{e_1} a x_{e_2}. Mohlo se stát, že přidáním proměnné x_{e_2} došlo ke ztrátě významnosti proměnné x_{e_1}. Proto provedeme zpětnou eliminaci proměnných. Pro náš model spočítáme logaritmickou věrohodnost $L_{e_1 e_2}^{(2)}$. Dále zkusíme každou proměnnou x_{e_j} vyřadit z modelu, odhadnout model bez ní a spočítat jeho logaritmickou věrohodnost $L_{-e_j}^{(2)}$. Opět provedeme modifikaci testu poměrem věrohodností, ve kterém

$$G_{-e_j}^{(2)} = -2\left(L_{-e_j}^{(2)} - L_{e_1 e_2}^{(2)}\right)$$

a $p_{-e_j}^{(2)}$ je příslušná p – hodnota, vše pro j = 1; 2. Označme x_{r_2} proměnnou, pro kterou je

$$p_{r_2}^{(2)} = \max\left(p_{-e_1}^{(2)}, p_{-e_2}^{(2)}\right).$$

Nyní zvolíme

$$p_R \in (0; 1).$$

O doporučeních pro tuto volbu pojednáme později. Pak proměnnou x_{r_2} vyřadíme z modelu, pokud

$$p_{r_2}^{(2)} > p_R,$$

v opačném případě proměnnou ponecháme v modelu. V obou případech pokračujeme dále.

Následuje dopředný výběr proměnných. Tato fáze probíhá stejně jako KROK (1), tedy přidáním každé proměnné zvlášť, odhadem modelu a výpočtem jeho logaritmické věrohodnosti, provedením modifikace testu poměrem věrohodností a vybráním jedné proměnné. Dále použijeme totéž kritérium pro rozhodnutí, zda danou proměnnou přidáme či nikoli. Pokud danou proměnnou přidáme, pak algoritmus pokračuje k dalšímu kroku, jinak končí.

<u>KROK (3)</u>: Postup v tomto kroku je stejný jako v předchozím kroku. Tedy nejprve odhadneme model, pak provedeme zpětnou eliminaci a dopředný výběr proměnných. Algoritmus takto probíhá, dokud nebudou splněny podmínky posledního kroku.

<u>KROK (S)</u>: Tento krok nastává ve dvou případech:

a) všechny proměnné jsme zařadili do modelu
b) všechny proměnné v modelu mají p – hodnotu pro vyřazení menší než p_R a proměnné, které nejsou zařazené do modelu, mají p – hodnotu větší než p_E.

Model v tom případě obsahuje proměnné, které jsou důležité s ohledem na kritéria p_R a p_E. Tyto proměnné mohou (nemusí) být zahrnuty v cílovém modelu. Při použití jakýchkoli hodnot p_R a p_E je vhodné zkontrolovat výběr proměnných na základě statistického testu, protože p_E není p – hodnota používaná při testování hypotéz. Existují dvě metody, jak daný test provést. Uvažujme $Q + 1$ vybraných proměnných včetně konstantního členu. Vezměme množinu všech vybraných proměnných, ve které budou všechny proměnné seřazené tak, jak jsme je přidávali do modelu. Označme $L_q, q = 1, \dots, Q$, jako logaritmickou věrohodnost modelu s q proměnnými. Dále zvolme hladinu α, nejčastěji se volí $\alpha = 0{,}05$.

První metoda porovnává dva „sousední“ modely. Předpokládejme, že jsme do modelu přidali q proměnných včetně konstantního členu. Položme:

$$G_q = 2\left(L_q - L_{q-1}\right).$$

Pokud

$$P\left(A > G_q\right) < \alpha,$$

pak $q = q + 1$ a pokračujeme k další proměnné, jinak končíme a v modelu použijeme pouze q proměnných, přičemž:

$$A \sim \chi^2(1).$$

Druhá metoda je pouze alternativou první metody. Položme

$$q = 0$$

a

$$G_q = 2\left(L_{Q+1} - L_q\right).$$

Pokud

$$P\left(B > G_q\right) < \alpha,$$

kde

$$B \sim \chi^2(Q + 1 - q),$$

pak $q = q + 1$ a pokračujeme k další proměnné, jinak končíme se stejným závěrem jako v případě první metody.

Tento algoritmus přináší dva problémy. Prvním problémem je velká početní náročnost. V případě velkého počtu proměnných a velkého množství dat může tento výběr trvat několik minut i hodin, v extrémním případě dnů. Ještě déle pak výběr trvá v případě, kdy tento algoritmus není v daném softwaru zabudovaný, protože v tom případě trvá výběr proměnných cca týden. Druhým problémem je volba p_R a p_E. Tato volba je zcela subjektivní. Velmi striktní volba je

$$p_E = 0{,}05.$$

Tato volba však může z modelu z modelu vyřadit proměnné, které jsou důležité. Nejčastěji se doporučuje zvolit

$$p_E \in \langle 0{,}15; 0{,}2 \rangle.$$

Chceme – li vybrat velké množství proměnných, pak můžeme zvolit i

$$p_E \geq 0{,}25.$$

Tato volba se však nepoužívá často. Druhou otázkou je volba p_R. Chceme – li vybrat velké množství proměnných, respektive vyloučit jich co nejméně, pak volíme

$$p_R = 0{,}9.$$

Mnohem lepší je však propojit volbu p_R s volbou p_E. Doporučuje se volit:

$$p_R \in \langle p_E + 0{,}05; p_E + 0{,}1 \rangle.$$

Tedy například pokud zvolíme $p_E = 0{,}15$, pak je vhodné zvolit $p_R = 0{,}2$ a podobně.

3.4. Verifikace modelu

Testy dobré shody

Testy dobré shody pro model logistické regrese sestavili Hosmer a Lemeshow v roce 1980 (viz též [4]). Model testuje nulovou hypotézu, která tvrdí, že model popisuje data, proti opačné alternativě. Při tomto testuje nejprve nutné vzít předpovědi modelu $\widehat{\pi(\mathbf{x}_i)}$ a seřadit je podle velikosti od nejmenší po největší. Takto seřazené předpovědi se pak rovnoměrně rozdělí do K skupin, tedy v první skupině bude nejmenších $\frac{1}{K}\%$ $\widehat{\pi(\mathbf{x}_i)}$, ve druhé skupině bude dalších $\frac{1}{K}\%$ $\widehat{\pi(\mathbf{x}_i)}$, takto pokračujeme v rozdělování dále a v poslední skupině bude největších $\frac{1}{K}\%$ $\widehat{\pi(\mathbf{x}_i)}$. Je zřejmé, že vzhledem k počtu pozorování bude každé skupině jiný počet odpovědí (bude se lišit o jedničku).

Označme C_j jako skupinu indexů i v j – té skupině. Buď o_j počet jedničkových odpovědí v C_j, tedy

$$o_j = \sum_{i \in C_j} y_i.$$

Dále označme n_j' jako počet indexů ve skupině C_j a $\overline{\pi}_j$ jako průměr odhadnutých pravděpodobností ve skupině C_j. Potom Hosmer – Lemeshowova statistika je

$$C = \sum_{i=1}^{K} \frac{\left(o_j - n_j' \overline{\pi}_j\right)^2}{n_j' \, \overline{\pi}_j \left(1 - \overline{\pi}_j\right)}. \tag{3.6}$$

Čím je C menší, tím je daný model lepší a významnější. Lze ukázat, že

$$C \sim \chi^2(K - 2).$$

Hodnota C se pak porovnává s (1 – α) – kvantilem daného rozdělení, přičemž obvykle bývá:

$$\alpha = 0{,}05.$$

Pokud je C menší než příslušný kvantil, pak odhad modelu považujeme za dobrý, respektive nezamítáme nulovou hypotézu, jinak jej nepovažujeme za dobrý. P – hodnota testu je

$$p = 1 - F(C),$$

kde F je distribuční funkce příslušného chí – kvadrát rozdělení. Tedy, pokud

$$p > \alpha,$$

pak říkáme, že odhad modelu je dobrý, a pokud

$$p < \alpha,$$

pak odhad modelu prohlásíme za špatný.

Parametr K vždy volíme, nejčastější volba je

$$K = 10,$$

ale při velkém množství dat jsou přípustné i jiné volby, například

$$K = 100.$$

Pokud model prohlásíme za špatný při rozdělení na deset skupin, můžeme provést dělení na více skupin. Pokud při rozdělení na více skupin zjistíme, že model je dobrý, pak jej můžeme považovat za dobrý. Volba

$$K > 100$$

se nepoužívá.

McFaddenův koeficient determinace

McFaddenův koeficient determinace je definován jako:

$$R^2 = 1 - \frac{L_p}{L_0},$$

kde L_p je logaritmická věrohodnost testovaného modelu a L_0 je logaritmická věrohodnost modelu pouze s konstantním členem. Čím je R^2 blíže jedné, tím je model lepší.

R^2 v praxi používáme k porovnávání dvou modelů. Uvažujme modely M1 a M2 s logaritmickými věrohodnostmi L_{p_1} a L_{p_2} a McFaddenovými koeficienty determinace R_1^2 a R_2^2. Zjevně platí, že lepší model má vyšší logaritmickou věrohodnost. Pokud je model M1 lepší, pak

$$L_{p_1} > L_{p_2}.$$

a tedy

$$R_1^2 > R_2^2.$$

4. Diverzifikační schopnosti modelu

V této části využíváme informací z [3].

Diverzifikační schopností modelu rozumíme schopnost modelu rozpoznat dobré a špatné klienty. Mějme tedy pozorování $y_i, i = 1; \dots; n$, a jim příslušné skóre, respektive odpovědi daného modelu, $s_i, i = 1; \dots; n$, v modelu logistické regrese se například může jednat o $\widehat{\pi(\mathbf{x}_i)}$. Uvažujme dvojice

$$\{y_i, s_i\}$$

a seřaďme je vzestupně podle s_i, neboť zjevně platí, že čím větší s_i, tím spíše $y_i = 1$, přičemž $y_i = 1$ znamená, že daný klient nesplatil. Označme

$$\{y_{(i)}, s_{(i)}\}$$

jako seřazené dvojice. Pro ideální model by zjevně existovalo $k \leq n$ takové, že

$$y_{(i)} = 0, \text{pro } i \leq k,$$

$$y_{(i)} = 1, \text{pro } i > k.$$

Takový model je ideální, ale v praxi nereálný. Diverzifikační vlastnosti modelu hodnotí, jak je daný model blízký ideálnímu modelu.

Pro potřeby výpočtů označme

$$B = \sum_{i=1}^{n} y_i$$

jako počet klientů, kteří nesplatili a

$$G = n - B$$

jako počet klientů, kteří splatili. Dále označme distribuční funkce B a G jako

$$F_B(s) = \frac{1}{B} \sum_{i=1}^{n} \mathbf{1}_{[s_i \leq s, y_i = 1]},$$

$$F_G(s) = \frac{1}{G} \sum_{i=1}^{n} \mathbf{1}_{[s_i \leq s, y_i = 0]},$$

kde

$$\mathbf{1}_A$$

je indikátor jevu A.

Klasifikační tabulky, ROC křivka

Intuitivní způsob, jak shromáždit výsledky odhadnutého modelu, je pomocí klasifikačních tabulek. Tato tabulka vznikne srovnáním výstupní proměnné y a odhadnuté hodnoty, která vychází z odhadu pravděpodobností pomocí odhadnutého modelu logistické regrese. K sestavení této odhadnuté hodnoty je nutné zvolit uzlový bod c a porovnat odhadnuté pravděpodobnosti s c . Pokud je odhadnutá pravděpodobnost větší než c, pak je odhadnutá hodnota rovna 1, jinak je nulová. Příklad klasifikační tabulky znázorňuje tabulka 7.

Odhadnutá hodnota	Pozorování		
	1	0	Celkem
1	A_{11}	A_{01}	O_1
0	A_{10}	A_{00}	O_0
Celkem	P_1	P_0	S

Tabulka 7 – příklad klasifikační tabulky

Z této tabulky spočítáme citlivost a specifičnost podle

$$\text{citlivost} = \frac{A_{11}}{P_1},$$

$$\text{specifičnost} = \frac{A_{00}}{P_0}.$$

Na základě citlivosti, specifičnosti a celé klasifikační tabulky již můžeme subjektivně posoudit kvalitu odhadu. Problém je ve volbě uzlového bodu c . Nejčastěji se doporučuje volba

$$c = 0{,}5,$$

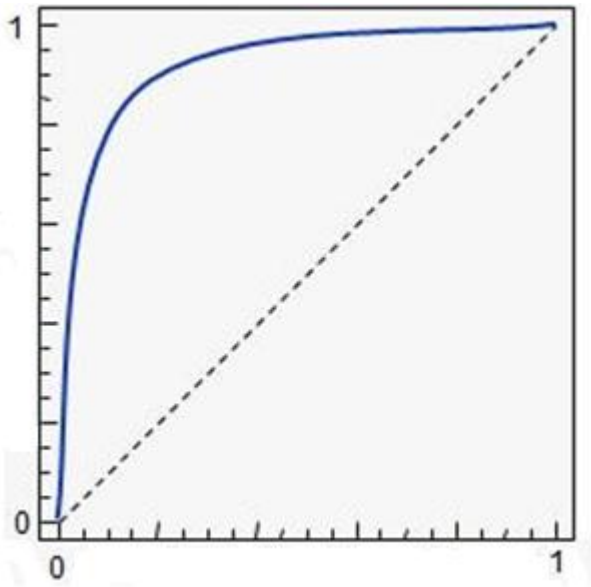
ale je možné volit c i jinak.

Další možnost, jak posoudit kvalitu modelu, je pomocí plochy pod ROC křivkou. K sestavení ROC křivky se používají klasifikační tabulky. Při sestavování klasifikační tabulky se volí různé uzlové body c, na jejichž základě se sestaví klasifikační tabulka a spočítá se citlivost a specifičnost. Též se spočítá položka

$$w = 1 - \text{specifičnost}.$$

Odsud se sestrojí ROC křivka jako graf citlivosti v závislosti na w, příklad ROC křivky ve srovnání s přímkou naivního modelu je na obr. 1, přičemž na vodorovné ose jsou hodnoty specifičnosti a na svislé ose jsou hodnoty citlivosti. Po sestrojení ROC křivky se spočítá plocha pod jejím grafem. Pokud je tato plocha menší nebo rovna 0,5, pak odhadnutý model odpovídá naivnímu modelu. Pokud je tato plocha

alespoň 0,7, pak již lze považovat odhadnutý model za kvalitní a takový model lze používat. Pokud je tato plocha alespoň 0,8, pak máme téměř perfektní model. Pokud je tato plocha alespoň 0,9, pak je náš model mimořádně vynikající. Poslední možnost je však v praxi nereálná.



Obr. 1 – příklad ROC křivky

Při sestavování klasifikačních tabulek pro ROC křivku je nutné zvolit několik uzlových bodů c . Počet a rozmezí pro volbu uzlových bodů je velmi subjektivní záležitost. Nejlepší možnost je začít s volbou na hodnotě 0,05 a pak hodnotu postupně zvyšovat, dokud specifičnost není rovna jedné. Velikost skoků je subjektivní volba, občas se volí 0,05.

Spočítat plochu pod ROC křivkou nemusí být vždy jednoduché. Proto je vhodné podívat se v jiném směru na pojem plocha pod ROC křivkou. Označme n_1 jako počet pozorování, pro která je $y = 1$, a n_0 jako počet pozorování, pro která je $y = 0$. Pak vytvořme $n_1 \times n_0$ dvojic tak, že ke každému pozorování s $y = 1$ přidáme všechna pozorování s $y = 0$. Pro těchto $n_1 \times n_0$ dvojic spočítáme relativní četnost případů, kdy pozorování s $y = 1$ má větší odhadnutou pravděpodobnost než pozorování s $y = 0$. Tuto relativní četnost můžeme považovat za plochu pod ROC křivkou. Interpretace ROC křivky, citlivosti a specifičnosti je uvedena na konkrétním numerickém příkladu v kapitole 5.3.

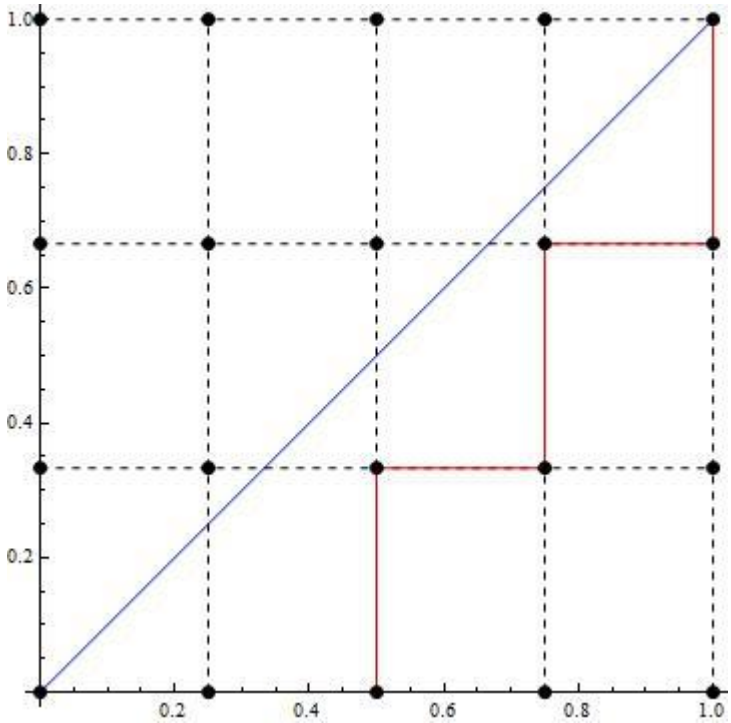
Zrcadlová ROC křivka spočívá ve využití $\{y_{(i)}, s_{(i)}\}$. V $y_{(i)}$ je v tom případě nutné spočítat nuly a jedničky. V grafu se pak vyznačí interval $(0; 1)$ na svislé i vodorovné ose. Nechť n_1 je počet pozorování, pro která $y = 1$, a n_0 je počet pozorování, pro která $y = 0$. Pak interval $(0; 1)$ na vodorovné ose rozdělíme na n_0 stejných dílů a na svislé ose jej rozdělíme na n_1 stejných dílů. Pak již zrcadlovou ROC křivku zakreslujeme podle:

- pokud je $y_{(i)} = 0$, pak v následujícím dílku postupujeme vodorovně,
- pokud je $y_{(i)} = 1$, pak v následujícím dílku postupujeme svisle,
- začátek je v bodě $(0; 0)$.

Takto sestavená zrcadlová ROC křivka je zrcadlově obrácená ROC křivka podle osy prvního kvadrantu. Plocha pod takto sestavenou zrcadlovou ROC křivkou je malá. Lze ji spočítat jako doplněk plochy pod ROC křivkou do jedné. Sestavování zrcadlové ROC křivky ilustruje následující příklad.

Příklad

Předpokládejme, že máme sedm klientů, z nichž tři selhali, tj. nesplatili úvěr, a ostatní úvěr splatili. To znamená, že rozdělíme jednotkový čtverec na dvanáct částí tak, jak je to znázorněno čárkovaně na obr. 2.



Obr. 2 – sestavování zrcadlové ROC křivky

Předpokládejme, že máme model, který nám spočítá skóre pro všechny klienty tak, že po seřazení od nejmenšího po největší budeme mít seřazena pozorování takto:

$$(0; 0; 1; 0; 1; 0; 1).$$

Při sestavování grafu zrcadlové ROC křivky začneme v bodě $(0; 0)$, pak dvakrát provedeme posun doprava, pak jednou nahoru, jednou doprava, jednou nahoru, jednou doprava a nakonec jednou nahoru.

Plocha pod zrcadlovou ROC křivkou se pak spočítá podle obdélníků pod jejím grafem. Jednotkový čtverec jsme rozdělili na dvanáct stejně velkých částí, to znamená, že každá část má obsah $\frac{1}{12}$. Pod zrcadlovou ROC křivkou jsou tři obdélníky a tedy plocha pod zrcadlovou ROC křivkou je $\frac{3}{12}$. Takový postup pro počítání plochy pod zrcadlovou ROC křivkou lze též uplatnit, ale při velkém množství dat již není tak snadný.

□

Giniho koeficient

Giniho koeficient (ozn. GINI) je definován následovně:

$$\mathrm{GINI} = 1 - 2 \cdot \mathrm{AUC},$$

kde AUC je plocha pod zrcadlovou ROC křivkou. Z

$$\mathrm{AUC} \in \langle 0; 1 \rangle$$

plyne

$$\mathrm{GINI} \in \langle -1; 1 \rangle.$$

Platí, že

$$\mathrm{GINI} = 1$$

značí dokonale správnou diverzifikaci,

$$\mathrm{GINI} = -1$$

značí dokonale chybnou diverzifikaci a

$$\mathrm{GINI} = 0$$

značí neutrální diverzifikaci. Čím je GINI blíže 1, tím je model lepší. Nicméně neexistuje hranice pro určení, zda je GINI malý nebo velký. Typické hodnoty Giniho koeficientu pro reálné v praxi použitelné modely se pohybují v rozmezí

$$0{,}5 < \mathrm{GINI} < 1.$$

Na počítání Giniho koeficientu existuje jednoduchý návod. Buď R_i pořadí $s_{(i)}$. Označme

$$S = \sum_{i=1}^{n} R_i y_{(i)}.$$

Dále označme

$$S_{min} = \frac{B}{2}(B + 1)$$

a

$$S_{max} = \frac{(G + 1) + n}{2} B$$

jako minimální a maximální hodnoty, kterých může nabývat S. Pak platí:

$$GINI = 2\frac{S - S_{min}}{S_{max} - S_{min}} - 1.$$

Kolmogorovova – Smirnovova statistika

Kolmogorovova – Smirnovova statistika (K – S) je definována jako

$$K - S = \max_{s \in \mathbf{R}}(|F_G(s) - F_B(s)|).$$

Pro výpočet této statistiky se používá ROC křivka. Uvažujme tečnu k zrcadlové ROC křivce, která je rovnoběžná s osou prvního kvadrantu (neboli s přímkou naivního modelu) a označme V jako vzdálenost této tečny od přímky naivního modelu. Pak platí:

$$K - S = \sqrt{2} \cdot V.$$

Pro výpočet se používá asymptotický vztah:

$$K - S \approx \frac{\sqrt{2}}{2} GINI.$$

Tento asymptotický vztah platí, pokud má zrcadlová ROC křivka hladký průběh. Pro většinu modelů je ROC křivky dostatečně hladkou křivkou, a tedy pro ně tento asymptotický vztah platí.

C – statistika

C – statistika je pravděpodobnost, že náhodně vybraný špatný klient má větší skóre než náhodně vybraný dobrý klient. Pro výpočet C – statistiky (nadále CS) se používá vztah:

$$GINI = 2\,CS - 1.$$

5. Parametr PD

5.1. Požadavky kladené na parametr PD

V této části postupujeme podle [1].

Hodnotící systém má několik dimenzí. V případě korporátního portfolia má dvě dimenze (PD a LGD). Je zde nutná strategie na provádění klasifikace. Jsou – li dvě expozice u jednoho dlužníka, pak se hodnotí stejně. U retailového portfolia musí být expozice rozdělené do homogenních poolů, které sdílí rizikové charakteristiky. V každém poolu je nutné odhadnout parametr PD (i LGD a EAD). Počet poolů a expozic v každém poolu musí být dostatečný k tomu, aby výsledky měly smysl. Musí být jasná a ucelená definice na rozdělení expozic do hodnotící stupnice. Je tedy nutná stálost napříč odvětvími, obory a zeměpisnými polohami a třetí strana by měla být schopna daný proces zopakovat. Banky by měly propojit všechny dostupné informace. Parametr PD se stanovuje na jeden rok, hodnocení však probíhá v delším horizontu.

Použití modelů má několik požadavků. V prvé řadě se mají použít jako primární vstup na hodnocení. Nesmí však sloužit jen jako pomůcka, protože pak bychom použili jen část dostupných informací. Dále je nutný lidský dohled na používání modelů. Banky musí ukázat sílu modelu a musí mít mechanismy na kontrolu kvality vstupů. Rovněž musí ukázat, že použitá data dostatečně reprezentují dlužníky. Lidský úsudek začleňuje části, které model nebere v potaz. Je též nutné neustále sledovat, potvrzovat a hodnotit vlastnosti modelu.

Je též nutné zadefinovat pojem selhání. Selhání se většinou definuje takto:

Definice: Selhání nastává, když se daná platba zpozdí o více než 90 dnů a zbývající dluh klienta je větší než 10 EUR, nebo i dříve za předpokladu, že je splacení nepravděpodobné.

Tato definice může být v praxi odlišná. Používá – li banka externí data, musí též znát definici selhání použitou při sběru dat (data zpravidla udávají, zda k selhání došlo, ale neudávají důvod selhání). Selhání lze zpravidla poznat dříve, než dojde k naplnění výše uvedené definice a to podle různých kritérií, např. zaopatření dlužníka, indikace bankrotu dlužníka apod.

I sám parametr PD musí splňovat několik požadavků. Předně musí být průměrem jednoletých měr selhání a tento průměr musí být dlouhodobý. Výjimkou z tohoto pravidla je PD pro retailové portfolio. Při splnění předpokladu srovnatelnosti dat se povoluje sdílení parametru PD. Odhad parametru PD se musí zakládat na praktických zkušenostech. Ekonomická situace v době sběru dat musí být vzata v potaz. Obvykle se zahrnuje rozdíl v chybě odhadu. Poslední požadavek nemusí být nutně splněn. Ze všech těchto požadavků plyne, že banka často nemá dostatek inter-

ních dat. To je klasický předsudek, který nemůže být použit ve vnitřním řízení rizik. Banka obvykle má dostatek dat na odhad parametru PD.

Odhad parametru PD musí též splňovat několik specifických požadavků. Pokud odhadujeme parametr PD pro korporátní portfolio, pak náš odhad musí splnit:

a) nutnost registrace vnitřní dlouhodobé praxe;
b) vlastní úsudek se musí vzít v potaz;
c) v případě použití externích dat je potřeba znát a zdokumentovat definici selhání a hodnotící kritéria používaná při tvorbě dat;
d) data použitá k odhadu musí být minimálně z posledních pěti let (tedy pokud bychom odhadovali parametr PD v roce 2012, potřebovali bychom data z let 2007 – 2012).

Při odhadu parametru PD pro retailové portfolio je třeba splnit:

a) přednostně se mají použít interní data;
b) v případě růstu portfolia se má začlenit sezónní efekt.

Při splnění těchto požadavků lze výsledný odhad použít pro modelování hodnot RWA funkce (viz kapitola 1.2).

5.2. Modelování parametru PD

V této části postupujeme podle [1].

5.2.1. Používané metody

Než banka začne modelovat, musí se rozhodnout, kterou metodu použije, a musí též sehnat data. Data mohou mít trojí původ. První možností jsou data, která podávají informaci o tom, zda daný subjekt selhal či neselhal. Taková data jsou nejobjektivnější, ale často nedostatková. Druhou možností jsou data získaná vnějším hodnocením. Tato data jsou často k dostání pro korporátní portfolio. Pomocí těchto dat lze nepřímo modelovat selhání. Pokud nelze získat žádná data, musí banka použít data z vnitřního hodnocení.

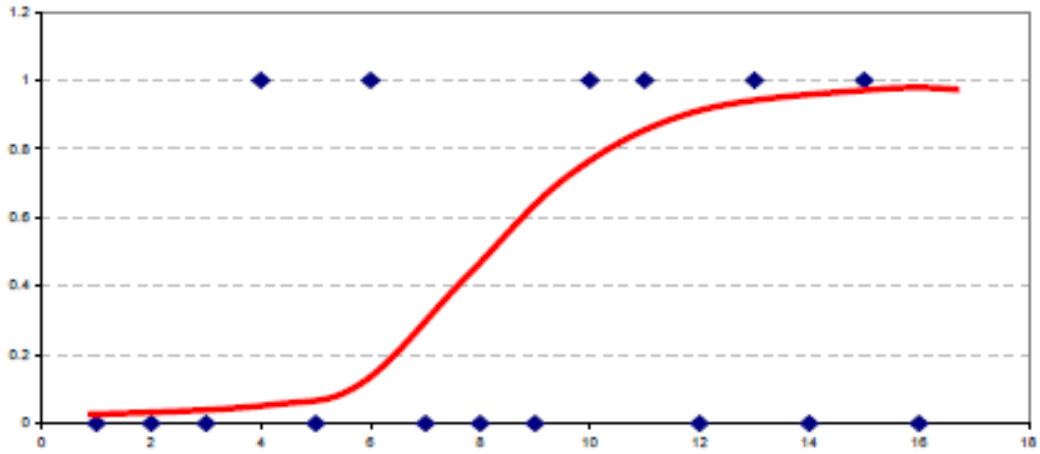
Má – li banka data, musí si vybrat metodu k modelování parametru PD. Nejčastěji používanou technikou je regrese. Používá se mnohorozměrná diskriminační analýza, metoda nejmenších čtverců nebo (nejčastěji) binární logistická regrese. V případě použití logistické regrese záleží též na charakteru dat. Máme – li data o selhání, pak se data binárně kódují například takto:

$$\text{subjekt selhal} = 1, \text{subjekt neselhal} = 0.$$

Pak je vhodné použít binární logistickou regresi. Máme – li data pocházející z vnějšího nebo vnitřního hodnocení, pak je třeba použít více tříd a k modelování použít takzvanou uspořádanou logistickou regresi. Kódování dat na základě hodnocení popřípadě ratingů může být například:

$$\mathrm{AAA} = 1, \mathrm{AA+} = 2, \ldots, \mathrm{B-} = 16, \mathrm{CCC} = 17.$$

V obou případech se však použije regrese, která připomíná vyrovnávání dat pomocí logistického trendu. Příklad tohoto vyrovnání pro data o selhání je na obr. 3, na kterém jsou modře znázorněna data a červeně vyrovnání pomocí logistické regrese.

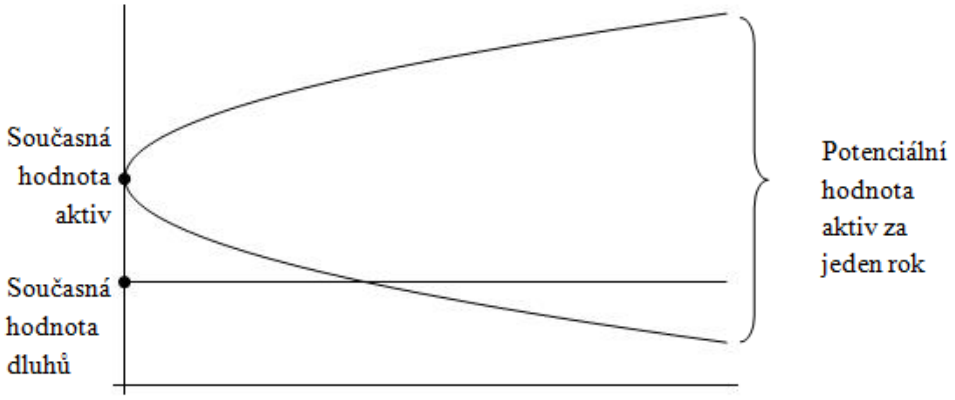


Obr. 3 – příklad logistické regrese

Další metody jsou méně časté a občas nejsou povolené legislativou nebo regulátorem. Jednou z možností jsou strukturální modely, které při modelování parametru PD sledují strukturu portfolia. Příkladem takových modelů je KMV model (viz též [10]). Označme g jako funkci tří proměnných. Pak pomocí KMV modelu modelujeme parametr PD takto:

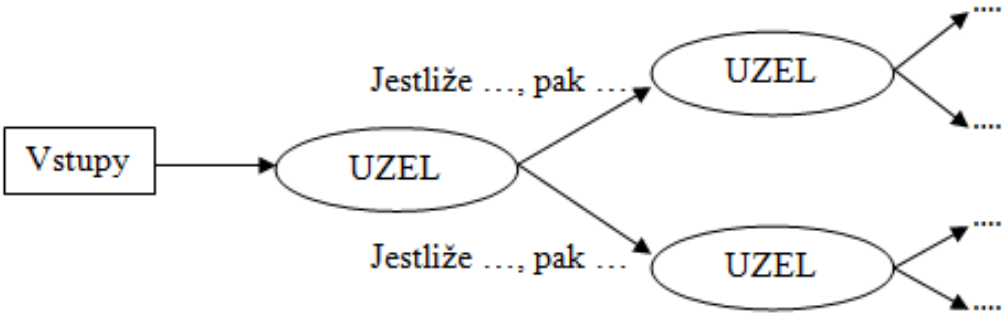
$$\mathrm{PD} = \mathrm{g}(\text{hodnota aktiv}, \text{volatilita aktiv}, \text{hodnota dluhu}).$$

Možné schéma strukturálního modelu je na obr. 4.

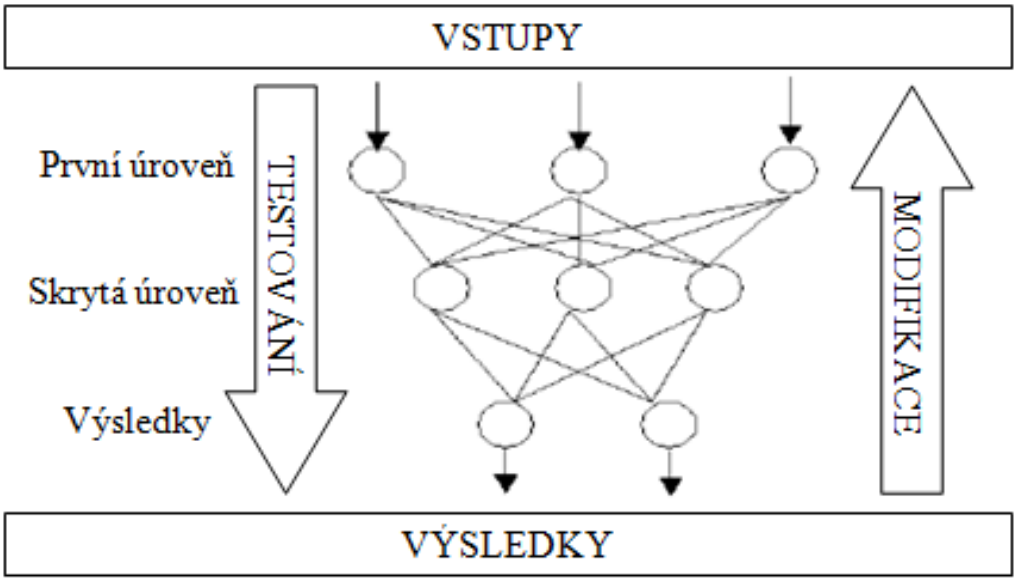


Obr. 4 – schéma strukturálních modelů

Další možností jsou takzvané expertní systémy. Příkladem expertního systému je rozhodovací strom, což znamená, že se v každém kroku rozhodujeme na základě současného stavu. Příklad binárního rozhodovacího stromu uvádí schéma na obr. 5. Poslední možností je použití induktivních metod a postupů. Příkladem těchto metod je metoda Neural Networks, jejíž příklad znázorňuje schéma na obr. 6.



Obr. 5 – příklad rozhodovacího stromu



Obr. 6 – příklad metody Neural Networks

5.2.2. Postup stavby modelu

Při stavbě modelu se postupuje podle předem daného postupu v osmi krocích. První krok spočívá ve sběru dat. Banka musí vědět, jaká data jsou k dispozici, co

obsahují, jak jsou spolehlivá a zda je vhodné je očistit například od vlivu prostředí, státu, pohlaví a podobně. Též je nutné doplnit chybějící data. Dále se provádí takzvaná univaried analysis. V rámci této analýzy se zkoumá vztah mezi každou proměnnou a průměrnou hodnotou PD, resp. průměrným hodnocením. Zkoumá se též tvar dané závislosti (zda je monotonní či nikoli). Součástí této analýzy je též výběr proměnných, které budou součástí modelu a které nikoli. Pak se provádí transformace poměru. Touto transformací se zpravidla docílí monotonie daného vztahu a dojde tak k očištění vztahu mezi selháním a danou proměnnou od nežádoucích vlivů.

Následuje regresní analýza. Nejčastěji se používá binární logistická regrese, která je pro modelování pravděpodobnosti nejvhodnější a byla detailně popsána v kapitole 3. Modeluje pravděpodobnost selhání jako:

$$\tilde{\pi}(\mathbf{x}) = 1 - \pi(\mathbf{x}) = \frac{1}{1 + \exp[-(b_1 x_1 + b_2 x_2 + \cdots + c)]}, \tag{5.1}$$

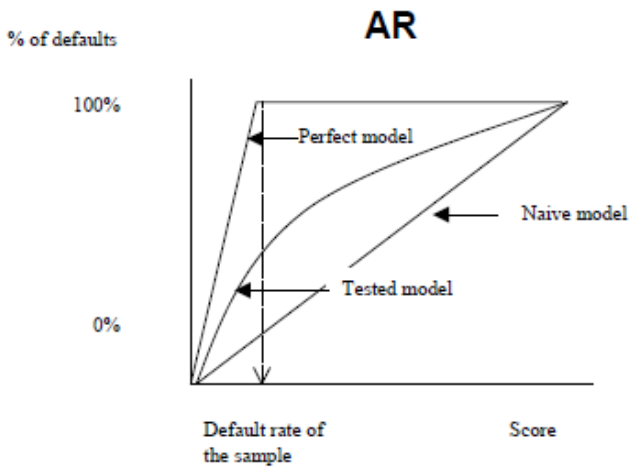
při použití

$$c = -\beta_0, b_1 = -\beta_1, b_2 = -\beta_2, \dots, b_p = -\beta_p.$$

Mimo logistické regrese se občas používá lineární regrese, která však není nejvhodnějším modelem (viz kapitola 3.1 a 5.3.5).

Následuje ověření správnosti modelu. K tomu se používají statistické a ekonomické testy. Často se používá p – hodnota, logaritmická věrohodnost, Waldův test, poměr přesnosti (získá se z počtu selhání), kumulativní stupeň rozdílu (získá se z ratingů), a testy a statistiky popsané v kapitole 3.4 a v kapitole 4. Zkoumá se též AR – diagram, jenž je na obr. 7. Čím je křivka testovaného modelu blíže přímce perfektního modelu, tím je model lepší. Výsledné hodnoty se rovněž porovnávají s hodnotami z veřejných modelů (např. Altmanův model), s komerčními modely, s hodnotami z modelů ostatních bank a s dalšími možnými výsledky.

Dalším krokem je kalibrace modelu. Při kalibraci záleží na charakteru použitých dat. Pokud banka použila data o selhání, musí ověřit, zda odpovídají definici dle BASEL II. Pokud neodpovídají, je nutné data upravit tak, aby odpovídala. Při použití dat o hodnocení se musí uvést zdroj informací o selhání a uvést definice selhání. Součástí kalibrace bývá též menší úprava modelu popsaná v kapitole 5.3.8. Dále následuje testovací fáze. V této fázi se ověřuje správnost všeho předchozího. Posledním krokem je sestavení odpovídajícího počítačového programu. Pak je již model hotov a může se snadno používat.



Obr. 7 – AR – diagram

5.3. Výstavba modelu binární logistické regrese a jeho alternativy

Model binární logistické regrese je jedním z příkladů modelů pro modelování parametru PD. V této kapitole sestavíme takový model z dat.

5.3.1. Data

Máme soubor dat od banky, která působí na bankovním trhu v České republice, který obsahuje cca 130 tisíc záznamů z portfolia v období říjen 2009 až říjen 2011. Data se vztahují k běžným účtům (nadále BÚ) a k nim připojeným kontokorentním úvěrům (nadále OVD). Každá záznam obsahuje charakteristiky, které se vztahují k BÚ nebo k OVD, popřípadě k vlastníkovi účtu:

1) id – Číslo daného záznamu
2) BalMinPast4Q (BÚ, nadále X_1) – Minimální zůstatek na účtu v průběhu posledního roku
3) DebtMonthCntPast4Q (BÚ, nadále X_2) – Počet měsíců posledního roku, ve kterých měl nějaký dluh na účtu
4) PmtOrderOutsideCnt (BÚ, nadále X_3)* – Počet platebních příkazů v aktuálním měsíci
5) LifeInsurance (BÚ, nadále X_4)**** – Platba za životní pojištění v aktuálním měsíci
6) PensionInsurance (BÚ, nadále X_5)**** – Platba za penzijní pojištění v aktuálním měsíci

7) BuildingSaving (BÚ, nadále X_6)**** – Platba na stavební spoření v aktuálním měsíci
8) StandOrderCnt_6M (BÚ, nadále X_7)* – Počet trvalých příkazů z účtu za posledních 6 měsíců
9) EBWithdCnt_3M (BÚ, nadále X_8)* – Počet čerpaní přes internetové bankovnictví za poslední 3 měsíce
10) CardStoreCntAvg_3M (BÚ, nadále X_9)* – Průměrný měsíční počet transakcí platební kartou v obchodech za poslední 3 měsíce
11) OutsideDeposMin_6M_cat (BÚ, nadále X_{10})* – Minimální měsíční vklad na účet během předchozích 6 měsíců
12) ODDelqStatPast (OVD, nadále X_{11})* – Delikvenční status na OVD účtě v minulem měsíci (1 = 0 – 30 DPD, 2 = 31 – 60 DPD, 3 = 61 – 90 DPD, atd.), DPD = dní po delikvenci
13) ODDebtMaxPast4Q (OVD, nadále X_{12})* – Maximální dluh na OVD za poslední rok
14) ODDelqCntPast1Q (OVD, nadále X_{13})* – Počet delikvencí na OVD za poslední 3 měsíce
15) ODUtilAmtAct (OVD, nadále X_{14}) – Vyčerpaná částka na OVD ke konci aktuálního měsíce
16) ODLimitAmtPast (OVD, nadále X_{15})* – Výše úvěrového limitu na OVD v předchozím měsíci
17) BalAvg (OVD, nadále X_{16})** – Průměrná vyčerpaná částka na OVD během aktuálního měsíce
18) DrawAvg (OVD, nadále X_{17})** – Průměrná výše čerpaní během aktuálního měsíce
19) DbCnt (OVD, nadále X_{18})** – Počet debetních operací na OVD účtu v aktuálním měsíci
20) PenIntPaid (OVD, nadále X_{19})*** – Zaplacené sankční úroky během aktuálního měsíce
21) Gender (nadále X_{20}) – Pohlaví (Male/Female)
22) ResidFlg (nadále X_{21}) - Resident (1/0) zda má trvaly pobyt v CR (1 = ano, 0 = ne)
23) bad - (1/0) Selhání klienta během následujících 12 měsíců (1 = ano, 0 = ne)

Položky označené jednou, dvěma, třemi nebo čtyřmi hvězdičkami obsahují chybějící hodnoty. Ty je nutné doplnit. V položkách označených čtyřmi hvězdičkami doplníme chybějící hodnoty nulami na základě ideje, že daný klient nemá životní pojištění, potažmo penzijní pojištění, respektive stavební spoření. V položkách označených jednou hvězdičkou doplníme rovněž nuly na základě ideje, že daná položka je nulová. Pro každou položku sestavíme histogram z dat s doplněnými nulami a histogram z dat bez vynechaných hodnot. Histogramy jsou prakticky totožné. Proto ponecháme toto doplnění jako směrodatné. V položkách označených dvěma nebo třemi hvězdičkami doplnění nul výrazně mění histogramy a tedy i celé rozdělení dat. V případě položek označených dvěma hvězdičkami proto doplníme chybějící hodno-

ty tak, že na místo chybějící hodnoty doplníme náhodně vybranou hodnotu z ostatních hodnot. Toto doplnění již histogram významně nezmění. V případě položky označené třemi hvězdičkami ponecháme doplnění nul jako směrodatné, protože téměř všechny hodnoty v této položce jsou nulové i bez doplnění nul na vynechaná místa.

Položka bad je sestavená dle definice selhání dle BASEL II, tedy k selhání došlo 90 dní po splatnosti. Tím je tato definice ověřena. Ještě dodefinujeme dodatečnou položku pohlaví. Význam této položky je zřejmý z názvu. Zavádíme ji, protože položka X_{20}, která má tentýž význam, je kódována systémem M / F. Toto kódování nelze použít pro výpočty. Proto položku X_{20} překódujeme do položky pohlaví podle klíče: 0 = muži, 1 = ženy. Nadále budeme používat položku pohlaví (nadále X_{20}).

Vzhledem k velkému množství dat by při použití všech pozorování mohlo dojít k přefitování modelu. Model by pak nemusel být správný. Z tohoto důvodu rozdělíme náhodně data na dvě poloviny. Na první polovině provedeme výběr proměnných a odhad modelu, na druhé polovině pak provedeme všechny testy. Rozdělení dat na dvě poloviny je zcela náhodné. Počet jedničkových pozorování v obou skupinách je téměř totožný, liší se o 55 pozorování.

5.3.2. Výběr proměnných pro odhad modelu

Máme k dispozici 21 proměnných. Z nich vybereme ty proměnné, které budou mít vliv na odhad modelu. Pro výběr proměnných použijeme algoritmus Stepwise. Pro potřeby algoritmu zvolíme

$$p_E = 0{,}15;\ p_R = 0{,}20.$$

Ještě je nutné vyřešit otázku počtu stupňů volnosti. Z popisu dat je zjevné, že většina proměnných není kategoriální, tedy uvažujeme pro ně jeden stupeň volnosti. Proměnné X_{20} a X_{21} jsou kategoriální se dvěma kategoriemi, což znamená, že pro tyto proměnné máme též jeden stupeň volnosti. Proměnná X_{11} se jeví jako kategoriální, protože uvádí delikvenční status, ale kategoriální být nemusí, protože tento status může být libovolně velký, neboť doba po delikvenci může být libovolně dlouhá. Pro tuto proměnnou tedy budeme uvažovat jedničkový počet stupňů volnosti. Výběr proměnných probíhá v následujících krocích:

KROK (0): Nejprve odhadneme model s konstantním členem. Jeho logaritmická věrohodnost je $-5738{,}56$. Chceme vybrat první proměnnou. Pro všechny proměnné odhadneme model, spočítáme logaritmické věrohodnosti a p – hodnoty a hledáme nejnižší hodnotu p – hodnoty. Minimální p – hodnota je téměř nulová (tj. velmi malá, řádově menší než 10^{-50}), což je menší než p_E, a sice pro několik proměnných. Algoritmus neuvádí, jak se v takovém případě máme zachovat. Proto proměnnou, kterou přidáme, zvolíme na základě logaritmických věrohodností modelu, protože model

s nejvyšší logaritmickou věrohodností má nejmenší p – hodnotu. Přidáme proměnnou X_2.

KROK (1): Logaritmická věrohodnost modelu je $-4992{,}67$. V tomto kroku zatím nevyřazujeme žádnou proměnnou, pouze zkoušíme přidat další. Spočítáme příslušné p – hodnoty a najdeme nejnižší, která je opět téměř nulová, což je menší než p_E, pro několik proměnných. Proměnnou, kterou přidáme, zvolíme stejně, jako výše. Přidáme proměnnou X_{10}.

KROK (2): Logaritmická věrohodnost modelu je $-4765{,}35$. Nejprve zkusíme nějakou proměnnou vyřadit. Maximální p – hodnota pro vyřazení je téměř nulová, tedy menší než p_R a tedy žádnu proměnnou nevyřadíme. Pak spočítáme p – hodnoty pro přidání další proměnné. Minimální p – hodnota je téměř nulová, což je menší než p_E, opět pro několik proměnných. Proto proměnnou, kterou přidáme, zvolíme stejně, jako výše. Přidáme proměnnou X_{14}.

KROK (3): Logaritmická věrohodnost modelu je $-4649{,}02$. Nejprve spočítáme p – hodnoty pro vyřazení. Nejvyšší p – hodnota je téměř nulová, tedy menší než p_R, a proto žádnou proměnnou nevyřadíme. Pak spočítáme p – hodnoty pro přidání další proměnné. Nejnižší p – hodnota je téměř nulová, což je menší než p_E, pro několik proměnných. Proměnnou, kterou přidáme, zvolíme stejně, jako výše. Přidáme proměnnou X_{15}.

KROK (4): Logaritmická věrohodnost modelu je $-4567{,}36$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné téměř nulová, tedy menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je téměř nulová, což je menší než p_E, pro několik proměnných, a proto proměnnou, kterou přidáme, vybereme stejně, jako výše. Přidáme proměnnou X_8.

KROK (5): Logaritmická věrohodnost modelu je $-4507{,}3$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné téměř nulová, tedy menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je téměř nulová, což je menší než p_E, pro proměnnou X_{13}, tedy tuto proměnnou přidáme.

KROK (6): Logaritmická věrohodnost modelu je $-4471{,}41$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné téměř nulová, tedy menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je rovna $2{,}30038 \cdot 10^{-13}$, což je menší než p_E, pro proměnnou X_9, tedy tuto proměnnou přidáme.

KROK (7): Logaritmická věrohodnost modelu je $-4444{,}54$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné rovna $2{,}77556 \cdot 10^{-15}$, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je rovna $5{,}9591 \cdot 10^{-11}$, což je menší než p_E, pro proměnnou X_{21}, tedy tuto proměnnou přidáme.

KROK (8): Logaritmická věrohodnost modelu je $-4423{,}13$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné $3{,}9968 \cdot 10^{-15}$, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je rovna $1{,}61138 \cdot 10^{-6}$, což je menší než p_E, pro proměnnou X_7, tedy tuto proměnnou přidáme.

KROK (9): Logaritmická věrohodnost modelu je $-4411{,}62$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné $1{,}2863 \cdot 10^{-8}$, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je rovna 0,000128, což je menší než p_E, pro proměnnou X_{12}, tedy tuto proměnnou přidáme.

KROK (10): Logaritmická věrohodnost modelu je $-4404{,}29$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné 0,000001, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je rovna 0,00439, což je menší než p_E, pro proměnnou X_6, tedy tuto proměnnou přidáme.

KROK (11): Logaritmická věrohodnost modelu je $-4400{,}23$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné 0,00013331, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je rovna 0,0091778, což je menší než p_E, pro proměnnou X_{18}, tedy tuto proměnnou přidáme.

KROK (12): Logaritmická věrohodnost modelu je $-4396{,}84$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné 0,003779, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je 0,01129, což je menší než p_E, pro proměnnou X_{20}, tedy tuto proměnnou přidáme.

KROK (13): Logaritmická věrohodnost modelu je $-4393{,}63$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné rovna 0,011, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je 0,1011, což je menší než p_E, pro proměnnou X_{17}, tedy tuto proměnnou přidáme.

KROK (14): Logaritmická věrohodnost modelu je $-4392{,}28$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné rovna 0,0116, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je 0,133, což je menší než p_E, pro proměnnou X_5, tedy tuto proměnnou přidáme.

KROK (15): Logaritmická věrohodnost modelu je $-4391{,}16$. V tomto kroku je nejvyšší p – hodnota pro vyřazení proměnné rovna 0,109, tedy je menší než p_R, takže žádnou proměnnou neodebereme. Nejmenší p – hodnota pro přidání proměnné je 0,21, což je větší než p_E, tedy žádnou další proměnnou nepřidáme a výběr proměnných tak končí.

Vybrali jsme patnáct proměnných.

Provedeme KROK (S) (viz kapitola 3.3), abychom otestovali, zda všechny vybrané proměnné mají význam. K tomu zvolme

$$\alpha = 0{,}05.$$

Uvedeme vždy pouze p – hodnoty pro daný krok, přičemž pokud bude daná p – hodnota velmi malá, uvedeme namísto konkrétní hodnoty pouze $< 0{,}0001$. V případě použití první metody dostáváme postupně výsledky:

- KROK (1) – KROK (9): $< 0{,}0001$
- KROK (10): $0{,}000128$
- KROK (11): $0{,}00439$
- KROK (12): $0{,}0092$
- KROK (13): $0{,}0113$
- KROK (14): $0{,}10112$
- KROK (15): $0{,}133$

Při použití druhé metody dostaneme výsledky:

- KROK (0) – KROK (10): $< 0{,}0001$
- KROK (11): $0{,}00115$
- KROK (12): $0{,}00993$
- KROK (13): $0{,}08451$
- KROK (14): $0{,}133$

Tedy z výsledků obou metod plyne, že poslední dvě proměnné (tj. proměnné X_{17} a X_5) nejsou na hladině 0,05 významné, a proto je do modelu nezahrneme. Všechny ostatní jsou významné. Máme tedy model se třinácti proměnnými.

5.3.3. Odhad modelu, jeho vlastnosti a ověřování

Odhadneme model binární logistické regrese ze vzorce (5.1) obsahující třináct výše vybraných proměnných. Výsledek odhadu parametrů na základě dat pro odhad je ve druhém sloupci tabulky 8 (význam položky „Konst.“ je zřejmý). Všechny hodnoty jsou zaokrouhlené na pět desetinných míst.

Máme odhadnutý model a všechny regresní parametry. Kromě odhadu též otestujeme nulovost jednotlivých parametrů na základě dat pro odhad. V třetím sloupci tabulky 8 jsou uvedeny p – hodnoty testu nulovosti, tedy testu poměrem věrohodností pro každý parametr. Z tabulky 8 je zjevné, že na hladině 0,05 hypotézu nulovosti zamítáme u všech parametrů. Rovněž pro všechny parametry sestrojíme intervaly spolehlivosti. Výsledky jsou v tabulce 8 ve čtvrtém a pátém sloupci.

Prom.	Odhad	P – hodnota	Interval spolehlivosti	
			Dolní mez	Horní mez
Konst.	$-3{,}11591$	$4{,}85598 \cdot 10^{-173}$	$-3{,}33368$	$-2{,}89813$
X_2	$0{,}21898$	$7{,}60389 \cdot 10^{-56}$	$0{,}19171$	$0{,}24624$
X_6	$-0{,}00008$	$0{,}01450$	$-0{,}00014$	$-0{,}00002$
X_7	$-0{,}01781$	$0{,}00003$	$-0{,}02618$	$-0{,}00944$
X_8	$-0{,}02697$	$7{,}96095 \cdot 10^{-8}$	$-0{,}03682$	$-0{,}01712$
X_9	$-0{,}10070$	$8{,}43434 \cdot 10^{-14}$	$-0{,}12715$	$-0{,}07426$
X_{10}	$-0{,}00004$	$6{,}17657 \cdot 10^{-20}$	$-0{,}00005$	$-0{,}00003$
X_{12}	$0{,}00008$	$0{,}00001$	$0{,}00004$	$0{,}00011$
X_{13}	$0{,}26946$	$2{,}7608 \cdot 10^{-12}$	$0{,}19390$	$0{,}34502$
X_{14}	$0{,}00007$	$1{,}29971 \cdot 10^{-33}$	$0{,}00006$	$0{,}00008$
X_{15}	$-0{,}00004$	$1{,}1863 \cdot 10^{-13}$	$-0{,}00005$	$-0{,}00003$
X_{18}	$0{,}01210$	$0{,}01023$	$0{,}00286$	$0{,}02133$
X_{20}	$-0{,}16506$	$0{,}01157$	$-0{,}29318$	$-0{,}03693$
X_{21}	$-0{,}54195$	$4{,}58957 \cdot 10^{-9}$	$-0{,}72316$	$-0{,}36073$

Tabulka 8 – odhad modelu binární logistické regrese a jeho vlastnosti

Na základě testovacích dat provedeme některé testy modelu uvedené v kapitole 3.4. Hosmer – Lemeshowova statistika pro tento odhad je

$$C = 54{,}726.$$

Kvantil chí – kvadrát rozdělení s osmi stupni volnosti na hladině 0,95 je 15,5. Tedy C je výrazně větší než příslušný kvantil, p – hodnota testu je

$$p = 4{,}99 \cdot 10^{-9},$$

a tedy náš model je špatný na hladině 0,05. O tom, proč je model špatný, pojednáme později. McFaddenův koeficient determinace je

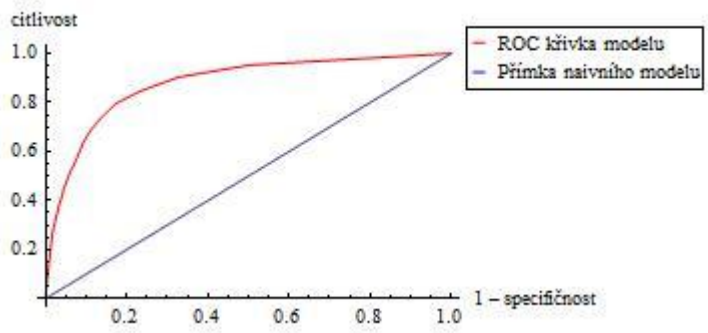
$$R^2 = 0{,}2028.$$

S těmito hodnotami budeme nadále srovnávat výsledky alternativních modelů.

Nyní prozkoumáme diverzifikační schopnosti modelu na základě dat pro test. Graf ROC křivky ve srovnání s přímkou naivního modelu je znázorněn na obr. 8. Spočítáme též plochu pod grafem ROC křivky podle návodu v kapitole 4. Vezmeme tedy všechny dvojice, ve kterých je vždy jedno pozorování nulové a druhé jedničkové. Těchto dvojic je 72 758 730. Dále spočítáme četnost případů, kdy pozorování s $y = 1$ má větší odhadnutou pravděpodobnost než pozorování s $y = 0$. Těchto případů je 63 982 592. Odsud spočítáme plochu pod ROC křivkou:

$$\frac{63\,982\,592}{72\,758\,730} = 0{,}87938.$$

Plocha po ROC křivkou je téměř 0,9, tedy odhadnutý model je perfektní.



Obr. 8 – ROC křivka modelu ve srovnání s přímkou naivního modelu

Interpretace citlivosti a specifičnosti vychází z dat a z tabulky 7. V tomto konkrétním případě máme

$$1 = \text{klient selhal}, 0 = \text{ klient neselhal}.$$

To znamená, že citlivost je poměr počtu klientů, kteří selhali a pro které model na určité hladině předpověděl, že selžou, a počtu klientů, kteří opravdu selhali. Specifičnost je pak poměr počtu klientů, kteří neselhali a pro které model na určité hladině předpověděl, že neselžou, a počtu klientů, kteří opravdu neselhali. Zjevně:

$$\text{citlivost} \leq 1, \text{specifičnost} \leq 1.$$

Položka w v kapitole 4 je pak doplněk specifičnosti do 1, a značí poměr počtu klientů, kteří neselhali, ale model pro ně na určité hladině předpověděl, že selžou, a počtu klientů, kteří ve skutečnosti neselhali. Jak již bylo zmíněno výše, ROC křivka znázorňuje závislost citlivosti na doplňku specifičnosti do 1. Uvažujme na vodorovné ose bod 0,05. Z obr. 6 odhadujeme, že hodnota ROC křivky v bodě 0,05 je 0,45. To znamená, že pokud model u 5 % klientů, kteří neselhali, předpoví selhání, pak u 45 % klientů, kteří selhali, předpoví selhání. Podobná úvaha lze provést i při jiné volbě bodu na vodorovné ose. Z toho plyne poměrně vysoká spolehlivost modelu.

Pomocné hodnoty pro Giniho koeficient jsou

$$B = 1067,$$

$$G = 68\,190,$$

$$S = 6{,}4552377 \cdot 10^7,$$

$$S_{\min} = 569\,778$$

a

$$S_{max} = 73\,328\,508.$$

Odsud

$$\mathrm{GINI} = 0{,}75876.$$

Kolmogorovova – Smirnovova statistika je asymptoticky rovna

$$K - S = 0{,}53653.$$

C – statistika je rovna

$$CS = 0{,}87938.$$

5.3.4. Alternativní možnosti k modelu

V této kapitole navrhneme alternativy k odhadnutému modelu. Budeme je navrhovat na základě dat i bez dat, tedy částečně teoretické. Alternativy navrhnuté na základě dat vždy odhadneme (uvedeme pouze vyrovnané hodnoty bez p – hodnot testu nulovosti a intervalů spolehlivosti) a porovnáme s výše odhadnutým modelem na základě plochy pod ROC křivkou a na základě Hosmer – Lemeshowovy statistiky při rozdělení na deset skupin a McFaddenova koeficientu determinace.

a) Alternativy navrhnuté na základě dat

Vezměme proměnné X_5 a X_4, které mají podobný význam, tedy platby za životní, resp. penzijní, pojištění, a sečtěme je po složkách. Tím zadefinujeme proměnnou pojištění a budeme zkoumat, zda takto zadefinovaná proměnná má význam pro model. Provedeme – li test poměrem věrohodností, tedy jeden krok algoritmu Stepwise, pro tuto proměnnou, dostaneme p – hodnotu 0,1981, což je větší než 0,15 a tedy tuto proměnnou do modelu vůbec nepřidáme. Namísto proměnné pojištění uvažujeme proměnnou indpojisteni (nadále N_1), která oznamuje, zda má klient nějaké pojištění, přičemž vynechané a nulové hodnoty v položkách X_5 a X_4 považujeme za indikaci toho, že klient dané pojištění nemá. Tato proměnná je dvoukategoriální s kategoriemi 0 = klient žádné pojištění nemá a 1 = klient má nějaké pojištění. Pro tuto proměnnou provedeme test poměrem věrohodností a dostaneme p – hodnotu rovnu 0,011879, což je menší než 0,15 a tedy tuto proměnnou již do modelu přidáme. Ještě zkontrolujeme, zda přidáním této proměnné nedošlo ke ztrátě významnosti jiné tak, jako jsme to provedli při samotném výběru proměnných. Nejvyšší p – hodnota pro vyřazení proměnné vychází 0,01306, což je menší než 0,20 a tedy žádnou proměnnou neodebereme. Odhadnuté hodnoty parametrů jsou v tabulce 9. Plocha pod ROC křivkou je 0,87989, samotná ROC křivka je stejná jako křivka původního modelu. Dále máme:

$$C_1 = 40{,}87$$

a

$$R_1^2 = 0{,}2033.$$

Tedy

$$C > C_1, R^2 < R_1^2,$$

tedy původní model je horší než tento model.

Prom.	Odhad
Konst.	3,11223
X_2	0,21862
X_6	−0,00008
X_7	−0,01248
X_8	−0,02575
X_9	−0,10118
X_{10}	−0,00004
X_{12}	0,00008
X_{13}	0,26433
X_{14}	0,00007
X_{15}	−0,00004
X_{18}	0,01253
X_{20}	−0,16177
X_{21}	−0,52403
N_1	−0,23181

Tabulka 9 – odhad prvního alternativního modelu

Vezměme proměnnou X_{13}, která značí počet delikvencí za poslední tři měsíce a nahraďme ji proměnnou inddelikvence (nadále N_2), pro kterou 0 = nedošlo k delikvenci za poslední tři měsíce, 1 = došlo k delikvenci za poslední tři měsíce. Tuto proměnnou zkusíme přidat do modelu. Opět provedeme jeden krok přidání do modelu a následné ověření, zda jiná proměnná tímto přidáním neztratila význam. Pro přidání máme p – hodnotu rovnu $6{,}3138 \cdot 10^{-13}$, tedy menší než 0,15 a tedy proměnnou přidáme, nejvyšší p – hodnota pro vyřazení je 0,01742, což je menší než 0,20 a tedy žádnou proměnnou neodebereme. Odhadnuté hodnoty parametrů jsou v tabulce 10. ROC křivka je stejná jako křivka původního modelu, plocha pod ROC křivkou je 0,8798. Dále máme:

$$C_2 = 53{,}1$$

a

$$R_2^2 = 0{,}2031.$$

Tedy

$$C > C_2, R^2 < R_2^2,$$

tedy původní model je horší než tento model.

Prom.	Odhad
Konst.	$-3{,}22079$
X_2	$0{,}23813$
X_6	$-0{,}00008$
X_7	$-0{,}01804$
X_8	$-0{,}02776$
X_9	$-0{,}10383$
X_{10}	$-0{,}00004$
X_{12}	$0{,}00008$
X_{14}	$0{,}00006$
X_{15}	$-0{,}00003$
X_{18}	$0{,}01366$
X_{20}	$-0{,}15465$
X_{21}	$-0{,}53720$
N_2	$0{,}59531$

Tabulka 10 – odhad druhého alternativního modelu

Prom.	Odhad
Konst.	$-2{,}80016$
X_2	$0{,}22316$
X_6	$-0{,}00010$
X_{10}	$-0{,}00004$
X_{12}	$0{,}00008$
X_{13}	$-0{,}00004$
X_{14}	$0{,}00007$
X_{15}	$-0{,}00004$
X_{18}	$0{,}00581$
X_{20}	$-0{,}19621$
X_{21}	$-0{,}51858$
N_3	$-0{,}44726$

Tabulka 11 – odhad čtvrtého alternativního modelu

Stejným nebo podobným způsobem lze nahradit kteroukoli jinou proměnnou. Uvažujeme odchozí transakce v minulosti. Ty jsou v proměnných X_7, X_8 a X_9, přičemž všechny jsou v modelu. Jejich nevýhoda je, že jsou brány za různé časové období, navíc nejsou všechny stejně interpretované (někdy je to přímo počet, jindy průměrná hodnota). Proto tyto proměnné nebudeme sčítat, ale budeme zkoumat indi-

kace odchozích plateb. Vše budeme uvažovat v horizontu posledních šesti měsíců, protože pokud nějaká transakce proběhla v posledních třech měsících, tak proběhla i v posledních šesti měsících, přičemž předpokládáme, že pokud k dané transakci nedošlo za poslední tři měsíce, tak k ní nedošlo ani za posledních šest měsíců. Z každé proměnné vezmeme indikátor toho, že došlo k platbě daným způsobem, pak tyto indikátory sečteme a dostaneme čtyř – kategoriální proměnnou indodchozichtransakci (nadále N_3) s kategoriemi:

0...nedošlo k žádné odchozí transakci,

1...docházelo k transakcím jedním způsobem (nezáleží na tom, jakým),

2...docházelo k transakcím dvěma způsoby (nezáleží na tom, jakými),

3...docházelo k transakcím všemi třemi způsoby.

Vyřadíme z modelu proměnné X_7, X_8 a X_9 a nahradíme je výše nadefinovanou proměnnou N_3. Nejprve však zjistíme, zda tuto proměnnou do modelu přidáme a zda přidání této proměnné nevyřadí z modelu jinou proměnnou. P – hodnota pro přidání proměnné je téměř nulová, což je menší než 0,15, tedy proměnnou přidáme, nejvyšší p – hodnota pro vyřazení je rovna 0,1968, což je menší než 0,20, tedy žádnou proměnnou neodebereme. Odhadnuté hodnoty parametrů jsou v tabulce 11. ROC křivka je stejná jako křivka původního modelu, plocha pod ROC křivkou je 0,8854. Dále máme:

$$C_3 = 49{,}4$$

a

$$R_3^2 = 0{,}20309.$$

Tedy

$$C > C_3, R^2 < R_3^2,$$

tedy původní model je horší než tento model.

Ještě uvedeme model bez konstantního členu, protože můžeme požadovat, aby model neobsahoval konstantní člen. Odhad modelu je v tabulce 12. ROC křivka tohoto modelu ve srovnání s původním modelem je na obr. 9. Plocha pod ROC křivkou je 0,85243, tedy tento model je o něco horší než původní model. Dále máme:

$$C_4 = 1894{,}88$$

a

$$R_4^2 = 0{,}1096.$$

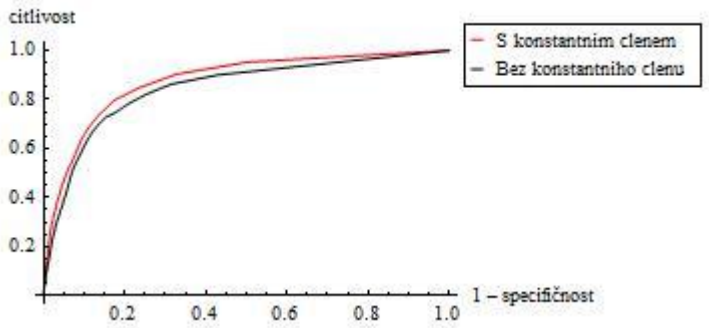
Tedy

$$C < C_4, R^2 > R_4^2,$$

tedy původní model je lepší než tento model.

Prom.	Odhad
X_2	0,18091
X_6	−0,00005
X_7	−0,01514
X_8	−0,01694
X_9	−0,18137
X_{10}	−0,00006
X_{12}	0,00009
X_{13}	0,08420
X_{14}	0,00012
X_{15}	−0,00012
X_{18}	−0,03509
X_{20}	−0,72932
X_{21}	−1,98484

Tabulka 12 – odhad modelu bez konstantního členu



Obr. 9 – srovnání ROC křivky modelu s konstantním členem a bez konstantního členu

Ze všech uvedených výsledků je zjevné, že odebrání konstantního členu má největší dopad na výsledky a jejich správnost, neboť v původním modelu je p – hodnota testu nulovosti nejmenší ze všech a výrazně menší než všechny ostatní. Z toho plyne, že konstantní člen bude významný, tedy jeho odebrání má vliv na výsledky, což se ukázalo. Dále:

$$R_4^2 < R^2 < R_3^2 < R_2^2 < R_1^2.$$

Tedy model bez konstantního členu je nejhorší a, dle výsledků výše, je výrazně horší než všechny ostatní modely. Model s indikátorem nějakého pojištění je lepší než původní model, model s indikátorem delikvencí je ještě lepší a model s indikátorem plateb za pojištění je nejlepší ze všech modelů. ROC křivka označuje všechny modely jako perfektní. Nicméně vzhledem k výsledkům Hosmer – Lemeshowovy statistiky nelze žádný model považovat za dobrý. Později pojednáme o tom, proč jsou všechny modely špatné.

b) Alternativy navrhnuté bez dat

V této části navrhneme alternativy, ke kterým nemáme dostupná data. Bude se jednat o další přidávání proměnných, nicméně navrhnuté proměnné nemusí být v modelu zahrnuté, neboť test poměrem věrohodností nemusí tyto proměnné zařadit do modelu. Zároveň přidání navrhovaných proměnných může mít za následek vyřazení proměnných, které jsou obsažené v modelu.

V případě informací o BÚ máme k dispozici mnoho informací. Nicméně by bylo možné přidat informaci o výběrech z bankomatu, popřípadě prostřednictvím pošty. Data obsahují informaci o transakcích platební kartou, přes internetové bankovnictví a prostřednictvím trvalých a jiných platebních příkazů. Výběry z bankomatu však provádí každý. Lidé, žijící na vesnicích jinou možnost nemají, protože obchody na vesnicích nejsou na jiný typ platby zařízené. Z tohoto důvodu by měla být tato položka uvažována, přičemž nezáleží na tom, zda uvažujeme výši výběru nebo počty výběrů.

Informace o OVD by bylo možné doplnit o informaci o úrokové míře uvalené na daný OVD. Pokud je však úroková míra neměnná v čase, pak tato informace ztrácí smysl. Proto by bylo vhodné zaznamenat tuto informaci v případě, kdy se ve sledovaném období mění úroková míra. Zároveň je tato proměnná zbytečná v případě, kdy je úroková míra pro všechny účty stejná.

Informaci o osobě klienta je vhodné doplnit informací o věku. Lze předpokládat, že lidé v určitých věkových kategoriích budou mít jinou platební morálku než lidé v jiných věkových kategoriích. Dále by bylo vhodné doplnit informaci o zaměstnání klienta, respektive o jeho platu, kterou získáme z výpisů. Nicméně tato možnost je poměrně kontroverzní z hlediska morálky a zákonů. Poslední uvedenou možností je uvedení bližší informace o bydlišti klienta. Tuto informaci podává proměnná ResidFlg, nicméně v případě residentů v České republice lze uvést navíc kraj, ve kterém resident bydlí. V případě, kdy víme, že všechny záznamy pocházejí z České republiky, můžeme uvažovat velikost obce, ve které klient bydlí.

5.3.5. Srovnání s modelem lineární regrese

V této části ukážeme, že lineární regrese není vhodná pro modelování parametru PD. Vezměme data, která jsme použili k odhadu modelu binární logistické regrese a na jejich základě odhadneme model lineární regrese. Budeme uvažovat pouze proměnné, které jsme ve výše uvedeném odhadu použili, i když použití algoritmu Stepwise je vhodné i pro lineární regresi. Odhad parametrů je v tabulce 13.

Ověřování předpokladů není cílem textu, a proto jej vynecháme. Výsledky jsou výrazně odlišné vzhledem k jinému typu modelu. Model lineární regrese má oproti modelu binární logistické regrese spoustu příjemných vlastností. Samotný odhad je jednodušší. Není nutné řešit soustavu rovnic, lze použít metodu nejmenších čtverců, na kterou není potřeba žádný speciální software. V případě použití programu Mathematica trvá odhad modelu lineární regrese cca 6 sekund, zatímco odhad modelu binární logistické regrese více než půl minuty. Odhad modelu binární logistické regrese je i paměťově náročnější než odhad modelu lineární regrese. Model lineární regrese má navíc srovnatelné diverzifikační schopnosti jako model binární logistické regrese, plocha pod ROC křivkou modelu lineární regrese na testovacích datech je v tomto případě 0,86535. Při použití jiných dat by byla tato plocha výrazně menší. ROC křivku neuvedeme.

Prom.	Odhad
Konst.	0,02801
X_2	0,01520
X_6	$-1{,}87932 \cdot 10^{-8}$
X_7	$-0{,}00011$
X_8	$-0{,}00021$
X_9	$-0{,}00054$
X_{10}	$-2{,}93163 \cdot 10^{-8}$
X_{12}	$5{,}74051 \cdot 10^{-6}$
X_{13}	0,01137
X_{14}	$4{,}23621 \cdot 10^{-7}$
X_{15}	$2{,}86847 \cdot 10^{-8}$
X_{18}	$-0{,}00027$
X_{20}	$-0{,}00240$
X_{21}	$-0{,}01703$

Tabulka 13 – odhad modelu lineární regrese

Z výše uvedeného je tedy model lineární regrese lepší než model binární logistické regrese. Nicméně pro modelování parametru PD je nevhodný, protože pomocí modelu lineární regrese můžeme získat předpovědi, které jsou záporné (v tomto případě 32 663 pozorování) nebo větší než jedna (v tomto případě žádné pozorování),

což je nesmysl vzhledem k tomu, že modelujeme pravděpodobnost, která je vždy v intervalu $\langle 0; 1 \rangle$, zatímco předpovědi pomocí modelu binární logistické regrese budou v tomto intervalu ze vzorce (3.2) nebo (5.1) (viz výše). Model lineární regrese je tedy bez dalších úprav nevhodný.

5.3.6. Kategorizace a faktorizace dat

Všechny modely binární logistické regrese vyšli jako špatné. Důvodem je nevhodné použití dat. Do modelu jsme dosadili spojité proměnné bez další úpravy. To znamená, že jsme neprovedli žádnou transformaci poměrů. V praxi se však nějaká transformace poměrů provést musí. Zpravidla se nejedná o logaritmické transformace, ale o kategorizaci dané proměnné a následný převod na dummy proměnné. Touto transformací docílíme monotonie vztahu mezi odpověďmi a danou proměnnou a též odhad parametrů bude lépe interpretovatelný.

Nejprve popíšeme samotnou faktorizaci dat. Důvodem pro tuto úpravu jsou poznatky z praxe. Ukazuje se, že je nevhodné dosazovat do modelu spojité proměnné a kategoriální proměnné s více než jednou kategorií bez dalších úprav. Předpokládejme, že máme model s jednou proměnnou, kterou je měsíční plat klienta, a pro jednoduchost předpokládejme, že daný model nemá konstantní člen. Odhadujeme tedy

$$\tilde{\pi}(x) = \frac{1}{1 + e^{-\beta x}}.$$

Nechť

$$\hat{\beta} < 0.$$

To však znamená, že se záporná znaménka vyruší, a tedy čím větší bude x, tím menší bude $\tilde{\pi}(x)$. Pokud by však klient prohlásil, že jeho plat je jeden milion, tj.

$$x = 1\,000\,000,$$

pak je $\tilde{\pi}(x)$ téměř nulové a tedy takový klient skoro jistě splatí úvěr, o který žádá. To však není zcela dobré, protože se může stát, že se plat klientovi naráz sníží, nebo je možné, že klient lhal. V takovém případě je však špatně předpovězena pravděpodobnost selhání. Problém spočívá v dosazení spojité proměnné do modelu bez další úpravy, protože je možné, že předpokládaná monotonní, respektive lineární, závislost mezi proměnnou plat a odezvami neplatí. Takové závislosti se docílí nejprve kategorizací spojitých proměnných. U spojité proměnné se nejprve nakreslí histogram z dat a na jeho základě se zvolí počet kategorií a jejich rozsahy. Na základě této volby se pak data do zvolených kategorií rozdělí. Volba počtu a rozsahu kategorií je zcela subjektivní. Nyní tedy máme všechny proměnné kategoriální.

Následuje převod všech kategoriálních proměnných na dummy proměnné neboli faktorizace proměnných. Dummy proměnné jsou proměnné, které nabývají pouze nulových a jedničkových hodnot. Jedná se de facto o indikátory nějakého jevu. Kategoriální proměnné s dvěma kategoriemi na dummy proměnné není nutné převádět, protože jsou již ze svého charakteru dummy proměnnými, ale je nutné je překódovat na nula – jedničkové proměnné. Pro jednodušší vysvětlení předpokládejme, že chceme faktorizovat proměnnu nejvyšší dosažené vzdělání, která má čtyři kategorie: základní vzdělání (nadále ZŠ), výuční list (nadále VL), středoškolské vzdělání s maturitou (nadále SŠ) a vysokoškolské vzdělání (nadále VŠ). Vzhledem k tomu, že proměnná má čtyři kategorie, budeme potřebovat čtyři dummy proměnné (pokud by měla tři kategorie, potřebujeme tři dummy proměnné a podobně). Příslušnou dummy proměnnou sestavíme jako indikátor dané kategorie v kategoriální veličině, tedy první dummy proměnná bude indikátorem ZŠ, druhá bude indikátorem VL, třetí proměnná bude indikátorem SŠ a čtvrtá proměnná bude indikátor VŠ. Tím je faktorizace proměnných ukončena a tyto proměnné vstoupí do modelu. Nemohou tam však vstoupit všechny, protože jejich součet je jednotkový vektor, který však do modelu vstoupil skrze konstantní člen. Znamenalo by to tedy vstup jedné proměnné vícekrát do modelu a model by pak nebylo možné odhadnout. Z toho důvodu se musí vybrat jedna dummy proměnná, která do modelu nevstoupí. Tato volba je opět subjektivní. Znamená to tedy, že se tímto způsobem převede kategoriální proměnná o L kategoriích na systém L – 1 dummy proměnných. Z toho důvodu není nutné faktorizovat proměnné se dvěma kategoriemi a stačí je překódovat. Pro interpretaci je však nutné vědět, co značí nula a co značí jednička.

Převod na dummy proměnné též zjednodušuje interpretaci odhadu parametrů, protože interpretace odhadu při použití spojitých proměnných není zcela srozumitelná. Interpretaci ilustrujeme na proměnné vzdělání definované výše. Předpokládejme, že jsme do modelu nezařadili proměnnou, která indikuje ZŠ, a ostatní proměnné jsme do modelu zařadili. Nezařazenou proměnnou občas nazýváme základní proměnnou. V tom případě odhadneme tři parametry, z nichž β_2 náleží k proměnné, která indikuje VL, β_3 náleží k proměnné indikující SŠ a β_4 náleží k proměnné, která je indikátorem VŠ. Interpretace pak souvisí s dummy proměnnou, kterou jsme do modelu nezařadili, tedy v tomto případě se ZŠ. Například parametr β_2 vyjadřuje efekt toho, že daný klient má VL, oproti tomu, že by měl jen ZŠ. Stejná interpretace je i v případě ostatních parametrů. Na základě stejné ideje se interpretují i odhady v případě jiných proměnných. Jiná možnost spočívá v použití fixních hodnot všech proměnných vyjma jedné, kterou chceme interpretovat, a dummy proměnné příslušné této proměnné postupně měnit a sledovat při tom výsledné odhady pravděpodobností. Dané pravděpodobnosti se pak navzájem porovnávají a na základě tohoto porovnání se posuzují různí klienti s danými vlastnostmi.

5.3.7. Odhad modelu při použití dummy proměnných

Použijeme proměnné, které jsme vybrali výše. Proměnná X_2 značí počet měsíců posledního roku, ve kterých klient měl nějaký dluh na účtu. Vzhledem k histogramu ji kategorizujeme na tři kategorie: 1. kategorie značí žádný měsíc, 2. kategorie značí jeden až tři měsíce, 3. kategorie značí zbytek. Převod na dummy proměnné je zřejmý. Ve třetí proměnné je oproti ostatním dvěma velmi málo jedničkových pozorování, a proto tuto proměnnou použijeme jako základní a do modelu ji nezařadíme. Tedy do modelu zařadíme první dvě proměnné, nadále je označme $X_{2,1}$ a $X_{2,2}$. Proměnná X_6 značí platbu na stavební spoření v aktuálním měsíci. Kategorizujeme ji do dvou kategorií: 1. kategorie značí částku, která je menší než jedna, druhá kategorie značí alespoň jednotkovou částku, na základě těchto kategorií vytvoříme dvě dummy proměnné. Vzhledem k tomu, že první dummy proměnná má více jednotkových pozorování, použijeme první proměnnou, nedále značíme $X_{6,1}$. Proměnnou X_7, která značí počet trvalých příkazů za posledních 6 měsíců, kategorizujeme do tří kategorií: 1. kategorie značí méně než pět příkazů, 2. kategorie značí alespoň pět příkazů a méně než 20 příkazů, 3. kategorie značí alespoň 20 příkazů, tedy vytvoříme tři dummy proměnné. V modelu použijeme první dvě proměnné, nadále značíme $X_{7,1}$ a $X_{7,2}$. Proměnnou X_8 (počet čerpání přes internetové bankovnictví za poslední 3 měsíce) kategorizujeme do tří kategorií:

1. kategorie = méně než pět čerpání,

2. kategorie = 5 až 25 čerpání,

3. kategorie = 25 čerpání a více.

Máme tedy tři dummy proměnné a v modelu použijeme první dvě, nadále $X_{8,1}$ a $X_{8,2}$. Proměnnou X_9 značící průměrný měsíční počet transakcí platební kartou v obchodech za poslední tři měsíce kategorizujeme do čtyř kategorií: 1. kategorie značí méně než jedno čerpání, 2. kategorie značí 1 až 5 čerpání, 3. kategorie značí 5 až 10 čerpání a 4. kategorie značí 10 čerpání a více, tedy máme čtyři dummy proměnné, přičemž v modelu použijeme první tři, nadále $X_{9,1}$, $X_{9,2}$ a $X_{9,3}$. Proměnnou X_{10} (minimální měsíční vklad na účet během posledních šesti měsíců) kategorizujeme do tří kategorií: 1. kategorie znamená vklad méně než 10 000, druhá kategorie znamená vklad, který je alespoň 10 000 a menší než 25 000, 3. kategorie značí ostatní výše vkladů, tedy máme tři dummy proměnné, přičemž v modelu použijeme první dvě kategorie, nadále značíme $X_{10,1}$ a $X_{10,2}$.

Proměnnou X_{12} (maximální dluh na OVD za poslední rok) kategorizujeme do dvou kategorií: 1. kategorie značí dluh, který je menší než jedna, 2. kategorie značí ostatní případy, tedy máme dvě dummy proměnné a v modelu použijeme první proměnnou, nadále značíme $X_{12,1}$. Proměnnou X_{13} (počet delikvencí na OVD za poslední 3 měsíce) kategorizujeme do dvou kategorií: 1. kategorie značí žádnou delikvenci, 2. kategorie značí kladný počet delikvencí, tedy máme dvě dummy proměnné a

v modelu použijeme první proměnnou (nadále $X_{13,1}$). Proměnnou X_{14} (vyčerpaná částka na OVD ke konci aktuálního měsíce) kategorizujeme do tří kategorií: 1. kategorie značí částku menší než 10, 2. kategorie značí částku alespoň 10 a menší než 15 000 a 3. kategorie značí ostatní částky, tedy máme tři dummy proměnné a v modelu použijeme první dvě proměnné (nadále $X_{14,1}$ a $X_{14,2}$). Proměnnou X_{15} (výše úvěrového limitu na OVD v předchozím měsíci) kategorizujeme do tří kategorií: 1. kategorie značí limit menší než 10 000, 2. kategorie značí limit alespoň 10 000 a menší než 20 000 a 3. kategorie značí ostatní limity, tedy máme tři dummy proměnné a v modelu použijeme druhou a třetí proměnnou (nadále $X_{15,2}$ a $X_{15,3}$). Proměnnou X_{18} (počet debetních operací na OVD účtu v aktuálním měsíci) kategorizujeme do čtyř kategorií: 1. kategorie značí méně než 5 operací, 2. kategorie značí alespoň 5 operací a méně než 10 operací, 3. kategorie značí alespoň 10 operací a méně než 20 operací a čtvrtá kategorie značí alespoň 20 operací, tedy máme čtyři dummy proměnné a v modelu použijeme první tři proměnné (nadále $X_{18,1}$, $X_{18,2}$ a $X_{18,3}$).

Proměnné X_{20} (pohlaví) a X_{21} (indikace trvalého pobytu na území ČR) není nutné kategorizovat, protože mají dvě kategorie. Do modelu je tedy dosadíme bez úprav. Je nutné mít na paměti, že v proměnné X_{20} jednička značí ženy a v proměnné X_{21} jednička značí, že klient má trvalý pobyt v ČR. Výše popsaná kategorizace a výběr proměnných vždy probíhal na základě počtu dat v daných kategoriích. Vyřazená dummy proměnná obsahuje nejméně jedničkových pozorování ze všech.

Na takto upravených datech odhadneme model. Výsledky uvádí tabulka 14. Výsledky některých p – hodnot nezamítají hypotézu nulovosti. Proměnné byly vybrané na základě modelu se spojitými proměnnými. Navíc kategorizace a faktorizace dat a výběr dummy proměnných, které jsme zařadili do modelu, nemusí být optimální. Přesto tento výběr ponecháme jako směrodatný, protože vysoká p – hodnota nezamítá hypotézu nulovosti, ale to neznamená, že daný parametr je skutečně nulový.

Provedeme test dobré shody. Hosmer – Lemeshowova statistika na datech pro test při rozdělení na 10 skupin je

$$C_d = 12{,}9819,$$

tedy p – hodnota testu je

$$p = 0{,}11248,$$

tedy na hladině 0,05 při dělení na 10 skupin nezamítáme hypotézu, která tvrdí, že model je dobrý. Při dělení na sto skupin máme

$$C_d^{(100)} = 90{,}5312,$$

příslušný kvantil chí – kvadrát o 98 stupních volnosti je 122,108 a p – hodnota je

$$p^{(100)} = 0{,}69137,$$

tedy model lze považovat za dobrý na hladině 0,05. McFaddenův koeficient determinace je

$$R_d^2 = 0{,}2231,$$

tedy

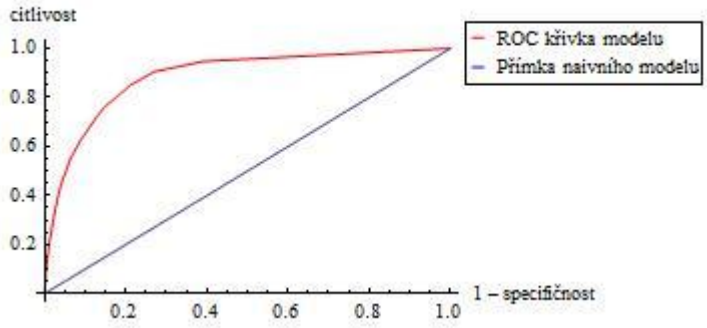
$$R_d^2 > R^2.$$

Prom.	Odhad	P - hodnota	Interval spolehlivosti	
			Dolní mez	Horní mez
Konst.	$-3{,}3982$	$1{,}3107 \cdot 10^{-27}$	$-4{,}0099$	$-2{,}7865$
$X_{2,1}$	$-1{,}5651$	$9{,}9544 \cdot 10^{-50}$	$-1{,}772$	$-1{,}358$
$X_{2,2}$	$-0{,}9732$	$6{,}7242 \cdot 10^{-34}$	$-1{,}1304$	$-0{,}816$
$X_{6,1}$	$0{,}43131$	$0{,}0004$	$0{,}1945$	$0{,}668$
$X_{7,1}$	$0{,}4696$	$0{,}0013$	$0{,}1834$	$0{,}7558$
$X_{7,2}$	$0{,}3209$	$0{,}02532$	$0{,}03967$	$0{,}6021$
$X_{8,1}$	$0{,}5267$	$0{,}002$	$0{,}192$	$0{,}8616$
$X_{8,2}$	$-0{,}1323$	$0{,}4501$	$-0{,}4756$	$0{,}211$
$X_{9,1}$	$1{,}1787$	$2{,}0433 \cdot 10^{-8}$	$0{,}7668$	$1{,}5906$
$X_{9,2}$	$0{,}7023$	$0{,}0009$	$0{,}2862$	$1{,}1185$
$X_{9,3}$	$0{,}3875$	$0{,}0964$	$-0{,}0693$	$0{,}8443$
$X_{10,1}$	$0{,}7914$	$6{,}3739 \cdot 10^{-7}$	$0{,}4799$	$1{,}1029$
$X_{10,2}$	$0{,}2316$	$0{,}144$	$-0{,}079$	$0{,}5424$
$X_{12,1}$	$-0{,}7936$	$2{,}2137 \cdot 10^{-11}$	$-1{,}026$	$-0{,}561$
$X_{13,1}$	$-0{,}10697$	$0{,}2269$	$-0{,}2805$	$0{,}0665$
$X_{14,1}$	$-1{,}9576$	$1{,}6342 \cdot 10^{-34}$	$-2{,}2708$	$-1{,}6445$
$X_{14,2}$	$-0{,}5183$	$1{,}2613 \cdot 10^{-7}$	$-0{,}7106$	$-0{,}3261$
$X_{15,2}$	$0{,}1431$	$0{,}1242$	$-0{,}0394$	$0{,}3256$
$X_{15,3}$	$-0{,}1301$	$0{,}3112$	$-0{,}3818$	$0{,}1217$
$X_{18,1}$	$-0{,}1279$	$0{,}2522$	$-0{,}3469$	$0{,}0911$
$X_{18,2}$	$-0{,}5515$	$2{,}8201 \cdot 10^{-7}$	$-0{,}762$	$-0{,}341$
$X_{18,3}$	$-0{,}4396$	$5{,}1053 \cdot 10^{-6}$	$-0{,}6285$	$-0{,}2507$
X_{20}	$-0{,}1761$	$0{,}0069$	$-0{,}3039$	$-0{,}0483$
X_{21}	$-0{,}4992$	$7{,}4224 \cdot 10^{-8}$	$-0{,}681$	$-0{,}3174$

Tabulka 14 – odhad modelu binární logistické regrese při použití dummy proměnných

To znamená, že při použití dummy proměnných dostaneme lepší model než při dosazení spojitých proměnných bez další úpravy. Navíc při použití dummy proměnných

dostaneme model, který lze považovat za dobrý, zatímco při použití spojitých proměnných dostáváme špatný model.



Obr. 10 – ROC křivka modelu při použití dummy proměnných

Plocha pod ROC křivkou je 0,89418. ROC křivka ve srovnání s přímkou naivního modelu je na obr. 10. Dále

$$\mathrm{GINI} = 0{,}788779;\ \mathrm{K-S} = 0{,}557449;\ \mathrm{CS} = 0{,}894176.$$

Model při použití dummy proměnných má lepší diverzifikační schopnosti.

Při použití dummy proměnných též odhadneme model lineární regrese. Výsledky jsou v tabulce 15. Výsledky všech ukazatelů jsou podobné jako při použití spojitých proměnných, tedy plocha pod ROC křivkou je 0,8829 a celkem dostáváme 44 693 záporných předpovědí a žádnou předpověď větší než jedna. Model je tedy nevhodný na modelování pravděpodobností selhání.

Odhadneme výše navržené alternativní modely při použití dummy proměnných. Odhad první alternativy (indikace pojištění) je v tabulce 16 v prvním sloupci, přičemž proměnná N_1 je přidaná proměnná. Faktorizace v případě této proměnné není nutná, protože tato proměnná je nula – jedničková, a tedy dummy proměnná. Odhad třetího alternativního modelu (odchozí transakce) je v téže tabulce ve třetím sloupci. Kategorizace dat není nutná, protože daná proměnná je kategoriální se čtyřmi kategoriemi. Faktorizace je zřejmá, přičemž použijeme druhou, třetí a čtvrtou dummy proměnnou (značení postupně N_1, N_2 a N_3). Čtvrtá alternativa (model bez konstantního členu) je v posledním sloupci v tabulce 16. Proškrtlé kolonky značí, že daná proměnná v modelu není. Druhá alternativa (indikace delikvencí) nemá smysl vzhledem k faktorizaci původních dat. Indikátor delikvence sice neodpovídá žádné dummy proměnné v modelu, ale proměnná $X_{13,1}$ je doplňkem indikátoru delikvence. Pro všechny alternativy provedeme test dobré shody. P – hodnoty při rozdělení na deset skupin vychází postupně

$$p_1 = 0{,}207, p_3 = 0{,}005, p_4 = 0{,}0027,$$

tedy první alternativu lze považovat za dobrou na hladině 0,05. Pro ostatní dvě alternativy provedeme stejný test při dělení na sto skupin a dostaneme:

$$p_3^{(100)} = 0{,}26, p_4^{(100)} = 0{,}006,$$

tedy třetí alternativu lze považovat za dobrou na hladině 0,05, ale čtvrtou alternativu považujeme za špatnou na hladině 0,05. Čtvrtá alternativa je však model bez konstantního členu, tedy v případě tohoto modelu lze předpokládat, že bude špatný. Dále

$$R_{d,1}^2 = 0{,}2233, R_{d,3}^2 = 0{,}216, R_{d,4}^2 = 0{,}209,$$

tedy první alternativa je mírně lepší než původní model, zatímco ostatní dvě alternativy jsou horší než původní model. Nejhorší je model bez konstantního členu. Plocha pod ROC křivkou je u všech modelů přibližně rovna 0,89. ROC křivky modelů jsou stejné jako ROC křivka původního modelu. Všechny modely tedy lze považovat za modely s perfektní diverzifikací.

Prom.	Odhad
Konst.	0,12918
$X_{2,1}$	−0,0946
$X_{2,2}$	−0,0807
$X_{6,1}$	0,00121
$X_{7,1}$	0,0047
$X_{7,2}$	−0,0006
$X_{8,1}$	0,0074
$X_{8,2}$	−0,0018
$X_{9,1}$	0,0105
$X_{9,2}$	0,0005
$X_{9,3}$	−0,0003
$X_{10,1}$	0,00898
$X_{10,2}$	−0,002
$X_{12,1}$	−0,0062
$X_{13,1}$	−0,0112
$X_{14,1}$	−0,0156
$X_{14,2}$	−0,0082
$X_{15,2}$	0,0058
$X_{15,3}$	0,0045
$X_{18,1}$	0,0042
$X_{18,2}$	−0,0034
$X_{18,3}$	−0,0039
X_{20}	−0,0028
X_{21}	−0,0145

Tabulka 15 – odhad modelu lineární regrese při použití dummy proměnných

Prom.	První alternativa	Třetí alternativa	Čtvrtá alternativa
Konst.	$-3{,}2764$	$-1{,}3892$	-----
$X_{2,1}$	$-1{,}5635$	$-1{,}6394$	$-1{,}6931$
$X_{2,2}$	$-0{,}9725$	$-1{,}0356$	$-1{,}1211$
$X_{6,1}$	$0{,}4243$	$0{,}47$	$0{,}1059$
$X_{7,1}$	$0{,}3801$	-----	$0{,}2073$
$X_{7,2}$	$0{,}2879$	-----	$0{,}0742$
$X_{8,1}$	$0{,}4954$	-----	$0{,}0831$
$X_{8,2}$	$-0{,}1506$	-----	$-0{,}6096$
$X_{9,1}$	$1{,}1823$	-----	$0{,}1867$
$X_{9,2}$	$0{,}7047$	-----	$-0{,}3154$
$X_{9,3}$	$0{,}3895$	-----	$-0{,}6427$
$X_{10,1}$	$0{,}7893$	$1{,}0142$	$0{,}4898$
$X_{10,2}$	$0{,}2339$	$0{,}4029$	$-0{,}1919$
$X_{12,1}$	$-0{,}7914$	$-0{,}7893$	$-0{,}7897$
$X_{13,1}$	$-0{,}1038$	$-0{,}1616$	$-0{,}1745$
$X_{14,1}$	$-1{,}9489$	$-1{,}979$	$-2{,}3355$
$X_{14,2}$	$-0{,}5135$	$-0{,}535$	$-0{,}9179$
$X_{15,2}$	$0{,}1428$	$0{,}1562$	$-0{,}1924$
$X_{15,3}$	$-0{,}1289$	$-0{,}1647$	$-0{,}8279$
$X_{18,1}$	$-0{,}1344$	$0{,}0187$	$-0{,}0778$
$X_{18,2}$	$-0{,}5534$	$-0{,}3693$	$-0{,}5468$
$X_{18,3}$	$-0{,}4374$	$-0{,}341$	$-0{,}4901$
X_{20}	$-0{,}1748$	$-0{,}2085$	$-0{,}2832$
X_{21}	$-0{,}4921$	$-0{,}4843$	$-0{,}8835$
N_1	$-0{,}1451$	$-0{,}4077$	-----
N_2	-----	$-0{,}858$	-----
N_3	-----	$-1{,}2175$	-----

Tabulka 16 – odhad alternativních modelů při použití dummy proměnných

Je zjevné, že modely, které jsou v případě použití spojitých proměnných špatné, jsou správné při použití dummy proměnných. V praxi se mnohem častěji používají dummy proměnné namísto spojitých proměnných. Důvodem je jejich nepraktičnost daná porušením předpokladu linearity vztahu k odpovědím, dále též složitější interpretace odhadů parametrů. Tato ilustrace navíc ukazuje, že použití spojitých proměnných může vést ke špatným výsledkům. Modely při použití dummy proměnných jsou lepší ve všech směrech. Mají mírně lepší diverzifikační schopnosti, dle výsledků McFaddenova koeficientu determinace jsou lepší. Navíc s výjimkou modelu bez konstantního členu jsou dle výsledků testu dobré shody správné.

5.3.8. Kalibrace modelu

Označme nejprve

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

jako průměr pozorovaných odezev, přičemž n je počet pozorování, a

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$$

jako průměr všech odpovědí modelu. Při odhadu modelu platí:

$$\bar{y} = \bar{\hat{y}}.$$

Tohle však platí pouze při použití dat pro odhad, ale při použití jiných dat (například dat pro test) již tohle neplatí. Nicméně snahou je, aby tato rovnost platila alespoň přibližně. Pokud je však jeden z průměrů výrazně větší, je nutno přistoupit ke kalibraci modelu.

Vezměme funkci g z kapitoly 3, respektive její odhad, a dosaďme do ní pozorované hodnoty, které použijeme pro kalibraci. Označme takto získané hodnoty jako $\hat{x}_i$. K těmto hodnotám přísluší hodnoty odezev y_i. Výsledný odhad pravděpodobnosti selhání pak je

$$\pi(\mathbf{x}_i) = \frac{e^{a+b\,\hat{x}_i}}{1 + e^{a+b\,\hat{x}_i}},$$

přičemž hodnoty a a b se odhadnou pomocí metody maximální věrohodnosti. Odhadujeme tedy nový regresní model s jednou proměnnou vycházející z původního odhadu a konstantním členem. Pro model s $\tilde{\pi}(\mathbf{x})$ probíhá kalibrace stejně, přičemž v čitateli je 1.

Provedeme kalibraci pro hlavní model se spojitými proměnnými a hlavní model s dummy proměnnými, pro model s $\tilde{\pi}(\mathbf{x})$, alternativní modely kalibrovat nebudeme. Jako data pro kalibraci použijeme data pro test. Nejprve provedeme kalibraci modelu se spojitými proměnnými. V tomto případě jednotlivé průměry vychází

$$\bar{y} = 0{,}01541$$

a

$$\bar{\hat{y}} = 0{,}01637$$

I přesto, že oba průměry jsou téměř stejné, provedeme kalibraci. Odhady příslušných parametrů uvádí tabulka 17, druhý sloupec. P – hodnota testu dobré shody pro takto kalibrovaný model je

$$p_s = 3{,}63516 \cdot 10^{-9}$$

tedy model je i po kalibraci stále špatný. V případě modelu s dummy proměnnými máme

$$\bar{\hat{y}} = 0{,}0162$$

I zde provedeme kalibraci, příslušné odhady jsou v tabulce 17, třetí sloupec. P – hodnota testu dobré shody pro takto kalibrovaný model je

$$p_d = 0{,}24216$$

tedy tento model je ještě lepší, než původní model. K testování jsme použili data pro test, tedy stejná data, která jsme použili pro odhad kalibrovaného modelu, a všechny testy jsme provedli při dělení na deset skupin.

Parametr	Model se spojitými proměnnými	Model s dummy proměnnými
a	$-0{,}16589$	$-0{,}08419$
b	$0{,}96777$	$0{,}99011$

Tabulka 17 – odhady proměnných při kalibraci

Poznámka: Při kalibraci jsme upravili modely a následně jsme pro ně nespočítali žádné ukazatele diverzifikačních schopností, přičemž může existovat obava z jejich narušení, respektive zhoršení. Tato obava je však zbytečná, neboť takováto kalibrace nezmění Giniho koeficient, tedy Giniho koeficient zkalibrovaných modelů je stejný, jako u modelů výše, a tedy diverzifikační vlastnosti modelů jsou stejné.

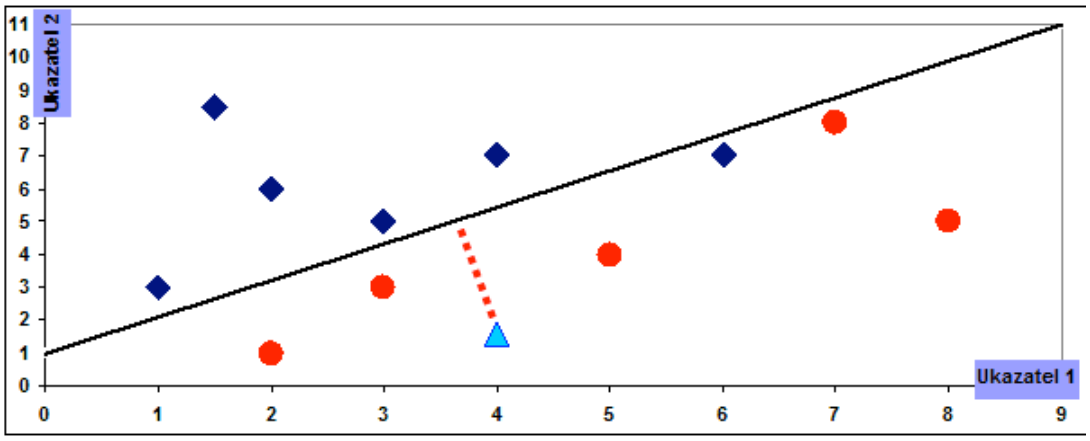
5.4. Další příklady modelů

5.4.1. Altmanův model

V této části využíváme informace z [5], [9] a [11].

Altmanův model, občas nazývaný Altmanovo z – skóre nebo Altmanova formule bankrotu, je dle [11] založen na diskriminační analýze. V diskriminační analýze se body několika různých tříd, v našem případě dvou, podniky s problémy a prosperující podniky, zakreslí do grafu, například obr. 11 (viz [11]), a nalezne se křivka nebo přímka, zvaná též diskriminační křivka, nebo též nadrovina popřípadě křivá plocha, která tyto třídy co nejlépe odděluje (na obr. 11 přímka, černě). Nové pozorování (na obr. 11 modrý trojúhelník) se zařazuje do třídy, které leží na příslušné straně od diskriminační křivky. Spolehlivost zařazení se pak posuzuje podle vzdálenosti od diskriminační křivky (na obr. 11 červeně tečkovaně). Občas se kolem diskriminační

křivky konstruuje pás spolehlivosti, zvaný též šedá zóna, který nedodává spolehlivé výsledky.



Obr. 11 – příklad grafu pro diskriminační analýzu

Obecný zápis Altmanova modelu je dle [11]:

$$Z = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6,$$

kde $a_i, i = 1; \dots; 6$ jsou diskriminační koeficienty a $X_i, i = 1; \dots; 6$ jsou poměrové ukazatele:

$$X_1 = \frac{\text{pracovní kapitál(oběžná aktiva} - \text{celkové krátkodobé dluhy)}}{\text{celková aktiva}},$$

$$X_2 = \frac{\text{zisk po zdanění} + \text{nerozdělený zisk minulých let}}{\text{celková aktiva}},$$

$$X_3 = \frac{\text{EBIT (zisk před odečtením úroků a daní, provozní hospodářský výsledek)}}{\text{celková aktiva}},$$

$$X_4 = \frac{\text{tržní hodnota akciového kapitálu}}{\text{celkové cizí zdroje}},$$

$$X_5 = \frac{\text{tržby}}{\text{celková aktiva}},$$

$$X_6 = \frac{\text{závazky po lhůtě spolatnosti}}{\text{tržby}}.$$

Altmanův model sestavil E. Altman v roce 1968 pro výrobní podniky s veřejně obchodovanými akciemi v USA, který dle [5] je

$$Z_1 = 1{,}2\,X_1 + 1{,}4\,X_2 + 3{,}3\,X_3 + 0{,}6\,X_4 + 0{,}999\,X_5, \text{přičemž } a_6 = 0. \quad (5.2)$$

Odsud lze spočítat pro danou firmu Altmanovo z – skóre. Dle [11], je – li $Z_1 > 2{,}99$, pak se společnost pohybuje v bezpečné zóně, je finančně silná a v následujícím roce nezbankrotuje, pokud je $Z_1 < 1{,}80$, je společnost v krizové zóně, má značné finanční potíže a v následujícím roce hrozí bankrot, a je – li $Z_1 \in \langle 1{,}81; 2{,}98 \rangle$, jedná se o šedou zónu, společnost může mít drobné finanční potíže a o jejím následujícím vývoji nelze nic říci. Dle [11] je nutné pro ekonomiku České republiky přidat do modelu ukazatel X_6, protože platební neschopnost má vliv na hospodaření společnosti. Altmanův model pro Českou republiku je dle [11]:

$$Z_{1_CZ} = 1{,}2\,X_1 + 1{,}4\,X_2 + 3{,}3\,X_3 + 0{,}6\,X_4 + 1{,}0\,X_5 + 1{,}0\,X_6,$$

přičemž interpretace výsledků je stejná jako v případě výsledků Z_1.

Praxe v průběhu let ukázala takový model jako nedostatečný, protože ho lze použít pouze pro akciové společnosti. Nebylo možné ho použít pro neakciové společnosti. Proto jej sám Altman v roce 1983 revidoval. Revize spočívala v celkovém přecenění modelu. V ukazateli X_4 se tržní hodnota akciového kapitálu nahradila vlastním jměním. Revidovaný model pak dle [11] je:

$$Z_2 = 0{,}717\,X_1 + 0{,}847\,X_2 + 3{,}107\,X_3 + 0{,}420\,X_4 + 0{,}998\,X_5, \text{přičemž } a_6 = 0,$$

přičemž se upravila i pásma klasifikace (dle [11]):

$$Z_2 < 1{,}23 \rightarrow \text{ krizová zóna,}$$

$$Z_2 > 2{,}90 \rightarrow \text{ bezpečná zóna,}$$

$$Z_2 \in \langle 1{,}23; 2{,}90 \rangle \rightarrow \text{ šedá zóna.}$$

Další revize proběhla v roce 1995. Takto revidovaný model je použitelný i pro nevýrobní podniky. Při této revizi se vypustil parametr X_5, vliv parametru X_6 je nadále nulový, parametr X_4 je ponechán ve verzi z roku 1983. Odhad pak dle [11] je:

$$Z_3 = 6{,}56\,X_1 + 3{,}26\,X_2 + 6{,}72\,X_3 + 1{,}05\,X_4$$

a pásma klasifikace (dle [11]):

$$Z_3 < 1{,}10 \rightarrow \text{ krizová zóna,}$$

$$Z_3 > 2{,}60 \rightarrow \text{ bezpečná zóna,}$$

$$Z_3 \in \langle 1{,}10; 2{,}60 \rangle \rightarrow \text{ šedá zóna.}$$

Altmanův model má spoustu příjemných vlastností, které jsou výhodné při jeho používání. Je již dávno odhadnutý, tedy při používání stačí pouze dosadit konkrétní data. Spolehlivost modelu je velmi vysoká. Model spěšně předpovídá bankrot v horizontu dvou let, v delším časovém horizontu je již úspěšnost menší. Navíc sběr

dat je snadný, protože všechny hodnoty jsou součástí účetních výkazů. Altmanův model má však též několik nevýhod. Jednou z nich je právě jeho stáří a fakt, že byl odhadnutý na základě dat z 60. let 20. století. Dnes by pravděpodobně vypadal jinak. Altmanův model lze aplikovat pouze na firmy, ne na fyzické osoby, pro které bychom rovněž rádi znali podobnou předpověď. Navíc Altmanův model pouze předpoví, zda firma zbankrotuje nebo nezbankrotuje, ale neukáže pravděpodobnost selhání, kterou je nutné znát pro potřeby RWA funkce. Altmanův model pro ekonomiku České republiky navíc nelze otestovat z důvodu malého množství dat.

5.4.2. Srovnání Altmanova modelu s modelem binární logistické regrese

Budeme srovnávat model logistické regrese (5.1) s původním Altmanovým modelem (5.2). Pro toto srovnání budeme do obou vzorců dosazovat stejné ukazatele získané z účetních výkazů. K tomuto srovnání je nejprve nutné odhadnout model binární logistické regrese. Ideální by bylo mít data, která měl v roce 1968 Altman. Ta jsou špatně dostupná, a proto použijeme simulaci, na jejímž základě odhadneme model binární logistické regrese. Dle [2] použil Altman skupinu 66 společností rovnoměrně rozdělených na bankrotující a nebankrotující. Průměrné hodnoty jednotlivých ukazatelů dle [5] znázorňuje tabulka 18.

K simulaci použijeme uvedené výsledky. Nejprve vygenerujeme 66 dat z alternativního rozdělení se střední hodnotou 0,5. Pak pro každou společnost vygenerujeme hodnoty všech pěti ukazatelů. Z definic všech ukazatelů je zřejmé, že ukazatelé X_1, X_2 a X_3 nabývají hodnot v intervalu $\langle -1; 1 \rangle$. Proto k jejich vygenerování použijeme rovnoměrné rozdělení, jehož střední hodnota bude rovna průměrným hodnotám z tabulky 18. Nosičem daného rozdělení bude vždy interval

$$(\text{průměrná hodnota} - \mathrm{a}; \text{průměrná hodnota} + \mathrm{a}),$$

přičemž

$$\mathrm{a} = 0{,}5 \text{ pro } X_1 \text{ a } X_3$$

a

$$\mathrm{a} = 0{,}3 \text{ pro } X_2.$$

Z definic dále plyne, že X_4 a X_5 nabývají kladných hodnot. Proto pro jejich vygenerování použijeme exponenciální rozdělení se střední hodnotou rovnou průměrným hodnotám. Odhad modelu s konstantním členem je

$$\pi = \frac{1}{1 + e^{-6{,}44598 + 9{,}8395\, X_1 + 54{,}9936\, X_2 + 1{,}90274\, X_3 + 10{,}2931\, X_4 + 1{,}39874\, X_5}}.$$

Odhad modelu bez konstantního členu je

$$\pi = \frac{1}{1 + e^{6,78648\,X_1 + 61,7669\,X_2 + 4,05515\,X_3 + 7,61374\,X_4 + 0,423075\,X_5}}.$$

Všechny tři modely nejprve srovnáme na datech o společnosti Škoda auto, a.s., z [12]. Tato společnost je akciová a výrobní, tedy pro ni použijeme vzorec (5.2), přičemž v proměnné X_4 použijeme vlastní jmění namísto tržní hodnoty akciového kapitálu. Výsledek z Altmanova modelu je 3,15149, tedy dle Altmanova modelu je společnost ekonomicky silná a nelze předpokládat bankrot. Pravděpodobnost selhání dle modelu s konstantním členem je $1,3488 \cdot 10^{-7}$ a dle modelu bez konstantního členu $1,74405 \cdot 10^{-8}$ a je zjevné, že firma skoro jistě nezbankrotuje. Časová i paměťová náročnost je u všech tří výpočtů zanedbatelná, tedy z hlediska výpočetní náročnosti jsou všechny tři modely stejné.

Ukazatel	Bankrotující podnik	Nebankrotující podnik
X_1	$-0,061$	0,414
X_2	$-0,626$	0,355
X_3	$-0,318$	0,154
X_4	0,401	2,477
X_5	1,500	1,900

Tabulka 18 – průměrné hodnoty ukazatelů

Dále srovnáme všechny modely na nasimulovaných datech. Budeme generovat proměnné vstupující do modelů, tedy X_1, X_2, X_3, X_4 a X_5. Ke generování proměnných X_1, X_2 a X_3 použijeme rovnoměrné rozdělení na intervalu $(-1; 1)$, pro proměnnou X_4 exponenciální rozdělení se střední hodnotou jedna a k proměnné X_5 exponenciální rozdělení se střední hodnotou dva. Takových hodnot vygenerujeme 1000 a na těchto hodnotách porovnáme všechny tři modely. Vezmeme výsledky z Altmanova modelu. Z nich vyplývá, že 398 simulovaných společností nezbankrotuje, 434 společností do jednoho roku zbankrotuje a o 168 společnostech nelze nic říci, protože jsou v šedé zóně. S těmito výsledky srovnáme výsledky z obou modelů binární logistické regrese.

Pro model s konstantním členem nejprve spočítáme odhady pravděpodobností selhání pro každou společnost. Z těchto odhadů vezmeme ty odhady, které odpovídají firmám, které dle Altmanova modelu nezbankrotují. Z těchto společností má 114 společností pravděpodobnost selhání vyšší než 0,5, 106 vyšší než 0,9, 96 vyšší než 0,99 a 90 vyšší než 0,999. Tedy z 398 společností, které dle Altmanova modelu nezbankrotují, je 90 společností, které dle modelu binární logistické regrese skoro jistě zbankrotují. Naopak ze společností, které dle Altmanova modelu zbankrotují, má 120 společností pravděpodobnost selhání nižší než 0,00001, tedy skoro jistě nezbankrotují. Obdobné výsledky dostaneme při použití modelu bez konstantního členu. Různost

výsledků je dána hlavně stavbou modelu, která probíhala na základě nasimulovaných dat, která neodpovídají skutečnosti. Navíc se v modelu používají spojité proměnné. O tomto přístupu je výše ukázáno, že není vhodný. Přesto je vhodné uvažovat výsledky z Altmanova modelu a porovnávat je s modely binární logistické regrese. Pokud bychom sestavili model binární logistické regrese z většího množství reálných dat, byla by správnost výsledků lepší a všechny modely by dávaly podstatně bližší výsledky.

5.4.3. KMV model

V této části postupujeme podle [7].

KMV jsou iniciály tvůrců modelu, kterými jsou Kealhofer, McQuown a Vašíček. Patří mezi strukturální prahové modely, tj. stav selhání se odvozuje od poklesu hodnoty aktiv pod určitou hodnotu. Používá alternativní rozdělení pro pravděpodobnost selhání. Je vhodnou aplikací lineární regrese na modelování pravděpodobnosti selhání. Pomocí tohoto modelu lze též řešit problém víceletého PD (viz dále).

Buď T horizont, ve kterém se předpovídají selhání klientů. Zpravidla klademe $T = 1$ rok. Uvažujeme dlužníka (firmu) i, i = 1;…;I. Označme A_t^i jako hodnotu aktiv i – tého dlužníka v čase t . Předpokládejme, že k selhání i – tého dlužníka dojde v časovém horizontu T, pokud $A_T^i < C_i$, kde C_i je předem daná částka, zpravidla výše dluhu, nebo výše splátek. Označme logaritmický výnos jako

$$r_i = \log\left(\frac{A_T^i}{A_0^i}\right).$$

Logaritmický výnos použijeme jako základní veličinu. Přirozeně platí:

$$P\left(A_T^i < C_i\right) = P\left(\frac{A_T^i}{A_0^i} < \frac{C_i}{A_0^i}\right) = P\left(r_i < \log\left(\frac{C_i}{A_0^i}\right)\right). \tag{5.3}$$

Namísto pravděpodobnosti selhání modelujeme logaritmický výnos za časový horizont T . Předpokládejme, že logaritmické výnosy klientů splňují

$$r_i = \beta_i \Phi_i + \varepsilon_i, i = 1; \dots; I, \tag{5.4}$$

kde Φ_i je kompozitní faktor dlužníka i, β_i je konstanta, která se též nazývá beta dlužníka i. Náhodné veličiny $\{\varepsilon_i\}$ jsou navzájem nezávislé a nezávislé na $\{\Phi_i\}$. Máme tedy regresní model. Platí:

$$\operatorname{var} r_i = \beta_i^2 \operatorname{var} \Phi_i + \operatorname{var} \varepsilon_i.$$

Koeficient určenosti regresní rovnice (5.4) je podle [7]:

$$R_i^2 = \frac{\beta_i^2 \operatorname{var} \Phi_i}{\operatorname{var} r_i},$$

přičemž $1 - R_i^2$ se interpretuje jako procentní míra dlužníka i.

Φ_i se rozkládá podle regionů a druhů podnikání:

$$\Phi_i = \sum_{k=1}^{K} w_{ik} \Psi_k \,, i = 1; \dots; I,$$

přičemž $\Psi_k, k = 1; \dots; K_0$, jsou indexy (ve smyslu ukazatel) pro obory podnikání, $\Psi_k, k = K_0 + 1; \dots; K$, jsou indexy pro regiony, w_{ik} jsou (zpravidla deterministicky určené) váhy. Ψ_k se dále vyjadřují pomocí globálních faktorů:

$$\Psi_k = \sum_{n=1}^{N} b_{kn} \Gamma_n + \delta_k, k = 1; \dots; K,$$

kde δ_k jsou vzájemně nezávislá rezidua, b_{kn} jsou beta oborů regionů podnikání, která se určující některou z metod hlavní komponenty, což jsou metody faktorové analýzy směřující k minimalizaci reziduí δ_k. Dosazením jednotlivých rozkladů do (5.4) dostáváme maticové zápisy

$$r = \beta W \Psi + \varepsilon,$$

$$r = \beta W (B\Gamma + \delta) + \varepsilon.$$

V obou případech tak máme regresní rovnici, ve které odhadujeme β . Původní model předpokládá normální rozdělení reziduí, v poslední době se však přihlíží k jejich empirickým distribučním funkcím.

Hlavním způsobem použití modelu je testování scénářů. Předpokládejme, že rozdělení globálních faktorů je vícerozměrné normální. Scénář pak může spočívat ve volbě jejich hodnot, nebo ve volbě realizací indexů, tedy

$$\Gamma_1 = \gamma_1; \dots; \Gamma_N = \gamma_N,$$

nebo

$$\Psi_1 = \psi_1; \dots; \Psi_K = \psi_K.$$

Budeme – li předpokládat, že náhodné veličiny $\{\varepsilon_i\}$ jsou rovněž normálně rozdělené, pak je rozdělení všech náhodných veličin v modelu normální a lze použít i různé formule. Bez jakéhokoli předpokladu na normalitu lze použít metody Monte Carlo.

Budeme – li předpokládat, že globální faktory mají vícerozměrné normální rozdělení, pak jsme schopni určit rozdělení kompozitních faktorů, a tedy i logaritmických výnosů. Pak již lze snadno určit hledanou pravděpodobnost, protože poslední pravděpodobnost v (5.3) je z definice hodnota distribuční funkce logaritmic-

kých výnosů v bodě $\log(C_i/A_0^i)$. Distribuční funkci pak lze určit pomocí pravidel pro určování rozdělení součtu. Za předpokladu, že globální faktory mají normální rozdělení, mají logaritmické výnosy rovněž normální rozdělení.

Výhodou modelu je použití lineární regrese, která je jednodušší než logistická regrese. Navíc s její pomocí získáme výsledky, které opravdu bude možné interpretovat jako pravděpodobnost. Nevýhodou modelu je jeho složitost a fakt, že jeho součástí jsou spekulace. Navíc samotné určování rozdělení součtu je časově náročné.

5.5. Problém víceletého PD

Parametr PD se zpravidla modeluje jeden rok. Bylo by vhodné mít nástroj, který modeluje parametr PD pro delší časový úsek. K tomu by se dal využít Kaplan – Mayerův odhad, ale jeho použití v současnosti není dostačující. Parametr PD pro více let se navíc nemodeluje. Namísto toho se parametr PD odhaduje každý rok zvlášť. V poslední době se však začínají objevovat hlasy, které vyžadují nějaké řešení tohoto problému.

Několik možností, řešící tento problém, je popsáno výše. Například Altmanův model dává spolehlivé výsledky ve dvouletém horizontu. I model binární logistické regrese může mít za určitých okolností predikční schopnost na více než jeden rok. Takový model by nesměl obsahovat informace o samotném úvěru nebo půjčce, neboť tyto informace se každým rokem mění. Každým rokem se může měnit částka, kterou ještě zbývá zaplatit, může se měnit výše úroků (myšleno části, která z dané splátky jde na zaplacení úroku), v extrémním případě se může měnit i samotná úroková míra. I informace získané každý rok z běžného účtu mohou být každý rok jiné. Takový model logistické regrese by tedy musel obsahovat pouze informace o klientovi (například obor podnikání, hodnotu kapitálu, sídlo nebo bydliště, informace, které získáme z účetnictví pro Altmanův model, v případě fyzických osob dále například věk, pohlaví, a podobně). Spolehlivost takového modelu by však byla pouze pro následující rok, nejvýše dva, pro delší časový horizont by spolehlivost klesala stejně jako v případě Altmanova modelu. KMV model tento problém řeší, protože není sestaven jen pro jednoletý horizont. Nevýhoda modelu spočívá v tom, že vychází z numerických simulací a testování scénářů. Používání tohoto modelu na delší časový horizont by tedy znamenalo nutnost dalších spekulací a scénářů. Tím by výrazně klesala spolehlivost modelu.

Všechny uvedené modely modelují pravděpodobnost selhání v nějakém budoucím okamžiku. Nechť T je okamžik selhání daného klienta (fyzické osoby či firmy), nechť je celočíselný a kladný, tedy T interpretujeme jako rok selhání klienta. Hledejme pravděpodobnost, že tento okamžik nastane do nějakého času t (včetně), tedy hledáme

$$P(T \leq t).$$

Selhání v čase T znamená, že klient neselhal v prvním roce, neselhal ve druhém roce, neselhal ani v (T – 1) – ním roce, ale selhal v T – tém roce. Tedy před selháním v roce T bylo T – 1 roků, ve kterých klient neselhal. Vezměme si tedy posloupnost délky T, ve které každý prvek bude značit, zda v daném roce k selhání došlo či nikoli. Zjevně prvních T – 1 prvků podá negativní odpověď a poslední prvek kladnou. Můžeme přejít ke geometrickému rozdělení. Geometrické rozdělení značí počet neúspěchů před prvním úspěchem v posloupnosti nezávislých pokusů. Jeho pravděpodobnostní předpis je

$$P(X = j) = p(1 - p)^j, j = 0; 1; 2; \dots, p \in (0; 1).$$

Položme

$$X = T - 1,$$

pak jsme přešli od hledání okamžiku selhání k hledání počtu let do selhání. Zjevně

$$T = X + 1.$$

Střední hodnota geometrického rozdělení je

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

a

$$ET = EX + 1 = \frac{1 - p}{p} + 1.$$

Známe – li EX, pak

$$p\,EX = 1 - p$$

a tedy parametr geometrického rozdělení je roven

$$p = \frac{1}{EX + 1}.$$

Tedy pokud bychom odhadli EX, což je průměrná doba do selhání, můžeme z toho odhadnout dané pravděpodobnosti pro náhodnou veličinu X. Pak již lze spočítat pravděpodobnost:

$$P(X = j) = P(T - 1 = j) = P(T = j + 1).$$

Hledaná pravděpodobnost pak je

$$P(T \leq t) = P(X + 1 \leq t) = P(X \leq t - 1),$$

což je distribuční funkce geometrického rozdělení, tedy

$$P(T \leq t) = \begin{cases} 1 - (1-p)^{\lfloor t-1 \rfloor + 1}, \text{pro } t - 1 \geq 0, \text{tedy } t \geq 1, \\ 0, \text{jinak.} \end{cases}$$

Doplňme, že $\lfloor a \rfloor$ značí dolní celou část čísla a . Tedy hledanou pravděpodobnost lze spočítat i v neceločíselném t. Spočítáním této pravděpodobnosti zjistíme pravděpodobnost selhání v nějakém konečném čase.

Tento model je však pouze teoretický a v praxi nepoužitelný. Jeho použití s sebou přináší několik problémů. První problém není tak zásadní. Spočívá v samotném modelování střední hodnoty EX, která je na modelování nejjednodušší. Na to je potřeba buď nějaký zobecněný lineární model, nebo je nutné použít odhad pomocí metody maximální věrohodnosti, popřípadě použít klasický výběrový průměr ze všech dob do selhání, přičemž v případech, kdy k selhání nedošlo, použít přímo splatnost daného úvěru. S tím též souvisí problém sběru dat. Další problém je fakt, že geometrické rozdělení modeluje v tento počet v nekonečném horizontu. Banku však většinou zajímá pravděpodobnost selhání pouze po dobu trvání úvěru nebo půjčky. Tento problém je možné vyřešit pomocí „useknutého“ geometrického rozdělení, o kterém pojednáme dále. Další problém je spekulace na výše zmíněnou nezávislost prvků dané posloupnosti. Od času, kdy došlo k selhání, již o žádné nezávislosti nemůžeme mluvit. Do roku, ve kterém došlo k selhání, by za jistých okolností mohla nastat nezávislost. V tom případě by muselo být chování klienta v jednotlivých letech nezávislé, což je jev v lidské společnosti nemožný. Tedy předpoklad nezávislosti je předpoklad, kvůli kterému tento model nelze použít v praxi. Dalším problémem je otázka spolehlivosti modelu. Ta v prvé řadě závisí na odhadu střední hodnoty. Lze získat z dat nestranný a konzistentní odhad střední hodnoty, ale z něj získaný odhad parametru p je konzistentní, ale není nestranný, protože dle Jensenovy nerovnosti nestrannost zachovávají pouze lineární funkce. Spolehlivost modelu též závisí na rozptylu náhodné veličiny X:

$$\operatorname{var} X = \frac{1-p}{p^2},$$

přičemž

$$\operatorname{var} T = \operatorname{var} X.$$

Tedy spolehlivost modelu závisí též na odhadu parametru p.

Buď M splatnost dané půjčky (v tomto případě máme na mysli smluvní splatnost půjčky, ne modelovanou splatnost z 2. kapitoly). Pak víme, že T může nabývat jakýchkoli kladných celočíselných hodnot, ale banku zajímají pouze hodnoty od 1 do M. Předpokládejme tedy, že T může nabývat pouze těchto hodnot. To znamená, že náhodná veličina X může nabývat hodnot od 0 do M – 1. Zúžíme – li její, původně geometrické, rozdělení na tyto hodnoty, dostáváme rozdělení

$$P(X = j) = \begin{cases} cp(1-p)^j, j = 0; 1; 2; \dots; M-1, p \in (0; 1), \\ 0, \text{jinak.} \end{cases}$$

Konstantu c dopočítáme z

$$\sum_{j=0}^{M-1} P(X = j) = 1.$$

Máme

$$\sum_{j=0}^{M-1} P(X = j) = \sum_{j=0}^{M-1} cp(1-p)^j = cp \sum_{j=0}^{M-1} (1-p)^j.$$

Vzhledem k tomu, že sčítáme konečnou geometrickou posloupnost s M sčítanci, dostáváme

$$\sum_{j=0}^{M-1} (1-p)^j = \frac{1-(1-p)^M}{1-(1-p)} = \frac{1-(1-p)^M}{p}.$$

Tedy

$$\sum_{j=0}^{M-1} P(X = j) = cp\frac{1-(1-p)^M}{p} = c[1-(1-p)^M].$$

Odsud

$$c = \frac{1}{1-(1-p)^M},$$

tedy

$$P(T = j+1) = P(X = j) = \frac{p(1-p)^j}{1-(1-p)^M}, j = 0; 1; 2; \dots; M-1, p \in (0; 1).$$

Pro takto určená j a p budou všechny členy kladné, součet všech členů je roven jedné, tedy z toho vyplývá, že tento předpis nám dává hodnoty menší než jedna, a tedy takto definované rozdělení má smysl. Nazvěme toto rozdělení useknutým geometrickým rozdělením.

Počítáme – li pro $t \in \langle 0; M)$

$$P(X \leq t) = \sum_{j=0}^{\lfloor t \rfloor} P(X = j) = \frac{p}{1-(1-p)^M} \sum_{j=0}^{\lfloor t \rfloor} (1-p)^j = \frac{1-(1-p)^{\lfloor t \rfloor}}{1-(1-p)^M},$$

dostaneme distribuční funkci X (a tedy i hledanou pravděpodobnost pro T):

$$P(T \leq t) = P(X \leq t-1) = \begin{cases} \dfrac{1-(1-p)^{\lfloor t-1 \rfloor}}{1-(1-p)^M}, \text{pro } t \in \langle 1; M+1), \\ 0, \text{pro } t < 1, \\ 1, \text{pro } t \geq M+1. \end{cases}$$

Hodnoty mimo interval $\langle 1; M+1)$ nejsou pro banku důležité. Jedničková pravděpodobnost pro $t \geq M+1$ interpretujeme jako splacenou půjčku, tedy zde může být pravděpodobnost jakákoli. Jedničkový výsledek dosazujeme pro splnění vlastností distribuční funkce. Střední hodnotu takového rozdělení spočítáme podle:

$$EX = \sum_{j=0}^{M-1} j\, P(X = j) = \frac{p}{1-(1-p)^M} \sum_{j=0}^{M-1} j(1-p)^j.$$

Položme

$$q = 1 - p.$$

Pak máme

$$\sum_{j=0}^{M-1} j\, q^j = q\, \frac{d}{dq}\left(\sum_{j=0}^{M-1} q^j\right) = q\, \frac{d}{dq}\, \frac{1-q^M}{1-q} = q\, \frac{-Mq^{M-1}(1-q) - (1-q^M)(-1)}{(1-q)^2}$$

$$= \frac{q}{(1-q)^2}\, (Mq^M - Mq^{M-1} + 1 - q^M),$$

takže

$$EX = \frac{(1-p)[M(1-p)^M - M(1-p)^{M-1} + 1 - (1-p)^M]}{p[1-(1-p)^M]}.$$

Odsud lze vyjádřit hodnotu p, přičemž v tomto případě dostaneme M řešení. Z nich vybereme pouze ta řešení, která jsou v intervalu $(0; 1)$. Pokud takové řešení neexistuje, pak tento model nemůžeme použít. Pokud je takových řešení více, získáme více modelů, které můžeme použít.

6. LGD: ztráta banky při selhání

6.1. Úvod do modelování parametru LGD

V této části postupujeme podle [1].

Parametr LGD se odhaduje pouze v rámci AIRB přístupu. Parametr LGD bývá často definován jako procentuální část z EAD, kterou banka ztratí při selhání klienta. Tato definice platí v případě, kdy banka jako příslušnou expozici zvolí dlužnou částku v okamžiku selhání. Obecně je tedy parametr LGD bezprostředně propojen s parametrem EAD. AIRB přístup klade na odhad několik požadavků:

a) v případě použití obnovených dat je nutná dokumentace jejich minulosti
b) důraz na opatrnost
c) započítává se hospodářský cyklus sebraných dat
d) soulad s definicí selhání dle BASEL II
e) LGD se odhaduje pro každou možnost selhání
f) minimálně musí být roven dlouhodobému váženému průměru ztrátových měr
g) v případě korporátního portfolia je nutné použít data z posledních sedmi let
h) v případě retailového portfolia je nutné použít data z posledních pěti let

Požadavků na parametr LGD a na jeho odhad je méně než v případě parametru PD.

Při pohledu na požadavky kladené na odhad parametru LGD se zdá, že tento odhad je snadný. To je dané tím, že parametr LGD není dosud prozkoumaný dostatečně do hloubky. Navíc na parametr LGD mají velký vliv specifika dané země. Při zkoumání detailů se významně zvyšuje složitost. Základní rozdíl mezi PD a LGD spočívá v tom, že parametr PD se zakládá na pevné události v daném čase, zatímco parametr LGD je dynamická záležitost, která může trvat několik let, protože vymáhání dluhů od klientů, kteří selhali, zabere nějakou dobu.

Další problém při modelování parametru LGD je sběr dat. Existují při typy zdrojů dat. Prvním typem jsou tržní data. Tato data jsou reprezentována cenami dluhopisů prodaných na sekundárním trhu. Jsou objektivní v případě jakéhokoli aktiva, které lze efektivně prodat. Sběr těchto dat je limitován, neboť není dostatek evidovaných dluhopisů. Druhým typem jsou implikovaná data, z nichž odhadujeme implikovanou hodnotu LGD. Tato data jsou odvozena z modelu oceňování aktiv, jehož standardizovaná verze neexistuje. Třetím typem jsou takzvaná pracovní data, ze kterých se odhadne pracovní LGD. Vznikají efektivními obnovami po pracovním procesu a diskontují se do času selhání. Jsou to interní historická data dostupná ve většině případů a jsou dostatečně objektivní. Parametr LGD se z nich odhaduje například na základě následující idee: ekonomická ztráta v sobě zahrnuje náklady a diskontní efekt. Označme r jako úrokovou míru. Nechť je tato míra konstantní v čase. Pak lze parametr LGD spočítat podle:

$$\mathrm{LGD} = 1 - \frac{\sum_t \frac{\text{návratnost}_t - \text{náklady}_t}{(1+r)^t}}{\mathrm{EAD}}. \tag{6.1}$$

Odsud je též zřejmé, že parametry LGD a EAD jsou navzájem propojeny. Oba se odhadují ve stejném čase.

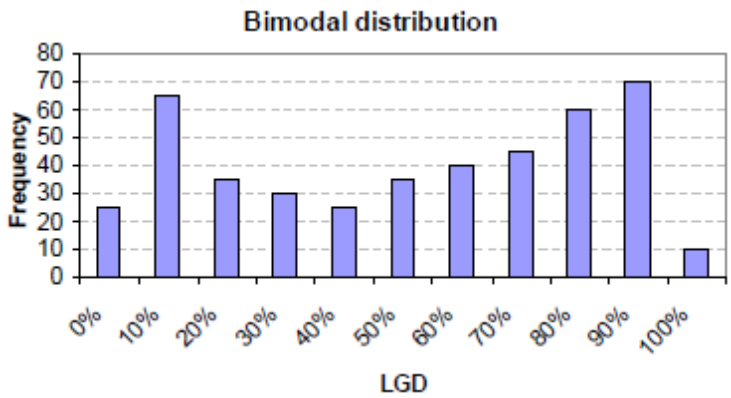
Příklad

Předpokládejme, že došlo k selhání firmy, která si půjčila 1 milion Euro. V době selhání je celkem splaceno 500 tisíc Euro (tj. EAD = 500 000). Jeden rok po selhání přichází banka o 1000 Euro. Dva roky po selhání je nad firmou prohlášen bankrot a banka dostane 200 tisíc Euro. Úroková míra je 5%. Za použití těchto informací použijeme vzorec (6.1) a dostaneme:

$$\mathrm{LGD} = 1 - \frac{\frac{-1000}{1{,}05} + \frac{200000}{1{,}05^2}}{500000} = 64\%.$$

□

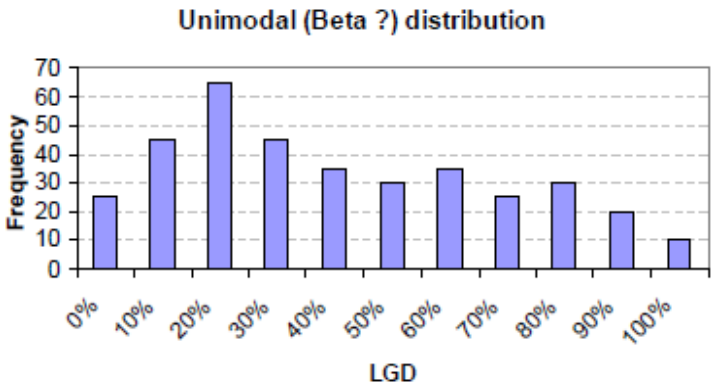
Zkoumá se též rozdělení parametru LGD a jeho vlastnosti. Rozdělení LGD se vyznačuje vysokou volatilitou (směrodatnou odchylkou). Nebývá normální, ale často nesymetrické nebo bimodální (viz histogram na obr. 12), respektive unimodální (viz histogram na obr. 13). To výrazně ztěžuje modelování parametru LGD.



Obr. 12 – příklad bimodálního rozdělení

Dále se zkoumá korelace mezi parametry PD a LGD. V průmyslu se jedná o nejklíčovější a nejkontroverznější záležitost. V řízení rizik se často modeluje jako fixní parametr. Většina studií se snaží ukázat, že parametry PD a LGD nejsou nekorelované. Tabulka 19 znázorňuje Altmanovu studii. Tato studie zkoumala rozdíly v hodnotě kapitálu při použití fixní hodnoty korelace mezi LGD a PD (v tabulce označeno jako fixní LDG), fixní hodnoty korelace mezi LGD a PD při započítání volatility LGD, tedy při započítání směrodatné odchylky do hodnoty parametru

LGD, (v tabulce volatilní LGD) a modelované hodnoty korelace mezi LGD a PD při použití statistických modelů, na dané hladině spolehlivosti. Rozdíl byl kolem třiceti procent.



Obr. 13 – příklad unimodálního rozdělení, pravděpodobně rozdělení Beta

Statistika	Fixní LGD	Volatilní LGD	Korelované LGD	Rozdíl fixní/korelované
99% VAR	435	437	564	29,6%
99,9% VAR	809	814	1053	30,1%

Tabulka 19 – Altmanova studie rozdílů mezi fixním LGD a korelovaným LGD

Parametru LGD se týkají další věci. Jednou z nich jsou náklady, neboť přímé i nepřímé náklady musí být započítány do parametru LGD. Započítání nákladů není důležité na úrovni portfolia, ale může být důležité na úrovni možností a vybavení, protože jejich započítání ztěžuje výpočet. Důležitá je též doba selhání, která může přinést problém z důvodu časových prodlev a nemožnosti dohledat data při selhání. Do parametru LGD též vstupuje úročení, protože parametr LGD se odhaduje na dobu delší než jeden rok. To znamená, že se změní hodnota peněz a je nutné započítat úrokovou míru. Parametr LGD pro celé portfolio může být nulový. Může se stát, že v případě nějakého úvěru vyjde záporná hodnota LGD. Hodnoty LGD se pak průměrují přes celé portfolio a dochází k nulovému LGD. K nulovému LGD může též dojít vlivem cenzorování dat, které je mnohdy nutné z nařízení regulátora.

6.2. Praktická metoda používaná k modelování

V této části popisujeme jednu z metod používanou k modelování parametru LGD. K modelování tohoto parametru lze použít regresní model (lineární regresi nebo zobecněný lineární model), ale tyto postupy nejsou příliš intuitivní. Intuitivnější metodou je metoda rozkladu parametru LGD, zvaný jako přístup nápravy a nenápravy k modelování parametru LGD, který používá banka působící na bankovním trhu v České republice. V následující části postupujeme podle jejích interních dokumentů. Základní vzorec při tomto přístupu je:

$$\mathrm{LGD} = \mathrm{PC} \cdot \mathrm{LGD}_{\mathrm{cured}} + (1 - \mathrm{PC}) \cdot \mathrm{LGD}_{\mathrm{irr}}, \tag{6.2}$$

kde PC je pravděpodobnost nápravy (z angl. Probability of curing), $\mathrm{LGD}_{\mathrm{cured}}$ je ztráta banky při nápravě (z angl. Loss given cured) a $\mathrm{LGD}_{\mathrm{irr}}$ je ztráta banky při nedobytnosti (z angl. Loss given irrecoverable). K zadefinování všech těchto pojmů je nutné nejprve definovat typy selhání.

6.2.1. Definice pojmů

Nejprve je nutné předefinovat pojem selhání klienta. Klienta lze považovat za klienta, který selhal, pokud je splněna alespoň jedna z následujících podmínek:

(i) klient pravděpodobně nesplatí;
(ii) klient má statut „DPD default“;
(iii)klient je nedobytný.

Všechny tři části je nutné speciálně definovat. Určit, že klient pravděpodobně nesplatí, lze na základě různých pozorování, například opakované krátké zpoždění plateb, vývoj na klientově běžném účtu, vysoká účetní ztráta dané firmy a podobně. Konkrétní podmínky, za kterých lze prohlásit, že klient pravděpodobně nesplatí, jsou interní záležitostí každé banky.

Pro definici statutu „DPD default“ je nejprve nutné zadefinovat pojem pseudo selhání. Pseudo selhání znamená, že klient má technicky zpoždění více než 90 dnů v dané splátce, která však není nedobytná, ale obnos, který klient poslal pozdě, nepřevyšuje nějakou danou hranici. Pseudo selhání není selhání a při výpočtech se s klienty, kteří splňují definici pseudo selhání, pracuje jako s klienty, kteří neselhali. „DPD statut“ nastává, pokud se daná splátka zpozdí o více než 90 dnů a není splněna definice pseudo selhání.

Klient je nedobytný, pokud je splněna alespoň jedna z následujících podmínek:

(i) daná úvěrová smlouva je ukončena bankou;
(ii) nezávislý soud rozhodl o odnětí dané záruky neplatiči.

Pro takové klienty se pak počítá ztráta banky z nedobytnosti. Pokud je klient nedobytný, bude nedobytný navždy, zatímco pokud má klient „DPD statut“ nebo pro něj lze usoudit, že pravděpodobně nesplatí, neznamená to, že klient všechny dluhy zpětně s nějakým dlouhým zpožděním nemůže splatit a banka jej pak může opět považovat za klienta, který neselhal.

Ztráta banky při nápravě souvisí s pojmem náprava. Náprava nastává právě tehdy, když klient již dříve splnil definici selhání, vše zpětně splatil a banka jej opět považuje za dobrého a spolehlivého klienta. S pojmem náprava souvisí pojem míra nápravy (ozn. CR, z angl. cure rate). Definice této míry je rozdílná pro portfolio s klienty, kteří selhali, a pro portfolio s klienty, kteří selhali. Nejprve je nutné zvolit den, ve kterém dojde k pozorování, takzvaný pozorovací den, a nějakou dobu, během které se sledují počty klientů, kteří selhali a u kterých došlo k nápravě (nadále pozorovací doba). Tato doba zpravidla následuje po pozorovacím dni. Uvažujme nejprve portfolio složené z klientů, kteří do pozorovacího dne neselhali. Označme A jako množinu těchto klientů v den pozorování. V pozorovací době sledujeme chování klientů. V té době někteří klienti selhali, z nich se někteří stali nedobytnými, u některých došlo k nápravě a v ostatních případech nelze říci nic o jejich dalším chování (nevíme, zda dojde k nápravě či zda se stanou nedobytnými). Proto zpravidla zvolíme ještě dodatečnou dobu, zpravidla dlouhou několik měsíců, ve které sledujeme totéž. Buď B množina klientů, o kterých nelze nic říci po pozorovací době. Buďte m počet klientů, kteří selhali během pozorovací doby, j počet klientů, u kterých došlo k nápravě během pozorovací doby, a k počet klientů, u kterých došlo k nápravě během dodatečné doby. Pak

$$\mathrm{CR} = \frac{\mathrm{j} + \mathrm{k}}{\mathrm{m}}.$$

Pro portfolio tvořené klienty, kteří jsou v pozorovacím dni považováni za klienty, kteří selhali, je výsledný vzorec pro CR stejný, ale definice jednotlivých prvků jsou jiné. A zde značí množinu všech klientů, kteří selhali, ale není jasné, zda dojde k nápravě nebo k nedobytnosti, B značí množinu všech klientů, kteří jsou ve stejné fázi po pozorovací době, m značí počet prvků A, j značí počet klientů, u kterých došlo k nápravě během pozorovací doby, a k značí počet klientů, u kterých došlo k nápravě během dodatečné doby.

Ještě označme $\mathrm{LGD}_{\mathrm{obs}}, \mathrm{EAD}_{\mathrm{obs}}, \mathrm{LGD}_{\mathrm{cured,obs}}, \mathrm{LGD}_{\mathrm{irr,obs}}$ jako pozorované hodnoty příslušných parametrů, které jsou protějškem k modelovaným hodnotám příslušných parametrů. Protějškem k PC je výše definovaná CR.

Ještě dodejme, že banka pro své klienty používá PD ratingy, s jejichž pomocí dané klienty hodnotí. Hodnocení od 1 do 9 znamená normální rating, hodnocení od 10 do 12 znamená selhání klienta, přičemž hodnocení 12 znamená nedobytnost, a ostatní hodnocení znamenají, že rating není uveden. Odsud vyplývá označení používané dále (například značení $\mathrm{W}_{10/11}$ znamená, že se jedná o danou charakteristiku klientů, jejichž PD rating je 10 nebo 11, a podobně).

6.2.2. Modelování PC

K modelování složky PC se zpravidla používá logistická regrese (viz kapitola 3). Hodnocení modelu na základě jeho diverzifikačních schopností probíhá stejně jako u modelů pro modelování parametru PD, stejným způsobem probíhá i kalibrace modelu. Jediným problémem spojeným s touto složkou je výběr proměnných k danému modelování. Pokud máme nějaké proměnné, pak lze snadno použít algoritmus Stepwise popsaný v kapitole 3. Transformace proměnných na dummy proměnné je též stejná. Je však nutné pečlivě vybrat proměnné, ze kterých posléze vybíráme proměnné pro odhad. Je nutné mít dostatek finančních dat o klientovi. S těmito daty souvisí různé problémy, například chyby v databázích nebo chybějící hodnoty obvykle doplňované nulami, dále například různé zdroje těchto dat a jejich rozporuplnost. Rovněž jsou důležitá nefinanční data o klientovi. Občas je též potřeba mít odhad parametru PD nebo různá data od jiných subjektů a registrů.

6.2.3. Modelování $\mathrm{LGD_{cured}}$

$\mathrm{LGD_{cured}}$ označuje ztrátu banky v případech, kdy klient selže a pak u něj dojde k nápravě a je zařazen z portfolia neplatících klientů zpět do normálního portfolia. Při odhadu tohoto parametru může nastat

$$\mathrm{LGD_{cured}} = 0\%.$$

To znamená, že klient sice selhal, ale pak zaplatil vše zpět včetně různých pokut, penále a nákladů banky na vymáhání.

Při odhadování této složky je nutné vyčíslit náklady banky spojené se selháním a následnou nápravou a rovněž vyčíslit ztráty banky při nápravě klienta. Označme $\mathrm{W_{10/11}}$ jako celkové náklady oddělení intenzivního sledování (neboli náklady na sledování) u klientů, u kterých nevíme, zda se stanou nedobytnými či napravenými (zpravidla se jedná o roční průměry). Nechť $\mathrm{A_{10/11}}$ je celková hodnota dlužných částek klientů, u kterých nevíme, co bude následovat (náprava nebo nedobytnost?), v daném čase. Obvykle $\mathrm{A_{10/11}}$ bývá průměrem těchto počtů za několik let. Buď $\mathrm{T_{until\ curing}}$ množstvím vážený průměr dob, během kterých u daného klienta došlo k nápravě. Jako váhy lze použít například $\mathrm{EAD_{obs}}$. Položme

$$\mathrm{IC_{cured}} = \frac{\mathrm{W_{10/11}}}{\mathrm{A_{10/11}}} \cdot \mathrm{T_{until\ curing}}.$$

$\mathrm{IC_{cured}}$ jsou průměrné náklady vyjádřené v procentech. Označme dále:

- $\mathrm{L_{10/11}}$ je celková ztráta způsobená klienty, u kterých nevíme, zda se stanou nedobytnými či zda u nich dojde k nápravě, za rok; obvykle se jedná o roční průměr;

- $L_{settlement/sale}$ jako celkovou ztrátu způsobenou různými dohodami nebo prodejem dlužních úpisů, které se následně ztrácí z portfolia během nápravy;
- $T_{until\ curing\ (excl\ settlement)}$ jako množstvím vážený průměr dob, během kterých u daného klienta došlo k nápravě, vyjma případů, u kterých došlo k domluvě nebo prodeji;
- $T_{until\ settlement}$ jako množstvím vážený průměr dob, během kterých u daného klienta došlo k dohodě nebo prodeji.

Jako váha se stejně jako výše používá EAD_{obs}. Průměrná ztráta v procentech je pak

$$L_{cured} = \frac{L_{10/11}}{A_{10/11}} \cdot T_{until\ curing\ (excl\ settlement)} + \frac{L_{settlement/sale}}{CR \cdot A_{10/11}} \cdot T_{until\ settlement},$$

přičemž $CR \cdot A_{10/11}$ je aproximace celkového množství klientů, kteří selhali a u kterých je možné, že dojde k nápravě.

Před samotným odhadem LGD_{cured} je vhodné provést diskuzi týkající se $A_{10/11}$. Je vhodné se přesvědčit, zda všechny prvky, použité k výpočtu $A_{10/11}$ jsou zaktualizované na střední návratnost. Pokud tomu tak nebude, pak dojde k přehodnocení $A_{10/11}$ a podhodnocení LGD_{cured}. Pak již lze odhadnout LGD_{cured} jako

$$LGD_{cured} = IC_{cured} + L_{cured}.$$

6.2.4. Modelování LGD_{irr}

Při modelování LGD_{irr} lze použít mnoho metod. Je možné použít regresní modely včetně zobecněných lineárních modelů, rozhodovací stromy, strojové výpočty a další metody. Ne všechny metody jsou dostatečně srozumitelné a intuitivní. Jedna z metod k modelování LGD_{irr} má tři až čtyři kroky a modeluje LGD_{irr} pomocí:

$$LGD_{irr} = \max\left[1 - R_{EAD} - \frac{\text{zajištěná návratnost}}{EAD} - \frac{\text{nezajištěná návratnost}}{EAD} + C_{irrecoverable}; 0\right].$$

V praxi jsou povoleny a používány tři výklady tohoto vzorečku:

(i) $LGD_{irr} = \max\left[\left((1 - R_{EAD})(1 - R_{unsecured}) - \sum R_{secured,asset} \frac{\text{cena aktiva}}{EAD} - \right.\right.$ Rsecured,securityhodnota cenného papíruEAD+Cirrecoverable;0*;*

(ii) $LGD_{irr} =$
$\max\left[\left(1 - R_{EAD} - \sum R_{secured,asset} \frac{\text{cena aktiva}}{EAD} - \right.\right.$

$$\sum R_{secured,security} \frac{\text{hodnota cenného papíru}}{EAD}\Big)(1 - R_{unsecured}) + C_{irrecoverable}; 0\Big];$$

(iii) $LGD_{irr} = \max[(1 - R_{EAD})(1 - R_{claim}) + C_{irrecoverable}; 0]$.

Pro portfolio, ve kterém se nezajištěné výše plateb zpravidla zaplatí před jakoukoli bankovní zástavou, se zpravidla používá první výklad. Druhý výklad je často používaný pro portfolio, ve kterém dochází k bankovním zástavám před jakýmkoli pokusem o zaplacení nezajištěných výší plateb. Pokud pozorování, u kterých došlo k domluvě nebo prodeji dlužních úpisů, tvoří nezanedbatelnou část všech pozorování s nedobytnými klienty nebo pokud první ani druhý výklad není kvůli pozorováním možný, použije se třetí výklad. Dodejme jen, že toto jsou pouze doporučení, kterými se banka nemusí řídit. Ve všech třech uvedených vzorcích:

- R_{EAD} značí rozdíl mezi pozorovanou expozicí při selhání EAD_{obs} použitou při modelování LGD_{irr} a hodnotou, které se banka efektivně domůže do domáhacího dne; domáhací den banka obvykle volí, pro ručitele je často spojen se dnem, kdy je vše vymoženo, jinak je tento den spojen se dnem, kdy se klient stane nedobytným;
- $R_{secured}$ odráží návratnost při zárukách:
  - pro záruky na aktivech (například dům) je tato míra relativní hodnotou daného aktiva ($R_{secured,asset}$);
  - pro záruky na cenných papírech (například ručitelé) se tato míra vyjadřuje jako procento hodnoty kontraktu ($R_{secured,security}$);
- $R_{unsecured}$ odráží nezajištěnou návratnost vyjádřenou buď jako procento efektivně vymožených dlužných hodnot (používá se při prvním výkladu) nebo jako procento vymožených dlužných hodnot po všech pokusech o jejich zajištění (používá se při druhém výkladu);
- R_{claim} odráží celkovou návratnost vyjádřenou jako procento efektivně vymožených dlužných hodnot;
- $C_{irrecoverable}$ značí náklady a ztráty pro nedobytné klienty.

Všechna R se nazývají míry návratnosti (z angl. recovery rates), vždy s danou vlastností. Následující řádky pojednávají o tom, jak všechny míry odhadnout. Veškeré kroky při odhadování LGD_{irr} spočívají v odhadu všech používaných měr. Ještě dodejme, že

$$LGD_{irr,obs} = \frac{EAD_{obs} - \text{návratnost} + \text{náklady}}{EAD_{obs}}.$$

Tento vzorec přímo souvisí se vzorcem (6.1), jediným rozdílem je diskontování ve vzorci (6.1).

Míra návratnosti EAD se modeluje pomocí pozorované hodnoty:

$$R_{EAD,obs} = \frac{EAD_{obs} - \text{vymožená hodnota}}{EAD_{obs}}.$$

Konkrétní vzorec pro odhad R_{EAD}uvedeme později společně se všemi ostatními měrami. Expozice v domáhacím dni se musí diskontovat do momentu selhání, vzorec pak je

$$R_{EAD,obs} = 1 - \frac{\text{expozice v domáhacím dni diskontovaná do momentu selhání}}{EAD_{obs}}.$$

Pro odhad R_{EAD} potřebujeme znát alespoň pozorované expozice při selhání (tedy EAD_{obs}), moment selhání klienta, částku, která byla efektivně vymožena a domáhací den.

Chráněné míry návratnosti $R_{secured}$ se odhadují pomocí

$$R_{secured,asset,obs} = \frac{\text{diskontovaná zajištěná návratnost}}{\text{hodnota aktiva}}$$

a

$$R_{secured,security,obs} = \frac{\text{diskontovaná zajištěná návratnost}}{\text{hodnota ceného papíru}}.$$

Pro odhad $R_{secured}$ je třeba znát alespoň hodnotu aktiv, respektive hodnotu cenného papíru, v pozorovacím dni, návratnosti na cenných papírech a dny, kdy k těmto návratnostem došlo, a momenty selhání klientů, kteří jako záruku použili cenné papíry.

Nechráněná míra návratnosti $R_{unsecured}$ se odhaduje pomocí

$$R_{unsecured,obs} = \frac{\text{diskontovaná nechráněná návratnost}}{\text{diskontovaná expozice v domáhacím dnu} - \text{diskontovaná návratnost}}.$$

Doplňme, že pojmem návratnost rozumíme vymoženou částku, kterou klient zaplatí. Pro odhad $R_{unsecured}$ je nutné znát alespoň záruky, které může daný klient poskytnout, nezajištěnou návratnost a příslušné dny, kdy k návratnosti došlo.

Domožená míra návratnosti R_{claim} se odhaduje pomocí

$$R_{claim,obs} = \frac{\text{diskontované návratnosti nebo výtěžky}}{\text{diskontovaná expozice v domáhacím dni}}.$$

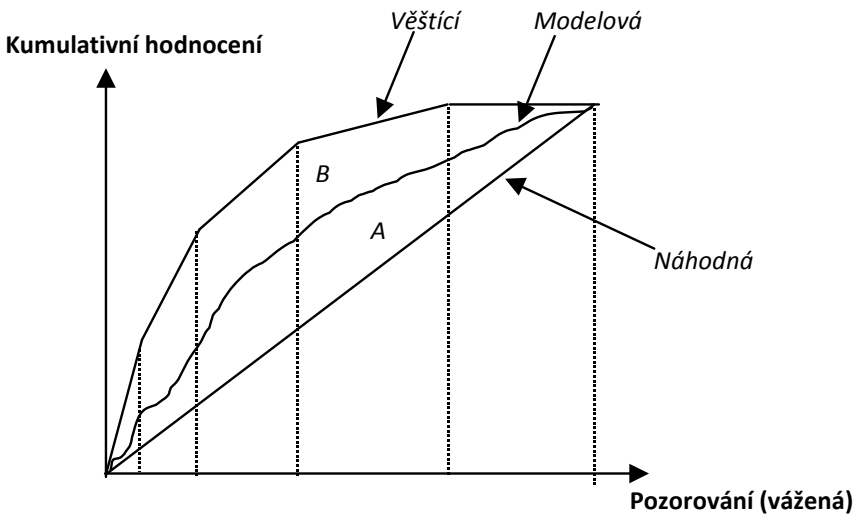
K odhadu R_{claim} stačí znát návratnosti nebo výtěžky u daných klientů a příslušná data návratnosti.

Výsledné míry návratnosti se odhadují pomocí formule

$$\sum_{i \in segment} \text{váha}_i \left(R_{segment} - R_{observed,i}\right)^2. \tag{6.3}$$

Ve formuli (6.3) $R_{segment}$ je odhadnutá míra návratnosti pro daný segment, přičemž jako $R_{segment}$ se bere postupně R_{EAD}, $R_{secured}$, $R_{unsecured}$ a R_{claim}. $R_{observed,i}$ je pozorovaná míra návratnosti pro i – té pozorování definovaná výše (definice $R_{EAD,obs}$ a další definice). Každá míra se ze vzorce (6.3) odhadne zvlášť minimalizací dané formule. Tedy pro každou míru budeme minimalizovat daný vážený součet. Váhy volíme pomocí následujících pravidel:

a) pro R_{EAD} použijeme EAD_{obs};
b) pro $R_{secured,asset}$ použijeme pozorovanou hodnotu aktiv;
c) pro $R_{secured,security}$ použijeme (tržní) hodnotu cenného papíru;
d) pro $R_{unsecured}$ použijeme pozorovanou hodnotu, které se můžeme domoci po vzetí v potaz všech možných chráněných návratností, tedy:
   $$\text{diskontovaná expozice v domáhacím dni} - \text{diskontovaná návratnost}$$
e) pro R_{claim} použijeme efektivně vymožitelnou dlužnou hodnotu, tedy diskontovanou expozici v domáhacím dni.



Obr. 14 – hodnocení modelů měr návratnosti

Odhady měr návratnosti lze hodnotit pomocí různých ukazatelů. Jedním z nich je takzvaná vážená číselná síla. Jedná se o ukazatel podobný Giniho koeficientu, rozdíl spočívá v tom, že Giniho koeficient má pouze dvě kategorie pro hodnocení, zatímco vážená číselná síla jich má více. Při počítání vážené číselné síly se nejprve všechny $R_{observed}$ rozdělí do skupin a přiřadí se jim sklon. Toto dělení a přiřazování uvádí tabulka 20. Pak se sestaví obrázek, jehož příklad je na obr. 14, který odpovídá AR diagramu používanému k hodnocení modelů používaných k modelování parametru PD. Tento obrázek obsahuje tři křivky, které začínají v počátku, tj. v bodu (0; 0),

a vedou do bodu, který odpovídá celkové velikosti portfolia na vodorovné ose a součtu všech sklonů vynásobených příslušnou váhou na svislé ose (kumulativní hodnocení). Namísto velikosti portfolia se často používá součet všech vah, tedy namísto příslušných pozorování se na vodorovnou osu nanášejí příslušné váhy, protože to pak lépe odpovídá realitě. Váhy se používají stejné jako při odhadování měr ve formuli (6.3).

Skupina	Skupiny $R_{observed}$	Sklon
1	$0\% \leq R_{observed} < 10\%$	9
2	$10\% \leq R_{observed} < 20\%$	8
3	$20\% \leq R_{observed} < 30\%$	7
4	$30\% \leq R_{observed} < 40\%$	6
5	$40\% \leq R_{observed} < 50\%$	5
6	$50\% \leq R_{observed} < 60\%$	4
7	$60\% \leq R_{observed} < 70\%$	3
8	$70\% \leq R_{observed} < 80\%$	2
9	$80\% \leq R_{observed} < 90\%$	1
10	$90\% \leq R_{observed} \leq 100\%$	0

Tabulka 20 – dělení $R_{observed}$ do skupin a přiřazování sklonů

Sestavení křivek v daném obrázku je pak velmi jednoduché. Věštící křivka vychází z pořadí pozorování, které souvisí s pozorovanými měrami návratnosti, respektive se skupinami v tabulce 20. Tedy příslušná skupina je závisle proměnnou. Věštící křivka se sestrojí tak, že s každým nárůstem vážené hodnoty daného pozorování na vodorovné ose se zvýší hodnota na svislé ose o hodnotu, která je rovna příslušnému sklonu vynásobenému danou váhou (váhy jsou stejné jako ve formuli (6.3)). Jak již je zmíněno výše, křivka začíná v počátku. Nejprve použijeme všechna pozorování z první skupiny, pak z druhé a tak pokračujeme dále k desáté skupině. Znamená to tedy, že věštící křivka bude nejprve velmi strmá a na konci již bude téměř vodorovná. Modelová křivka se sestaví stejně, jediný rozdíl spočívá v použití modelovaných hodnot, tedy $R_{segment}$, kterým se přiřadí příslušná skupina a sklon. Pak se sestaví modelová křivka. Náhodná křivka je v podstatě přímka vedoucí z počátku do cílového bodu. Její definice je v praxi složitější.

Takto sestavené křivky vymezují plochy A a B znázorněné na obr. 14. Pak dostáváme:

$$\text{vážená číselná síla} = \frac{A}{A + B}.$$

Čím je vážená číselná síla větší, tím je model lepší. Mnohdy je lepší ukázat obr. 14 a na jeho základě posoudit kvalitu modelu. Je zjevné, že čím je modelová křivka blíže

věštící křivce, tím je vážená číselná síla větší a model je tím lepší. Důvodem pro tohle je velmi složité spočítání daných ploch.

Míry návratnosti není nutné odhadovat všechny. Záleží vždy na použitém výkladu. Nicméně po jejich odhadnutí a posouzení kvality daných odhadů je ještě nutné odhadnout $C_{irrecoverable}$. K tomu nejprve zavedeme značení:

- $W_{10/11}$ jsou celkové náklady oddělení intenzivního sledování (neboli náklady na sledování) u klientů, u kterých nevíme, zda se stanou nedobytnými či napravenými (zpravidla se jedná o roční průměry);
- $A_{10/11}$ je celková hodnota dlužných částek klientů, u kterých nevíme, co bude následovat (náprava nebo nedobytnost?), v daném čase, opět se zpravidla jedná o roční průměr;
- $T_{until\ irrecoverable}$ je množstvím vážený průměr dob, během kterých u daného klienta došlo k nedobytnosti;
- $L_{10/11}$ je celková ztráta způsobená klienty, u kterých nevíme, zda se stanou nedobytnými či zda u nich dojde k nápravě, za rok; obvykle se jedná o roční průměr;
- $T_{until\ irrecoverable\ (excl\ settlement)}$ je množstvím vážený průměr dob, během kterých u daného klienta došlo k nedobytnosti, vyjma případů, u kterých došlo k domluvě nebo prodeji;
- W_{12} jsou celkové náklady vymáhacího oddělení (vnější i vnitřní) na nedobytné klienty za rok; zpravidla se bere průměr za několik let;
- A_{12} je celková hodnota dlužných částek u nedobytných klientů v daném čase, opět se zpravidla jedná o roční průměr;
- $T_{in\ irrecoverable}$ je množstvím vážený průměr dob do uzavření případů u nedobytných klientů;
- $L_{settlement/sale}$ jako celkovou ztrátu způsobenou různými dohodami nebo prodejem dlužních úpisů, které se následně ztrácí z portfolia během nedobytnosti;
- $T_{until\ settlement/sale}$ je množstvím vážený průměr dob mezi dnem selhání a dnem, kdy došlo k dohodě nebo prodeji.

Jako váha se zpravidla používá EAD_{obs}. Mnoho značených veličin odpovídá veličinám používaným k modelování LGD_{cured}, ale nemusí se stejným způsobem počítat. Stejně jako u modelování LGD_{cured} je i zde vhodné provést diskuzi týkající se $A_{10/11}$ a A_{12}. Je vhodné se přesvědčit, zda všechny prvky, použité k výpočtu $A_{10/11}$ a A_{12} jsou zaktualizované na střední návratnost. Pokud tomu tak nebude, pak dojde k přehodnocení $A_{10/11}$ a A_{12} a podhodnocení $C_{irrecoverable}$.

Při takto zavedeném značení označme dále

$$IC_{until\ irrecoverable} = \frac{W_{10/11}}{A_{10/11}} \cdot T_{until\ irrecoverable}$$

jako minulé procentuální náklady na nedobytné klienty. Průměrné procentuální ztráty z nedobytných klientů jsou

$$L_{\text{until irrecoverable}} = \frac{L_{10/11}}{A_{10/11}} \cdot T_{\text{until irrecoverable (excl settlement)}}.$$

Průměrné procentuální náklady na vymáhání od nedobytných klientů jsou

$$C_{\text{in irrecoverable}} = \frac{W_{12}}{A_{12}} \cdot T_{\text{in irrecoverable}}.$$

Ztráta způsobená různými dohodami nebo prodejem dlužních úpisů, které se následně ztrácí z portfolia během nedobytnosti, vstupuje do nákladů jako průměrná procentuální ztráta:

$$L'_{\text{settlement/sale}} = \frac{L_{\text{settlement/sale}}}{A_{12}} \cdot T_{\text{until settlement/sale}}.$$

Odsud se již pouze spočítá $C_{\text{irrecoverable}}$ pomocí:

$$C_{\text{irrecoverable}} = IC_{\text{until irrecoverable}} + L_{\text{until irrecoverable}} + C_{\text{in irrecoverable}} + L'_{\text{settlement/sale}}.$$

Pak již lze snadno spočítat LGD_{irr}.

Jak již bylo zmíněno výše, při modelování LGD_{irr} je nutné diskontování do pozorovacího dne. K tomu je nejprve nutné zvolit diskontní míru. Existuje spousta literatury, jak tuto míru zvolit. Doporučuje se zvolit diskontní míru jako místní míru financování zdrojů navýšenou o rizikovou přirážku 2%. Samotné diskontování je pak zřejmé.

6.2.5. Smíšený model, jeho kalibrace a testování

Pokud je již odhadnuté vše, co jsme potřebovali k modelování parametru LGD, lze odhadnout LGD pomocí vzorce (6.2). Takový model je však ještě nutné kalibrovat a otestovat. Na to je nejprve nutné získat data použitelná k výpočtu LGD_{obs}. Při tomto výpočtu je vždy nutné rozlišit, jakého máme klienta. Pokud se jedná o klienta, u kterého došlo k nápravě, pak LGD_{obs} se počítá jako:

$$LGD_{\text{obs}} = LGD_{\text{cured,obs}} = IC_{\text{cured}} + \frac{L_{\text{obs}}}{EAD_{\text{obs}}},$$

kde L_{obs} je pozorovaná ztráta u daného klienta a ostatní dvě veličiny jsou definované v kapitole 6.2.1 a 6.2.3. V případě, kdy není k dispozici pozorovaná ztráta u každého klienta, u kterého došlo k nápravě, je k dispozici ztráta celého portfolia, respektive je nutné ji určit, klientů, u kterých došlo k nápravě, a

$$\mathrm{LGD_{obs}} = \mathrm{LGD_{cured,obs}} = \mathrm{IC_{cured}} + \frac{\sum_i \mathrm{L_{obs,i}}}{\sum_i \mathrm{EAD_{obs,i}}},$$

přičemž i probíhá dané portfolio. Máme – li klienta, který se stal nedobytným, pak

$$\mathrm{LGD_{obs}} = \mathrm{LGD_{irr,obs}} = \frac{\text{pozorovaná expozice v domáhacím dni} - \text{pozorované návratnosti} + \mathrm{L_{obs}}}{\mathrm{EAD_{obs}}} + \mathrm{IC_{until\ irrecoverable}} + \mathrm{C_{in\ irrecoverable}},$$

kde $\mathrm{L_{obs}}$ je pozorovaná ztráta u daného klienta a ostatní veličiny jsou definované v kapitole 6.2.1 a 6.2.4. Tím získáme pro každého klienta zvlášť pozorovanou hodnotu parametru LGD. Pak již je možné provést kalibraci modelu.

Nejprve se provádí kalibrace pod – modelů, tedy PC modelu a $\mathrm{LGD_{irr}}$ modelu. Kalibrace PC modelu je uvedena v kapitole 6.2.2 a kalibrace $\mathrm{LGD_{irr}}$ se provádí stejným způsobem, jako v případě smíšeného modelu. Model pro $\mathrm{LGD_{cured}}$ se nekalibruje. Popíšeme tedy pouze kalibraci finálního LGD modelu.

Před kalibrací LGD modelu je nejprve nutné určit kalibrační cíl (CaTa, z angl. Calibration Target). K tomu se zpravidla používá aritmetický průměr všech pozorovaných hodnot $\mathrm{LGD_{obs}}$:

$$\mathrm{LGD_{obs}} = \frac{\sum_{i=1}^{n} \mathrm{LGD_{obs,i}}}{n},$$

nebo vážený průměr:

$$\mathrm{LGD_{(obs)}} = \frac{\sum_{i=1}^{n} \mathrm{outst\ exp_{obs,i}} \cdot \mathrm{LGD_{obs,i}}}{\sum_{i=1}^{n} \mathrm{outst\ exp_{obs,i}}},$$

přičemž $\mathrm{outst\ exp_{obs,i}}$ je hodnota nezaplacené expozice v pozorovacím dni, n je velikost vybrané části klientů, a i probíhá přes všechna taková pozorování.

Kalibrační funkce misí splňovat tři kritéria:

1) musí být ohraničena 0 a 100%;
2) vážený průměr modelu po kalibraci, respektive odhadnutých hodnot parametru LGD, vybrané části portfolia musí být přibližně roven CaTa;
3) funkce je kladným násobkem modelu před kalibrací navýšeným o konstantu.

Znamená to tedy, že kalibrační funkce je:

$$\mathrm{LGD_{cal,i}} = \begin{cases} 0, & \alpha \cdot \mathrm{LGD_{pre,i}} + \beta < 0; \\ \alpha \cdot \mathrm{LGD_{pre,i}} + \beta, & 0 \leq \alpha \cdot \mathrm{LGD_{pre,i}} + \beta \leq 1\,; \\ 1, & \alpha \cdot \mathrm{LGD_{pre,i}} + \beta > 1. \end{cases}$$

Příslušné omezení pak je:

$$\frac{\sum_{j=1}^{m} \text{outst exp}_{obs,j} \cdot \text{LGD}_{cal,j}}{\sum_{j=1}^{m} \text{outst exp}_{obs,j}} \approx \text{CaTa}.$$

Ve dvou posledních vzorcích je použito značení:

- $\text{LGD}_{cal,i}$ je hodnota parametru LGD pro i – tého klienta po kalibraci;
- $\text{LGD}_{pre,i}$ je hodnota parametru LGD pro i – tého klienta před kalibrací;
- α je kladné číslo;
- β je jakékoli reálné číslo;
- $\text{outst exp}_{obs,i}$ je i – tá pozorovaná hodnota expozice;
- i je index daného pozorování
- j probíhá přes vybranou část portfolia, přičemž m je její velikost.

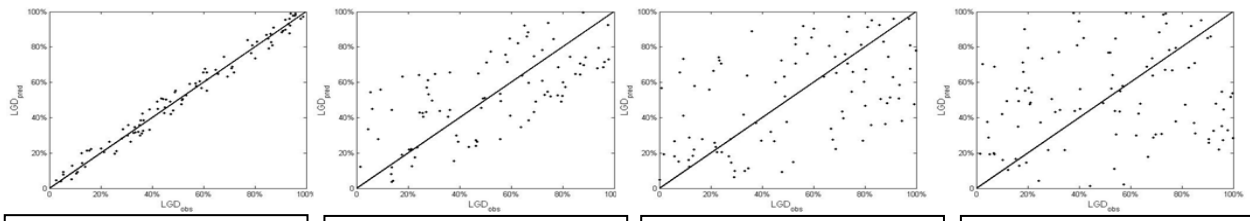
Hodnoty α a β se volí tak, aby příslušné omezení platilo co nejlépe, přičemž znak ≈ znamená, že dané hodnoty jsou si co neblíže, tedy jsou si co nelépe rovny.

Hodnoty LGD	Stupeň
Jednoduchá stupnice	
$0\% \leq \text{LGD} < 10\%$	1
$10\% \leq \text{LGD} < 30\%$	2
$30\% \leq \text{LGD} < 50\%$	3
$50\% \leq \text{LGD} < 70\%$	4
$70\% \leq \text{LGD} < 90\%$	5
$90\% \leq \text{LGD} \leq 100\%$	6
Rozšířená stupnice	
$0\% \leq \text{LGD} < 3{,}34\%$	1a
$3{,}34\% \leq \text{LGD} < 6{,}67\%$	1b
$6{,}67\% \leq \text{LGD} < 10\%$	1c
$10\% \leq \text{LGD} < 16{,}67\%$	2a
$16{,}67\% \leq \text{LGD} < 23{,}34\%$	2b
$23{,}34\% \leq \text{LGD} < 30\%$	2c
$30\% \leq \text{LGD} < 36{,}67\%$	3a
$36{,}67\% \leq \text{LGD} < 43{,}34\%$	3b
$43{,}34\% \leq \text{LGD} < 50\%$	3c
$50\% \leq \text{LGD} < 56{,}67\%$	4a
$65{,}67\% \leq \text{LGD} < 63{,}34\%$	4b
$63{,}34\% \leq \text{LGD} < 70\%$	4c
$70\% \leq \text{LGD} < 76{,}67\%$	5a
$76{,}67\% \leq \text{LGD} < 83{,}34\%$	5b
$83{,}34\% \leq \text{LGD} < 90\%$	5c
$90\% \leq \text{LGD} < 93{,}34\%$	6a
$93{,}34\% \leq \text{LGD} < 96{,}67\%$	6b
$96{,}67\% \leq \text{LGD} \leq 100\%$	6c

Tabulka 21 – příklad dělení parametru LGD

V praxi se část hodnoty parametru LGD kategorizují a daným kategoriím se přiřazuje stupeň, protože tohle je intuitivnější. Je možné zvolit několik kategorií, příklady uvádí tabulka 21 (jednoduchá stupnice a rozšířená stupnice).

Po kalibraci (nebo i před kalibrací) je vhodné otestovat model a posoudit tak kvalitu odhadu. Testuje se PC model, $\mathrm{LGD_{irr}}$ model a smíšený model, přičemž $\mathrm{LGD_{irr}}$ model a smíšený model se testují pomocí stejných ukazatelů. $\mathrm{LGD_{cured}}$ model se netestuje. K testování PC modelu se zpravidla používá Giniho koeficient a další ukazatelé (viz kapitola 4). K testování smíšeného modelu se, kromě jiného, používá vážená číselná síla. Ta se většinou používá v případě kategorizování parametru LGD, přičemž stupnice se použije podobná, jako v případě testování měr návratnosti. Nicméně vážená číselná síla je vhodná k testování měr návratnosti, zatímco k testování celého LGD modelu lze použít i jiné ukazatele. Jednou z možností, jak zhodnotit kvalitu modelu je pomocí takzvaného scatterplotu. Scatterplot ukazuje závislost LGD (občas značeno jako $\mathrm{LGD_{pred}}$) na svislé ose a $\mathrm{LGD_{obs}}$ na vodorovné ose a říká, jaká je predikční schopnost modelu vzhledem k realitě. Příklady jsou na obr. 15. Každý scatterplot znázorňuje situaci a text pod ním popisuje, zda je model dobrý či nikoli.



Velmi dobrý model	Dobrý model	Slabý model	Velmi slabý model

Obr. 15 – příklady scatterplotů

Dalším ukazatelem kvality modelu je R^2 statistika. Označme pozorování jako x_i a odhadnuté hodnoty jako y_i, nechť je takových hodnot n. Označme průměrné hodnoty jako

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

a

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Výpočet R^2 spočívá ve výpočtu součtů čtverců. Označme

$$SS_{tot} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

jako součet čtverců v pozorovaných hodnotách,

$$SS_{model} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

jako součet čtverců v odhadnutých hodnotách a

$$SS_{err} = \sum_{i=1}^{n} (x_i - y_i)^2$$

jako součet čtverců mezi pozorovanými a odhadnutými hodnotami. Pak

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}.$$

Často se používá vážená R^2 statistika, do níž vstupují vážené součty čtverců. V tomto případě máme značení:

$$SS_{tot,weighted} = \sum_{i=1}^{n} EAD_{obs}(x_i - \bar{x})^2,$$

$$SS_{model,weighted} = \sum_{i=1}^{n} EAD_{obs}(y_i - \bar{y})^2$$

a

$$SS_{err,weighted} = \sum_{i=1}^{n} EAD_{obs}(x_i - y_i)^2.$$

Odsud pak máme

$$R^2_{weighted} = 1 - \frac{SS_{err,weighted}}{SS_{tot,weighted}}.$$

Čím je R^2 blíže jedné, tím je model lepší. Nicméně informace podaná statistikou R^2 nemusí být zcela relevantní. Lze najít dva modely se stejnou hodnotou statistiky R^2 a s různými scatterploty a správností.

6.2.6. Volba pozorovacího dne

Je zjevné, že před sběrem dat je nutné zvolit pozorovací den a následnou pozorovací dobu, během které budeme sledovat selhání klientů, jejich chování a všechny výše uvedené veličiny. Na základě toho určíme, zda u klienta došlo k nápravě či se stal nedobytným, spočítáme všechny potřebné míry návratnosti a příslušné hodnoty LGD a odhadneme pravděpodobnost nápravy. Též je nutné vyřešit chybějící data a data, která nedávají smysl. K tomu lze zvolit nějakou fixní hodnotu (nulovou, průměr, medián) nebo použít nějaké sofistikovanější metody, kterých je několik a o kterých nebudeme pojednávat.

Nicméně nejprve je nutné zvolit pozorovací den. Pozorovací den se zpravidla volí jako začátek roku (tedy 1. leden), ale je možné zvolit první den jakéhokoli měsíce. Na tento den navazuje pozorovací doba, které je zpravidla jeden rok dlouhá. Veškeré informace o klientech získáváme v pozorovacím dni, i když k jejich selhání dojde později. Během pozorovací doby pak všechny klienty sledujeme a po skončení pozorovací doby již máme dostatek informací k modelování, tedy víme, zda klient selhal či nikoli, a u některých klientů víme, zda došlo k nápravě. Jak již bylo zmíněno výše, na pozorovací dobu může navázat dodatečná doba dlouhá několik měsíců, protože po pozorovací době zpravidla máme několik klientů, u kterých nevíme, zda dojde k nápravě nebo k nedobytnosti. Během dodatečné doby již nesledujeme selhání klientů. K modelování pak použijeme pouze informace o klientech, kteří selhali.

6.3. Ilustrace používané metody

V této části provedeme numerickou ilustraci metody popsané v části 6.2. Pro ilustraci použijeme data poskytnutá bankou, která působí na bankovním trhu v České republice. Tyto data poskytují informace o vymáhání dluhů od malých podniků a soukromých podnikatelů. Obsahují informace o daném podniku (například informace získané z účetních výkazů), o jeho úvěrové angažovanosti (tedy informace o typu úvěru a o dlužných částkách), o inkasech (tj. platbách od podniku) při použití zajištění a o inkasech bez použití zajištění. Konkrétní datová položka bude vždy vysvětlena při jejím použití.

Nejprve je nutné zvolit den pozorování a pozorovací dobu, během které budeme pozorovat selhání klientů. K dispozici máme roky selhání. Daný soubor obsahuje informace o klientech, kteří selhávali od roku 2000 do současnosti. Tedy stačí zvolit jeden rok a vybrat pozorování, u kterých došlo během tohoto roku k selhání. Tedy jako pozorovací den zvolíme první leden nějakého roku. Nejstarších pozorování (tedy s datem selhání před rokem 2005) je velmi málo k tomu, aby se na jejich základě dal odhadnout model. Nejnovějších pozorování (tedy s datem selhání po roce 2008 je nejvíce, ale tato pozorování mohou být zkreslena vlivem ekonomické recese, která započala na podzim roku 2008 v USA. Proto jako pozorovací den tedy volíme

1. leden roku 2007 a pozorovací dobu volíme jako rok 2007. Použijeme tedy ta pozorování, u kterých došlo k selhání během roku 2007 a jejich chování budeme pozorovat do současnosti, tedy do roku 2013. Znamená to tedy, že dodatečná doba je neomezená. Na základě těchto dat odhadneme model.

Z dat nejprve vybereme ta pozorování, u kterých došlo k nápravě, pozorování, u kterých došlo k nedobytnosti, a ostatní pozorování. K tomu použijeme položku aktuální PD rating. Tato položka obsahuje aktuální rating daného klienta. Klienti, u kterých došlo k nápravě, mají rating 1 – 9, nedobytní klienti mají hodnocení 12 a klienti, u kterých nevíme, co nastane, mají ostatní hodnocení (do této skupiny klientů zařadíme i klienty, u kterých není uveden PD rating). V tomto roce celkem selhalo 385 klientů, z toho u 9 došlo k nápravě a 96 klientů se stalo nedobytnými. Odsud dostáváme, že

$$\mathrm{CR} = 0{,}02338.$$

Nejprve odhadneme $\mathrm{LGD}_{\mathrm{cured}}$. Pro modelování této části potřebujeme i informace, které nejsou obsaženy v datech. Jednou z nich je informace o nákladech sledovacího oddělení na klienty, kteří selhali a my stále nevíme, co u nich dále nastane, tedy $\mathrm{W}_{10/11}$. Připouští se, vzít celkové náklady daného oddělení a ty pak rovnoměrně rozdělit mezi jednotlivé klienty. Nicméně ani tuto informaci nemáme. Máme sice k dispozici externí náklady, ale to nejsou náklady na sledování klientů. Proto zafixujeme náklady příslušného oddělení na hodnotě 50 000 000. Ty pak rozpočítáme mezi klienty, u kterých nevíme, zda došlo k nápravě nebo k nedobytnosti a pak tyto náklady u těchto klientů sečteme. Celkem dostáváme:

$$\mathrm{W}_{10/11} = 3\,636\,364.$$

Dále zafixujeme celkovou ztrátu u klientů, u kterých nevíme, co dále nastane na hodnotě 120 000 000, tedy:

$$\mathrm{L}_{10/11} = 120\,000\,000,$$

protože v tomto případě nic nemusíme rozpočítávat. Dále vezmeme položku zůstatek angažovanosti při selhání, tedy dlužné částky v okamžiku selhání. Z těchto částek vybereme ty částky, které se týkají klientů, u kterých nevíme, co nastane, a všechny tyto částky sečteme. Tím získáme celkovou hodnotu těchto dlužných částek, tedy

$$\mathrm{A}_{10/11} = 415\,857\,860{,}93.$$

Dále předpokládejme, že nedocházelo k dohodám ani prodejům, tedy

$$\mathrm{L}_{\mathrm{settlement/sale}} = 0, \mathrm{T}_{\mathrm{until\ settlement}} = 0,$$

a tedy

$$\mathrm{T}_{\mathrm{until\ curing}} = \mathrm{T}_{\mathrm{until\ curing\ (excl\ settlement)}}.$$

K výpočtu $T_{\text{until curing}}$ použijeme zůstatky angažovanosti při selhání jako příslušné váhy, rok selhání 2007 a rok, kdy došlo k poslední změně statutu, který budeme považovat za okamžik nápravy. Odsud dostaneme příslušnou hodnotu

$$T_{\text{until curing}} = 0{,}4244.$$

Každý klient má nějaký status. O statutech ještě později pojednáme. Nyní již máme vše potřebné k odhadu LGD_{cured}. Z dosud spočítaných hodnot dostáváme:

$$IC_{\text{cured}} = 0{,}03711,$$

$$L_{\text{cured}} = 0{,}12247.$$

Odsud

$$LGD_{\text{cured}} = 0{,}15959.$$

Dále odhadneme model pro modelování PC. K tomu použijeme logistickou regresi. Z výše uvedeného je zřejmé, jak bude vypadat vysvětlovaná proměnná. Z dat, která máme k dispozici, použijeme bez výběru algoritmem Stepwise následující informace (uvádíme i s případnou kategorizací, převodem na dummy proměnné, resp. s uvedením toho, které kategorie použijeme, a s označením, které budeme nadále používat):

- informace o tom, zda má klient angažovanost u jiné banky. Jedná se tedy o informaci o úvěrech u jiných bank. Tato proměnná je kategoriální se třemi kategoriemi (ano/ne/nevíme), Nadále budeme používat kategorie ano a ne (značíme $X_{1,a}$ a $X_{1,b}$);
- aktiva společnosti dle posledních účetních výkazů. Tuto proměnnou kategorizujeme podle pravidla: pokud je hodnota aktiv menší než jedna nebo není uvedena, jedná se o první kategorii, jinak se jedná o druhou kategorii. Budeme používat první kategorii (nadále X_2);
- tržby společnosti dle posledních účetních výkazů. Tuto proměnnou kategorizujeme stejně jako předchozí proměnnou. Stejně jako u předchozí proměnné použijeme první kategorii (nadále X_3);
- vlastní jmění společnosti dle posledních účetních výkazů. Tuto proměnnou kategorizujeme stejně jako předchozí dvě proměnné. Použijeme první kategorii (nadále X_4);
- informace o tom, zda je klient v NSS, tato proměnná je kategoriální se dvěma kategoriemi, přičemž 1 znamená ano, 0 znamená ne. Odsud je zřejmé, kterou proměnnou použijeme (nadále X_5);
- rating klienta před zařazením do ohrožených úvěrů, tedy před tím, než byl klient prohlášen za rizikového. Má – li klient rating od 1 do 9, nachází se v první kategorii, pokud má rating 10 – 12, nachází se ve druhé kategorii, jinak se nachází ve třetí kategorii. Použijeme druhou a třetí kategorii (nadále $X_{6,a}$ a $X_{6,b}$);

– status klienta. Klient může mít status od 2 do 9, přičemž 2 znamená mimosoudní řešení, 3 znamená soudní řešení, 4 znamená vrácení případu do předchozího segmentu, 5 znamená ukončeno vymáhání, 6 znamená plně dobrovolně splaceno, 7 znamená postoupeno, 8 znamená vrácení případu do předchozího segmentu s podmínkou, 9 znamená soudní řešení s vymožením po odpisu do podrozvahových položek. Tuto proměnnou rozdělíme do tří kategorií: první kategorie znamená statut 2, 5 nebo 6, druhá kategorie znamená status 3, 7 nebo 9, třetí kategorie znamená status 4 nebo 8. Použijeme první dvě kategorie (nadále $X_{7,a}$ a $X_{7,b}$).

Odhad příslušných parametrů uvádí tabulka 22.

Prom.	Odhad	p – hodnota
Konst.	0,23688	0,93212
$X_{1,a}$	0,49076	0,70356
$X_{1,b}$	−0,53471	0,70529
X_2	22,12085	0,99942
X_3	−1,05664	0,99997
X_4	−22,3174	0,99803
X_5	0,51703	0,74878
$X_{6,a}$	−18,0036	0,99902
$X_{6,b}$	0,30635	0,84692
$X_{7,a}$	−2,54619	0,15018
$X_{7,b}$	−21,0717	0,98997

Tabulka 22 – odhad PC modelu

Pro takto odhadnutý model provedeme test dobré shody. Při dělení na 10 skupin dostáváme Hosmer – Lemeshowovu statistiku rovnu

$$C = 1{,}56457,$$

tedy příslušná p – hodnota je

$$p = 0{,}99158,$$

tedy nezamítáme hypotézu a model považujeme za dobrý. McFaddenův koeficient determinace je

$$R^2 = 0{,}44178.$$

Giniho koeficient pro tento model je

$$\mathrm{GINI} = 0{,}90485,$$

tedy model lze považovat za perfektní.

Při odhadování LGD_{irr} nejprve odhadneme položku $C_{irrecoverable}$, která je při použití jakéhokoli přístupu stejná. Nejprve spočítáme všechny potřebné hodnoty. Stejně, jako při výpočtu LGD_{cured}, některé hodnoty zafixujeme, protože data neposkytují dané informace. Proměnné $W_{10/11}$, $L_{10/11}$, $A_{10/11}$, $L_{settlement/sale}$ a $T_{until\ settlement}$ jsou stejné, jako výše. Náklady na vymáhání u nedobytných klientů data neposkytují, a proto je zafixujeme na hodnotě 10 000 000, tedy

$$W_{12} = 10\,000\,000.$$

Hodnotu A_{12} spočítáme stejným způsobem, jako hodnotu $A_{10/11}$. Dostaneme

$$A_{12} = 520\,365\,546{,}01.$$

Dále

$$T_{until\ irrecoverable\ (excl\ settlement)} = T_{until\ irrecoverable}.$$

Hodnotu $T_{until\ irrecoverable}$ spočítáme stejně jako $T_{until\ curing}$ a dostaneme

$$T_{until\ irrecoverable} = 1{,}6911.$$

Nakonec předpokládejme, že banka nedobytné klienty vyřídí za jeden rok, tedy položme

$$T_{in\ irrecoverable} = 1.$$

Odsud dostáváme:

$$IC_{until\ irrecoverable} = 0{,}14788,$$

$$L_{until\ irrecoverable} = 0{,}488,$$

$$C_{in\ irrecoverable} = 0{,}01922,$$

$$L'_{settlement/sale} = 0.$$

Odsud

$$C_{irrecoverable} = 0{,}6551.$$

Dále odhadneme všechny míry návratnosti. Pro míru R_{EAD} potřebujeme pozorovanou hodnotu parametru EAD, kterou dostaneme použitím zůstatků angažovanosti při selhání. Dále potřebujeme expozici v domáhacím dni, kterou dostaneme odečtením hodnot inkas banky od pozorované hodnoty EAD. Tuto hodnotu pak budeme diskontovat při použitím úrokové míry 5% a rozdílu v letech, ke kterým použijeme nejpozdější datum inkasa a datum selhání, které použijeme jako domáhací den. Pak spočítáme příslušné hodnoty $R_{EAD,obs}$ a minimalizací formule (6.3) dostaneme

$$R_{EAD} = 0{,}40777.$$

Dodejme jen, že k minimalizaci formule (6.3) stačí položit derivaci dané formule rovnu nule a vyřešit vše jako rovnici. Tím získáme bod, ve kterém má daná formule minimum, protože se jedná o konvexní kvadratickou rovnici.

Při výpočtu $R_{secured}$ zjednodušujeme situaci tím, že nebudeme rozlišovat druh zajištění, tedy

$$R_{secured} = R_{secured,asset} = R_{secured,security}.$$

Stačí tedy vzít hodnoty zajištění, které data poskytují ve zvláštní kolonce, dále vzít hodnoty inkasa při zajištění, výše definovanou úrokovou míru a rozdíly v letech spočítané stejně, jako výše. Stejným postupem jako výše dostaneme

$$R_{secured} = 0{,}95963.$$

Již je zjevné, jak spočítat $R_{unsecured}$, která vychází

$$R_{unsecured} = 6{,}06619 \cdot 10^{-44},$$

i jak zjistit R_{claim}, která je

$$R_{claim} = 0{,}31437.$$

Veškerá data získáme stejným způsobem jako výše, tedy použijeme roky inkas, inkasa a zůstatky angažovanosti.

Pak již lze odhadnout LGD_{irr}. Tato hodnota bude pro každého klienta jiná (s výjimkou třetího výkladu), a proto žádnou hodnotu nebudeme uvádět. Odhad LGD_{irr} nebudeme testovat ani kalibrovat, vše provedeme u smíšeného modelu. Zároveň nebudeme kalibrovat PC model, protože bychom ke kalibraci použili stejná data jako k odhadu, a tedy by kalibrace neměla smysl. Před odhadem celkového modelu ještě spočítáme kalibrační cíl a hodnoty, se kterými budeme porovnávat výsledky smíšeného modelu. Tyto hodnoty lze spočítat jen pro napravené a nedobytné klienty. Pro ostatní klienty není možné s jistotou cokoli spočítat. Pro napravené klienty je nejprve nutné určit ztrátu. Předpokládejme, že pro napravené klienty je celková ztráta rovna 2 000 000, ostatní hodnoty již známe, tedy LGD_{obs} pro napravené klienty bude pro všechny stejné ve výši

$$LGD_{obs} = 0{,}20538.$$

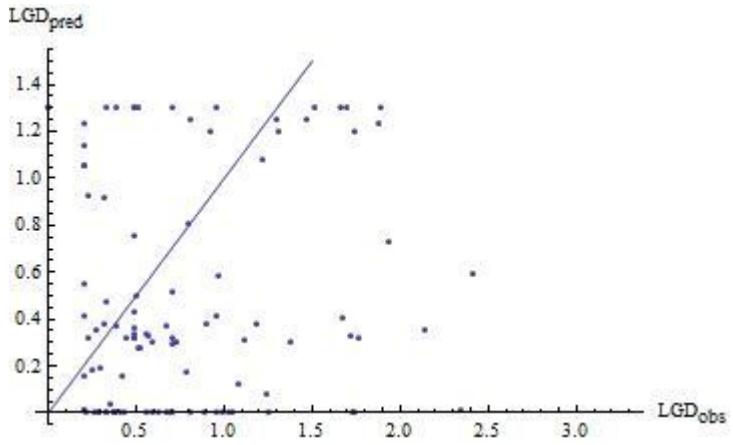
Pro výpočet LGD_{obs} pro nedobytné klienty je opět nejprve nutné spočítat danou ztrátu. Předpokládejme, že tato ztráta je celkově rovna 150 000 000, po jejím rovnoměrném rozdělení mezi klienty dostaneme ztrátu u jednoho klienta ve výši 1 562 500. Ještě je nutné určit pozorovanou návratnost u jednotlivých klientů. V tomto případě se pojmem návratnost rozumí návratnost po domáhacím dni. Spočítat tuto návratnost je složité, proto zvolíme, že ji spočítáme podělením expozice v domáhacím dni dvanácti a následným diskontováním. Ostatní položky již známe. Odsud dostaneme hodnoty pro porovnání. Celkem dostaneme kalibrační cíl

$$\mathrm{CaTa} = 0{,}45644,$$

přičemž k výpočtu tohoto cíle jsme použili vážený průměr, kde jako váhy jsme použili příslušnou expozici, tedy zůstatky angažovanosti při selhání.

Odhadneme smíšený model a provedeme test pomocí scatterplotu a statistiky R^2_{weighted}. Použijeme všechny tři výklady používané k modelování $\mathrm{LGD_{irr}}$, tedy provedeme též srovnání těchto tří postupů. Při použití prvního výkladu a výběru dat potřebných pro test (tedy pouze dat, která souvisí s daty použitými pro výpočet CaTa) dostaneme scatterplot na obr. 16 a

$$R^2_{\mathrm{weighted}} = 0{,}14829.$$



Obr. 16 – scatterplot modelu při prvním výkladu

Při kalibraci volíme

$$\alpha = 1{,}35;\ \beta = 0{,}09.$$

Po provedení kalibrace dostáváme

$$R^2_{\mathrm{weighted}} = 0{,}302.$$

Při použití druhého výkladu dostáváme stejné výsledky jako při použití prvního výkladu. Při použití třetího výkladu dostáváme jiný model. Scatterplot tohoto modelu je na obr. 17, dále

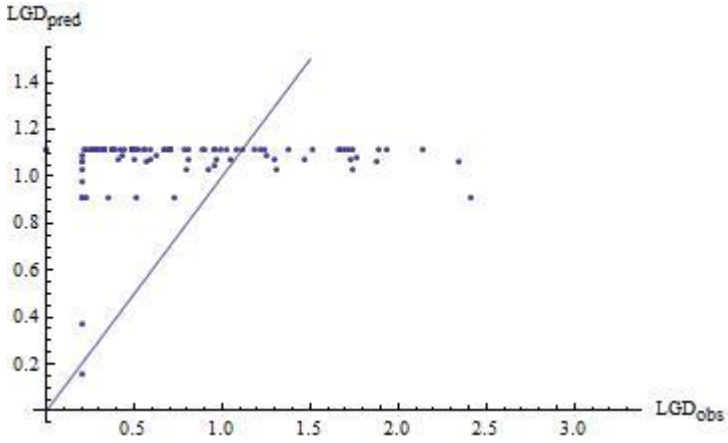
$$R^2_{\mathrm{weighted}} = 0{,}41488,$$

při kalibraci volíme

$$\alpha = 0{,}445;\ \beta = -0{,}03$$

a po kalibraci máme

$$R^2_{\mathrm{weighted}} = 0{,}56102.$$



Obr. 17 – scatterplot modelu při třetím výkladu

Ještě dodejme, proč mohou být hodnoty parametru LGD větší než jedna a záporné. Tato vlastnost velmi výrazně zhoršuje model po kalibraci, tedy znamená to, že model je po provedení kalibrace horší než před kalibrací právě kvůli této vlastnosti. Tato vlastnost navíc odporuje základní definici parametru LGD (procento z EAD, které banka ztratí při selhání). Nicméně svůj smysl má. Uvažujme klienta, který selhal. Má tedy nějaký neuhrazený dluh. Pokud klient selhal, banka má nějaké náklady na monitoring a na alespoň částečné vymožení zbývajícího dluhu. Většinou klient zaplatí nějakou část zbývajícího dluhu. Může se však najít klient, který selhal, pak u něj došlo k nápravě a banka od něj bez větších komplikací vymohla nejen všechny peníze, které klient dlužil, ale klient zaplatil i nějakou část nákladů, které banka měla na monitoring a vymáhání. Tato možnost nejčastěji nastává při exekucích. Právě náklady způsobují, že obvykle

$$\mathrm{LGD}_{\mathrm{cured}} \neq 0,$$

i když klient vše splatil. Uvažujme naopak klienta, který selhal a banka od něj nevymohla žádnou část jeho dluhu ani pomocí exekuce, například pro nulovou hodnotu klientova majetku. V tom případě klient nic nezaplatil a banka navíc měla náklady, které rovněž nikdo nezaplatí. V tom případě je pak ztráta ještě vyšší, než dlužná částka, a tedy

$$\mathrm{LGD} > 1.$$

V extrémním případě může takový klient dlužit již velmi málo a náklady banky mohou být vyšší než dluh klienta. To pak odůvodňuje případy, kdy

$$\mathrm{LGD} \in (2; 3).$$

Tyto případy však kalibrace vylučuje, a tedy model před kalibrací je lepší než model po kalibraci.

6.3.1. Meziroční srovnání výsledků modelu

V této části použijeme data z let 2005, 2006, 2008, 2009, 2010 a 2011, na jejich základě odhadneme hodnoty parametru LGD pro klienty, kteří v těchto letech selhali, a výsledky navzájem srovnáme. Použijeme při tom výše odhadnuté hodnoty všech měr návratnosti i odhadnutý PC model. Hodnoty, které jsme výše zafixovali pro rok 2007, ponecháme fixní na stejných hodnotách jako pro rok 2007, tedy:

- nedocházelo k prodejům ani dohodám;
- náklady na sledování všech klientů, kteří selhali, jsou 50 000 000 za rok;
- náklady na vymáhání u nedobytných klientů jsou 10 000 000 za rok;
- celková ztráta u napravených klientů je 2 000 000, celková ztráta u nedobytných klientů je 150 000 000, celková ztráta u klientů, u kterých nevíme, co nastane, je 120 000 000.

Při odhadu LGD_{irr} použijeme první výklad. Vždy nejprve odhadneme příslušnou hodnotu parametru LGD pro každého klienta, pak spočítáme příslušný kalibrační cíl, odhad kalibrujeme a otestujeme a nakonec srovnáme na základě mediánu, průměru a odhadu rozptylu. Výsledky srovnání jsou v tabulce 23, která uvádí kalibrační konstanty, $R^2_{weighted}$ po kalibraci, počet selhání v daném roce a zmíněné ukazatele pro všechny výše zmíněné roky i pro rok 2007 spočítané z odhadů ze zkalibrovaného modelu.

Rok	Kalibrační konstanty		$R^2_{weighted}$	Počet selhání	Průměr	Medián	Rozptyl
	α	β					
2005	1	0	0,62407	80	0,85281	1	0,11561
2006	30	0,175	0,40075	172	0,84869	1	0,09171
2007	1,35	0,09	0,302	385	0,79161	1	0,10557
2008	0,85	0	0,4873	741	0,57195	0,73535	0,08617
2009	1	0	0,17314	1312	0,5211	0,66924	0,07225
2010	1	0	0,91407	1124	0,54018	0,68192	0,07183
2011	7	0,11	0,89633	833	0,85503	1	0,10717

Tabulka 23 – meziroční srovnání

Z tabulky je vidět, že v posledních dvou letech máme velmi kvalitní odhady. Dále je zjevný zvyšující se počet selhání v jednotlivých letech. To je způsobeno dvěma fakty. Prvním faktem je, že v roce 2005 ještě neplatila smlouva BASEL II, tedy ještě oficiálně neexistoval IRB přístup, a proto data z tohoto roku nejsou zcela přesná ani úplná, totéž platí pro rok 2006. Nárůst po roce 2008 je způsoben mezinárodní ekonomickou krizí, následný pokles v roce 2011 je způsoben jednak postupným odezníváním ekonomické recese a též faktem, že data jsou velmi nová (v době

sepisování práce necelé dva roky stará). Není proto možné mít z tohoto roku kompletní data. Dle průměru se však zlepšuje vymáhání dluhů, neboť průměr postupně klesá. Nárůst v roce 2011 je spíše způsoben větším množstvím neuzavřených případů, což plyne z faktu, že data jsou velmi nová. V mnoha případech však došlo k tomu, že banka ztratila vše. To plyne z hodnot mediánů, které jsou prostředním prvkem v setříděném výběru daných hodnot. Dané rozptyly pouze značí přesnost daného průměru, tedy značí, že data nejsou od průměru příliš odchýlena.

7. Parametr EAD: expozice v selhání

V této části postupujeme podle [1].

Za parametr EAD se zpravidla považuje dlužná částka v okamžiku selhání, ale připouští se i jiná volba dané expozice. BASEL II uvaluje na parametr EAD několik požadavků. Především je nutné shromáždit data o expozici v selhání a vytvořit z nich databázi. Vždy se musí dbát na opatrnost. Rovněž je nutné vzít v potaz podnikatelský cyklus získaných dat. Nutný je též soulad s definicí selhání dle BASEL II. Vždy je potřeba odhadnout CCF (takzvaný kreditní konverzní faktor, z angl. Credit Conversion Factor). V případě rozvahových položek musí platit:

$$\min \mathrm{EAD} = \text{expozice}.$$

U pod – rozvahových položek musí parametrem odhadnout možný přibližný počátek objemu peněz při selhání a též následující vývoj po selhání. Dále se musí uvažovat dlouhodobý průměr váženého selhání s bezpečnostním rozdílem a to v případě, že je nenulová korelace mezi parametry PD a EAD. Při sběru dat je potřeba projít dlouhou historii. V případě korporátního portfolia je nutné sebrat data z posledních sedmi let, u retailového portfolia z posledních pěti let.

Existují dva typy parametru EAD. První typ je spojený s chování klienta. Vždy se ptáme, zda se klient bude držet nějakého směru, nebo zda bude dodržovat své sliby. V tomto případě se musí odhadnout CCF a sice pro jakýkoli typ produktu. Tabulka 24 uvádí příklad odhadů CCF při použití základního přístupu.

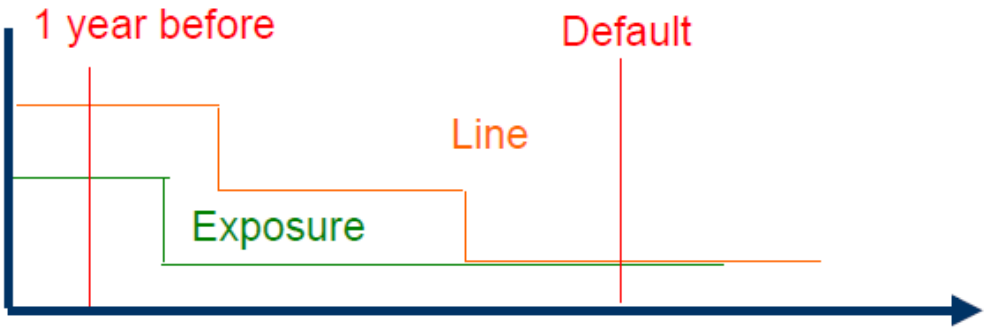
CCF	Typ produktu
0%	– závazky, které lze zrušit bez předchozího upozornění
20%	– krátkodobé možnosti, které byly obchodně uzavřené a mohou se sami splatit (např. písemný úvěr, který je zajištěný nějakým podkladovým aktivem) – závazky se splatností menší než 1 rok
50%	– transakce spojené s možnostmi (např. splnění obligace) – závazky se splatností větší než 1 rok
100%	– přímé úvěrové náhražky – prodané a zpětně zakoupené smlouvy – prodaná aktiva pomocí forwardu – zajišťovací půjčky

Tabulka 24 – příklad odhadů CCF v základním přístupu

V tomto případě platí:

$$\mathrm{EAD} = \text{současné nebezpečí} + \mathrm{CCF} \cdot \text{nevyčerpaná část úvěrové čáry}.$$

Úvěrová čára v časovém horizontu jednoho roku před selháním je znázorněna oranžově na obr. 18. Na tomto obrázku je též ve stejném časovém horizontu zeleně znázorněna úroveň nebezpečí. Z tohoto obrázku plyne, že při modelování parametru EAD nás též zajímá minulost, tedy se musíme dívat na okamžik selhání a na dobu po selhání, ale i na dobu před selháním, zatímco v případě modelování parametru LGD nás zajímá pouze doba po selhání.



Obr. 18 – úvěrová čára a úroveň nebezpečí v horizontu jednoho roku před selháním

Druhým typem parametru EAD je EAD spojený s instrumenty finančních derivátů. Jsou to funkce vývoje jejich rizikových parametrů (např. úrokové míry apod.). S finančními deriváty souvisí účetní metoda MtM (Mark – to – Market), pomocí které se oceňuje daná finanční transakce fair hodnotou. Hodnota MtM může být kladná nebo záporná. Pokud je kladná, pak je hodnota transakce kladná a transakce je v penězích, naopak, pokud je MtM záporné, pak je hodnota transakce záporná a transakce je mimo peníze. V případě OTC derivátů (tj, derivátů, obchodovaných přes přepážku) je samotné kreditní riziko funkcí dvou hlavních komponent:

- současné kreditní riziko, které je rovné současnému MtM,
- potenciální budoucí kreditní riziko nebo nějaká jiná přídavná funkce, což závisí na možném budoucím vývoji MtM.

Obr. 19 – složení celkového kreditního rizika

Celkové kreditní riziko pak je

$$\text{celkové kreditní riziko} = \text{současné kreditní riziko} + \text{přídavná funkce}.$$

Obr. 19 toto složení znázorňuje graficky v závislosti na zbytkové době trvání daného derivátu, přičemž za přídavnou funkci volíme možnou hodnotu MtM v budoucnosti. Modelování budoucího odrazu vývoje kreditního rizika OTC derivátů je spojeno s tržním rizikem. Modeluje je v závislosti na budoucí volatilitě faktorů tržního rizika, které jsou základní pro daný finanční derivát a kontrakt, který se tímto derivátem uzavírá. Je to však „věštění z křišťálové koule“, tedy nepřesná a nespolehlivá záležitost. Proto se používají takzvané kreditní ekvivalenty, v jejichž souvislosti se používají regulatorní přídavné funkce. Jejich použití demonstruje následující příklad.

Příklad

Zbývající splatnost	Úroková míra	Zlato a míra pro výměnu	Čisté jmění	Drahé kovy	Ostatní komodity
≤ 1 rok	0,0%	1,0%	6,0%	7,0%	10,0%
$1 - 5$ let	0,5%	5,0%	8,0%	7,0%	12,0%
≥ 5 let	1,5%	7,5%	10,0%	8,0%	15,0%

Tabulka 25 – hodnoty regulatorní přídavné funkce v závislosti na zbývající splatnosti derivátu a druhu podkladového aktiva

Tabulka 25 udává hodnoty regulatorní přídavné funkce. Banka uzavřela tříletý úrokový swap (tj. podkladové aktivum je úroková míra) s jinou bankou o hypotetické ceně 1 000 Euro. Tržní hodnota swapu v současnosti je 10 Euro. Buď

$$\text{PFE} = 1\,000\ \text{EUR} \cdot 0{,}5\%.$$

Dále máme

$$\text{hodnota MtM} = 10\ \text{EUR}.$$

Pak kreditní ekvivalent je

$$10\ \text{EUR} + 1\,000\ \text{EUR} \cdot 0{,}5\% = 15\ \text{EUR}.$$

□

Hodnota PFE z příkladu se obecně počítá podle vzorce:

$$\text{PFE} = \sum_{i} \text{N}_i\ \text{CCF}_i, \tag{7.1}$$

kde N_i je nominální cena i – tého kontraktu a CCF_i je příslušná hodnota kreditního konverzního faktoru, přičemž lze použít i hodnoty regulatorní přídavné funkce. Kreditní ekvivalent pak lze použít jako odhad parametru EAD nebo jako odhad celkového kreditního rizika.

Chce – li banka omezit kreditní riziko ekonomických derivátů, může použít několik faktorů, které dané riziko zmírní (takzvané tišící faktory). Jedním z nich jsou deriváty řízené managementem, což jsou cenné papíry vydané za účelem pokrytí čistého MtM nad danou úrovní. Dále je možné použít takzvané zvenčí obalené deriváty síťovinou. Znamená to vyrovnání pohledávek a pasiv v případě selhání protistrany. Též lze zrušit celé ujednání, což je oboustranná podmíněná možnost zrušit danou smlouvu před datem vypršení. Nejčastější možností je vyrovnání pohledávek a pasiv v případě selhání protistrany. Použití této možnosti je vždy omezeno regulátorem, který stanovuje pravidla pro takové očištění. Při tomto vyrovnání se pak modeluje nebezpečí selhání. K tomu je nutné spočítat kreditní ekvivalent a PFE. K tomu je nutné upravit PFE spočítanou dle (7.1), označme to jako PFE(bez očištění). Nechť

$$\text{PFE(bez očištění)} = 1.$$

Pak očištěné PFE (označme PFE(s očištěním)) je

$$\text{PFE(s očištěním)} = 0{,}4 + 0{,}6\ \text{NGR},$$

kde

$$\text{NGR} = \frac{\text{očištěné MtM}}{\text{součet kladných hodnot MtM}},$$

přičemž

$$\text{očištěné MtM} = \max\left\{0; \sum_i \text{MtM}_i\right\},$$

kde MtM_i je MtM i – tého kontraktu. Je – li

$$\text{PFE(bez očištění)} \neq 1,$$

pak

$$\text{PFE(s očištěním)} = (0{,}4 + 0{,}6\ \text{NGR})\ \text{PFE(bez očištění)}.$$

Toto počítání demonstruje následující příklad.

Příklad

Banka uzavřela tři kontrakty. Chceme spočítat kreditní ekvivalent. Vše, co je nutné vědět, obsahuje tabulka 26.

Číslo kontraktu	Hypotetická cena	CCF	MtM
1	1 000 EUR	1,0%	+100 EUR
2	2 000 EUR	5,0%	–30 EUR
3	3 000 EUR	6,0%	–40 EUR

Tabulka 26 – soupis nutných informací o kontraktech

Z tabulky plyne:

$$\text{součet kladných hodnot MtM} = 100 \text{ EUR},$$

$$\text{očištěné MtM} = 100 - 30 - 40 = 30 \text{ EUR}.$$

Odsud

$$\text{NGR} = \frac{30}{100} = 0{,}3.$$

Dále

$$\text{PFE(bez očištění)} = 1000 \cdot 1\% + 2000 \cdot 5\% + 3000 \cdot 6\% = 290 \text{ EUR}.$$

Tedy

$$\text{PFE(s očištěním)} = (0{,}4 + 0{,}6 \cdot 0{,}3) \cdot 290 \text{ EUR} = 168{,}2 \text{ EUR}.$$

Dále z tabulky plyne, že současná expozice (neboli současné kreditní riziko, resp. hodnota MtM) je 30 EUR. Odsud

$$\text{kreditní ekvivalent} = 30 \text{ EUR} + 168{,}2 \text{ EUR} = 198{,}2 \text{ EUR}.$$

□

Méně používanou možností jsou deriváty řízené managementem. Jejich použití má několik důsledků. Denně se počítá MtM. Denní marže se vyhlašují v případě, kdy čistá současná hodnota MtM je vyšší než předem daná hranice, která je obvykle nulová. Nulovost hranice se odůvodňuje omezením kapitálové spotřeby. Za deriváty řízené managementem lze považovat pouze peněžní deriváty a deriváty na obligace vydané vládami v OECD (viz též [15]). Přídavné funkce se v tomto případě nepoužívají příliš striktně, používají se pouze jednoměsíční přídavné funkce. Nejméně používanou možností je zrušení celé smlouvy. V tom případě je velmi důležitá přídavná funkce. Vůči této funkci se pak uvažuje celková splatnost daného aktiva. Důležité je též datum rozvázání smlouvy, protože k tomuto datu vešker expozice v selhání končí. Situaci znázorňuje obr. 20.

Obr. 20 – schéma rozvázání smlouvy

Závěr

Tato práce komplexně představuje IRB přístup. Pojednává o parametrech, které banky modelují na základě vlastních dat pomocí regulátorem schválených metod a postupů. Krátce pojednává o modelování parametrech M (splatnost) a EAD (expozice v selhání). Dále srovnává přístupy k modelování parametru PD (pravděpodobnost selhání klienta). Věnuje se též problému víceletého PD. Pojednává o parametru LGD (ztráta banky při selhání klienta) a o praktické metodě používané k jeho modelování. Pojednává též o logistické regresi a o diverzifikačních schopnostech skóringových funkcí.

Práce odhaduje model binární logistické regrese, který se používá při odhadu parametru PD, při použití reálných dat a na základě dat k němu navrhuje alternativní modely. Návrhy modelů provádí úpravou, přidáním nebo odebráním proměnných. Jedním z výsledků je srovnání odhadnutého modelu používaného pro modelování parametru PD při použití spojitých a dummy proměnných. Práce ukazuje, že odhad modelu při použití spojitých proměnných má horší vlastnosti než odhad modelu při použití dummy proměnných. Dále práce srovnává Altmanův model s modelem binární logistické regrese. V rámci numerické ilustrace používané metody k modelování parametru LGD též srovnává použití různých přístupů k odhadu jedné ze složek parametru LGD. Srovnává též meziroční výsledky parametru LGD od roku 2005 do roku 2011. Teoretickým výsledkem je návrh „useknutého“ geometrického rozdělení, které je součástí pojednání o problému víceletého PD. V rámci pojednání o binární logistické regresi též připomíná postup k odhadu modelu binární logistické regrese včetně odvození věrohodnostních rovnic.

Práce představuje různé přístupy k modelování některých parametrů. Je vhodná pro banky, které tyto parametry musí odhadovat. Práce ukazuje, které přístupy jsou vhodné a které nikoli. Navrhuje též odhady modelů a některé alternativy. Některé navrhnuté modely lze v praxi využít. Z hlediska celkového IRB přístupu je zajímavé pojednání o parametru M a aplikace střední doby splatnosti na jeho odhad.

Práci by prospěla přesnější data. Při numerické ilustraci metody používané k modelování parametru LGD se práce potýká s problémem nedostupnosti dat. Jejich dostupnost by numerickou ilustraci zlepšila. Též eliminace chybějících hodnot v rámci odhadu modelu pro parametr PD by daný odhad vylepšila.

Seznam použité literatury

[1] Balthazar, L.: Credit Risk Internal Rating Based Approach http://www.scribd.com/doc/27119285/Credit-Risk-Irb-Approach2

[2] Dluhošová, D., Zmeškal, Z.: Metodologie Z – metrics a vývoj ratingových modelů dle Altmana. http://www.ekf.vsb.cz/miranda2/export/sites-root/ekf/konference/en/okruhy/frpfi/Conference_proceedings/prispevky_plne_verze/Dluhosova.Dana.pdf

[3] Hosmer, D. W., Lemeshow, S.: Applied Logistic Regression, 2. vydání. John Wiley & Sons, New York 2000

[4] Hosmer, D. W., Lemeshow, S.: A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics A10:1043-1069, 1980

[5] Konečný, A.: Aplikace Altmanova modelu a Indexu bonity na telekomunikační odvětví v ČR za účelem srovnání jejich vypovídací schopnosti http://www.econ.muni.cz/~alois/publikace/KonecnyAplikace%20Altmanova%20modelu%20a%20Indexu%20bonity.pdf

[6] Krýcha, D.: Ekonomické scénáře v pojišťovnictví http://www.nfvp.cz/res/data/000150.pdf

[7] Mandl, P., Mazurová, L., Justová, I.: Matematika a řízení rizik 2009/10. Matfyzpress, Praha 2010

[8] Mc Cullagh, P., Nelder, J.A.: Generalized Linear Models, 2. vydání. Chapman and Hall, London 1989

[9] Mejstřík, M., Pečená, M., Teplý, P.: Základní principy bankovnictví. Karolinum, Praha 2009

[10] Vasicek, O.: Probability of loss on loan portfolio. Research report KMV 1987

[11] Informační bulletin České statistické společnosti, ročník 20, číslo 4, prosinec 2009. http://www.statspol.cz/bulletiny/ib-2009-4.pdf

[12] ŠKODA Výroční zpráva 2011.

https://or.justice.cz/ias/ui/vypis-sl.pdf?subjektId=isor%3a19710&dokumentId=B+332%2fSL106%40MSPH&partnum=0&variant=1&klic=x64btv

[13] http://www.bis.org/bcbs/about.htm

[14] http://www.bis.org/bcbs/qis/qis3results.pdf

[15] http://www.oecd.org/