Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Scoring a vyhodnocení bonity klienta

Plzeň, 2012 **Jana Tíkalová**

vložit originál zadání !!!!!

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 26. června 2012

Jana Tíkalová

Poděkování

Děkuji především pracovníkům CRM ze společnosti, která mi poskytla data pro tuto práci a cenné rady ohledně jejich sběru a ekonomickém pohledu na scoring, a vedoucímu diplomové práce Ing. Patrice Markovi Ph. D. zejména za jeho trpělovist a rady především v oblasti statistiky.

Abstrakt

Tato práce pojednává o vyhodnocování bonity klientů žádajících o poskytnutí úvěru, především se zaměřením na scoring jako základní formu vyhodnocování bonity klienta a úvěrového rizika poskytovatele úvěrů. Scoring je metoda vybudovaná na vyhodnocování údajů o klientovi a údajů o úvěrovém produktu, o který žádá. Cílem scoringu je na základě daných informací posoudit, zda žádající klient je dostatečně bonitní a riziko spjaté s poskytnutím úvěru je tak pro úvěrovou instituci únosné. Scoring je založen na historických datech a statistických metodách, pomocí nichž je predikováno riziko u jednotlivých nových žadatelů.

V této práci budou prezentovány statistické metody, na kterých je scoring vystavěn a taktéž jeho ekonomický význam pro poskytovatele úvěru.

Klíčová slova

úvěrové riziko, poskytovatel úvěru, scoring, úvěruschopnost, bonitní klient, default klienta, logitová regrese

Abstract

This diploma thesis deals with evaluation creditworthiness of clients who apply for a credit, especially this thesis focuses on scoring as the basic form of evaluation of clients bonity and credit risk of credit grantor. Scoring is method built on evaluation of data about clients and data about credit product, which is applied for. The aim of scoring is, on the basis of existing data, to recognize if applying client has good standing and the risk regarding credit grant is profitable (tolerable). Scoring is based on historical data and statistic methods, which are able to predict risk in cases of each new applicant.

In this thesis there will be statistic methods, which the scoring is built on, and moreover economic significance of scoring will be presented.

Key Words

credit risk, credit grantor, scoring, creditworthiness, clients bonity, clients default, logistic regression

Obsah

Kapitola 1

Úvod

Cílem této diplomové práce je vyhodnocení bonity klienta ve vztahu k jeho úvěrovým závazkům. Práce je zpracovávána z pohledu poskytovatele úvěru [1] a jemu dostupných dat, nikoliv z pohledu samotného klienta. Jako prostředek sloužící k naplnění uvedeného cíle byl zvolen scoring.

V teoretické části je popsána obecně teorie úvěrového rizika, kterému čelí každý poskytovatel úvěru, jakožto i možnosti řízení úvěrového rizika a jeho regulatorní rámec.

V další části jsou poté vymezeny statistické nástroje, které jsou používány při sledování úvěrového rizika, zejména scoringu. Pozornost je věnována především logitové regresi, která je k modelování scoringu využívána nejčastěji.

V praktické části je nejprve stručné seznámení s poskytovatelem dat k této práci a jeho řízením úvěrového rizika, na základě čehož je vymezen cíl práce konkrétněji, především s přihlednutím k jeho očekáváním a požadavkům. V samotné aplikační části je poté vyhodnocen užívaný model scoringu a navrženy změny v modelu využívaném zadavatelem této práce.

Veškeré podkladové soubory a výpočty jsou přiloženy na datovém CD.

[1] S ohledem na to, že v českém právním řádu neexistuje jednotná definice poskytovatele úvěru, budeme v této práci za poskytovatele úvěru považovat obecně osobu, která půjčuje peníze za úplatu.

Kapitola 2

Rizika poskytovatele úvěru a vyhodnocení bonity klienta

V této kapitole bude proveden stručný náhled na teorie týkající se rizika poskytovatele úvěru. Cílem této kapitoly není podat ucelený výklad teorie úvěrového rizika, nýbrž jen ukázat na jeho hlavní pilíře, které si často mnozí neuvědomují.

V prvé řadě by mělo být patrné, že úvěrové riziko poskytovatele úvěru je jedním z jeho zásadních podnikatelských rizik, od kterého se odvíjí úspěch veškerých jeho podnikatelských aktivit a také jeho ziskovost. Druhý bod představuje regulatorní rámec činnosti poskytovatele úvěru, to znamená, že veškeré aktivity v této oblasti nejsou a nemohou být rozhodovány jen dle čisté vůle poskytovatele úvěru. Více podrobností k této kapitole lze dohledat v literatuře [9] až [15].

2.1 Vymezení rizika poskytovatele úvěru

Každý podnikatelský subjekt na sebe výkonem své činnosti přebírá riziko neúspěchu při naplňování svého podnikatelského záměru.[2] U poskytovatelů úvěru se vedle tohoto podnikatelského rizika objevuje ještě riziko finanční.

[2]Mezi rizika vznikající při podnikatelské činnosti se řadí jak riziko vyplývající z obchodní strategie společnosti, tak rizika provozní, která zahrnují selhání lidského či technického faktoru, podvody, přírodní vlivy a jiné neplánované události.

Finanční riziko je dle Jílka [9] riziko potencionální finanční ztráty subjektu, tj. nikoli již existující realizovaná či nerealizovaná finanční ztráta, ale ztráta v budoucnosti vyplývající z daného finančního či komoditního nástroje nebo finančního či komoditního portfolia. Finanční rizika by se měl každý podnik, nejen poskytovatel úvěru, snažit kontrolovat a aktivně řídit. Finanční riziko lze dále dělit na riziko tržní, likvidní a úvěrové.

- tržní riziko = riziko změny některého faktoru trhu - úroková míra, kurz měny, cena akcie, obligace či jiného obchodovatelného instrumentu - a dopad této změny na hodnotu aktiv, závazků nebo čistého jmění.

- likvidní riziko = riziko nedostatku volných finančních prostředků nutných k řádnému splnění splatných závazků. V bankovní teorii se dále rozlišuje i riziko solventnosti, tj. schopnosti svým kapitálem pokrýt vzniklé ztráty, respektive pokles hodnoty aktiv pod tržní hodnotu závazků. V bankovnictví musí být v této oblasti dodržována pravidla stanovena obecně závaznými právními předpisy[3].

- úvěrové riziko = riziko nesplnění závazku ze strany obchodního partnera, ať už úplného či částečného. Toto riziko je významně propojeno s rizikem likvidním i rizikem tržním.

A právě úvěrové riziko a jeho aktivní řízení by mělo být základní pracovní náplní úvěrového oddělení, tzv. Credit Risk Managmentu každého poskytovatele úvěru. Řízení úvěrového rizika tvoří taktéž předmět této práce.

2.2 Řízení úvěrového rizika poskytovatele úvěru

Řízení úvěrového rizika je u poskytovatele úvěru zpravidla úkolem specializovaného úvěrového oddělení, toto bývá zpravidla označováno zkratkou CRM z anglického Credit Risk Management. CRM identifikuje mezi úvěrovými riziky nejen riziko samotného

[3]V České republice je tato problematika upravena Vyhláškou č. 123/2007 Sb., o pravidlech obezřetného podnikání bank, spořitelních a úvěrních družstev a obchodníků s cennými papíry, ve znění pozdějších předpisů, zejména pak Vyhlášky č. 282/2008 Sb. ze dne 1. srpna 2008 [16]. Tato vyhláška provádí v českém prostředí směrnici Evropského parlamentu a Rady 2006/49/ES, o kapitálové přiměřenosti investičních podniků a úvěrových institucí, která vznikla na základě činnosti Basilejského výboru pro bankovní dohled

defaultu obchodního partnera (žadatele o úvěr), ale také riziko zajištění, riziko koncentrace, v případě účelového úvěru či leasingového financování pak také riziko zůstatkové hodnoty.

- Riziko defaultu a jeho vyhodnocení je hlavním obsahem této práce, proto se mu budeme věnovat níže podrobněji. Samotné slovo default znamená selhání, zmeškání, nedodržení závazků. V teorii úvěrového rizika se poté používá *default* pro označení situace, kdy klient selže v plnění svých závazků z úvěrové smlouvy, tj. zejména ve splácení, aniž by nás zajímalo, z jakého důvodu selhal.

- Riziko zajištění (collateral risk, recovery risk) se týká ocenění zajištění kryjící případné ztráty, při čemž musí být vyhodnocena likvidita a vymahatelnost jednotlivých typů zajišťovacích instrumentů.[4] Poskytovatel úvěru musí při využívání zajišťovacích instrumentů také přihlížet k daňovým aspektům jednotlivých řešení zajištění a tím optimalizovat své riziko.

- Zpravidla každá úvěrová instituce si definuje určitý kvalifikační práh svého ekonomického kapitálu (kapitál nutný ke krytí neočekávaných ztrát) a aby nedocházelo k přílišné koncentraci rizika, žádný úvěrový limit schválený pro jednoho klienta by tento práh neměl překročit. Při řešení těchto otázek hovoříme o riziku koncentrace.

- Riziko zůstatkové hodnoty je poté riziko poklesu tržní ceny předmětu, na který byl poskytnut účelový úvěr, a dále riziko jejího neuhrazení na konci financování. Toto riziko lze v praxi přenést na třetí subjekt uzavřením smlouvy o budoucí smlouvě kupní.

Řízení úvěrového rizika poskytovatele úvěru by mělo být uceleným procesem, který bude vyhodnocovat veškeré výše popsané aspekty rizika.

Úvěrové riziko taktéž významně působí na stabilitu finančního trhu, proto se jeho problematikou zaobírají i dohledové orgány finančního trhu a některé jeho aspekty jsou regulovány i legislativním způsobem, ať už na státní úrovni nebo na úrovni Evropské unie. Nejvýznamnější současný regulatorní akt je známý pod názvem Basel II.

[4]Mezi typické zajišťovací instrumenty patří zástavní právo, zajišťovací převod práva a ručení. Všechny typy zajišťovacích instrumentů, které lze užít v České republice ke snížení rizika poskytovatele úvěru, jsou vymezeny českým právním řádem, konkrétně pak v části osmé, oddílu pátém zákona č. 40/1964 Sb., občanský zákoník, ve znění pozdějších předpisů a v části třetí, hlavě první, dílu třetím zákona č. 513/1991 Sb., obchodní zákoník, ve znění pozdějších předpisů.

2.2.1 Basel II

Snahu stanovit jednotná pravidla pro měření finančního rizika bank je možné najít již v polovině osmdesátých let minulého století. Prvotní dohoda o kapitálu z této doby nesla označení Basel I (1988) - Basel Capital Accord. Jméno nese po švýcarském městě Basel, kde sídlí Bank for International Settlements (Banka pro mezinárodní platby), jejíž součástí je i výbor pro bankovní dohled Basel Committee on Banking Supervision.

Hlavním předmětem Basel I byla kapitálová přiměřenost - výška minimálního regulatorního kapitálu byla stanovena na 8 procent z rizikově vážených aktiv (RWA), tj. aktiv banky vážených příslušnou rizikovou vahou, která odpovídá stupni úvěrového rizika konkrétního druhu aktiva. Za nedostatky této dohody bylo považováno vymezení jen dvou druhů rizik (úvěrové a rizika přesunu kapitálu mezi zeměmi), při čemž měření úvěrového rizika probíhalo velmi jednoduše pomocí rizikových vah (RW) určených pro jednotlivé státy.

V devadesátých letech se tak vyvíjela aktivita na novelizaci této dohody, výsledkem byl v roce 1999 návrh Basel II, jehož finální verze byla publikována v roce 2004. Tato dohoda vedle kvantitativních požadavků na banky stanovuje také požadavky kvalitativní. Velmi podrobně jsou v dohodě stanoveny minimální kapitálové požadavky (pilíř 1).[5] Basel II vymezuje tři způsoby měření úvěrového rizika:

- standardizovaný přístup (STA) - aplikuje se na základě koeficientů stanovených přímo regulátorem, poskytovatel úvěru tak neposuzuje riziko samostatně. Proces stanovení ratingu a rizikových vah je určen externími ratingovými agenturami, ratingy jednotlivých agentur pak podléhají schválení regulátora trhu. Rizikové váhy jsou zde oproti Basel I stanovovány kvalifikovaněji a jsou zde také vymezeny techniky na zmírnění úvěrového rizika.

- základní IRB přístup (FIRB - Foundation Internal Ratings-Based Approach) - založený na interním ratingu poskytovatele úvěru, nespornou výhodou tohoto přístupu je možnost přizpůsobení ratingu konkrétnímu produktu poskytovatele úvěru (banky), z čehož vyplývá nižší potřeba vlastního kapitálu.

[5] Dohoda se celkem skládá z pilířů tří, vedle nejvýznamnějšího prvního pilíře zaměřeného na kapitálové požadavky se v druhém pilíři věnuje postupu dohledových orgánů a ve třetím stanovuje základní pravidla chování bank v tržním prostředí, zejména sdílení informací. Více informací lze dohledat v [12] a [13].

- pokročilý IRB přístup (AIRB - Advanced Internal Ratings-Based Approach) - založený na interním ratingu poskytovatele úvěru s využitím podrobnějších analýz.

Basel II rozeznává jako základní rizikové prvky, pomocí nichž vypočítává riziko vážených expozic, pravděpodobnost selhání klienta (Probability of Default), expozice při selhání (Exposure at Default), ztráta v případě selhání (Loss Given Default) a splatnost (Maturity). Zatímco při použití interního ratingu FIRB banka používá svůj vnitřní matematický model jen pro výpočet pravděpodobnosti selhání klienta a jako ostatní údaje využívá ty stanovené regulátorem, při využití metody AIRB si banka sama určuje všechny rizikové prvky.

Jako reakce na makroekonomickou situaci v roce 2008, při níž se ukázalo, že regulace finančního sektoru není postačující, Basilejský výbor pro bankovní dohled vypracoval Basel III. Tento regulatorní prostředek rozšiřuje minimální kapitálové požadavky a zavádí nové požadavky na likviditu bank a bankovní pákový efekt. Z hlediska úvěrového rizika Basel III. hovoří o posilování úvěrových standardů a řízení rizika, spíše než na vyhodnocení defaultu klienta, na což je zaměřena tato práce, je ale zaměřen na riziko zajištění a riziko kapitálové přiměřenosti. Jeho zavedení v praxi by mohlo spíše ovlivnit obchodování s cennými papíry a možnosti jejich zajištění a hedgingu. Basel III. byl představen v roce 2011 a časový harmonogram pro jeho implementaci je stanoven až do roku 2019.

2.3 Vyhodnocení bonity klienta

Proces vyhodnocení žadatele o úvěr by měl vždy probíhat v širších souvislostech. Neměl by být vyhodnocován jen daný subjekt jako individualita, ale vždy by měly být zohledněny i globální podmínky a podmínky odvětví, ve kterém žadatel o úvěr působí. Obecně je pro poskytovatele úvěru rizikovější, mít veškeré své klienty ze stejného odvětví (zejména citlivého na hospodářský cyklus) než mít pestrou škálu klientů.

I přes výše uvedené skutečnosti nejvýznamnějším faktorem pro posouzení žadatele o úvěr zůstává individuální faktor. I zde lze vymezit obecné pravidlo, že roste úvěrové riziko s výší poskytnutého kapitálu a s délkou úvěrového kontraktu.

Při vyhodnocování úvěrového rizika daného klienta je nutné zohlednit důvody, které může mít klient k nezaplacení úvěru řádně a včas. Lze rozlišit v zásadě tři situace, kdy klient

z ekonomických důvodů nebude schopen platit, kdy neplatit je již původní postoj žadatele a kdy žadatelem je významný klient, který neplacením bude vytvářet nátlak v rámci své vyjednávací pozice.

Samotné vyhodnocení klienta pak může probíhat různými metodami. Ke klientům, jejichž podíl na úvěrovém portfoliu poskytovatele je nepatrný, lze přistupovat na základě celého portfolia klientů s obdobnými charakteristikami a vyhodnocovat je v rámci portfoliového rizika metodou nazývanou scoring. U větších klientů, jejichž angažovanost je pro poskytovatele úvěru významnou, je již nutné hodnotit žadatele o úvěr individuálně a využijeme metody ratingu. Každý poskytovatel úvěru si zpravidla stanoví hranice úvěrového rámce, který ještě spadá do nevýznamných portfoliových klientů a který naopak již musí být vyhodnocován individuálně.

2.3.1 Vymezení defaultu

Před samotným procesem vyhodnocování klienta je nutné stanovit, co rozumíme defaultem klienta. Opakem defaultního klienta je poté klient bonitní, který je schopný dostát svým závazkům řádně a včas. Každá banka či jiná úvěrová instituce může mít default nastaven jinak. Vyhláška [16] chápe default (selhání klienta) jako situaci, kdy je splněna alespoň jedna z následujících podmínek:

- lze předpokládat, že dlužník pravděpodobně nesplatí svůj závazek řádně a včas, aniž by věřitel přistoupil k uspokojení své pohledávky ze zajištění,
- alespoň jedna splátka jistiny nebo příslušenství jakéhokoliv závazku dlužníka vůči věřiteli je po splatnosti déle než 90 dnů; k této podmínce povinná osoba nemusí přihlížet, pokud částka po splatnosti není významná s tím, že práh významnosti stanoví povinná osoba s ohledem na to, jakou částku nevymáhá při odpisu pohledávky.

Vyhláška [16] dále definuje i faktory avizující, že dlužník (žadatel) nesplní své závazky řádně. Mezi tyto faktory se řadí např. zahájení insolvenčního řízení s dlužníkem, dlužník v úpadku, neuplatnění časového rozlišení u expozice (pohledávka věřitele vůči dlužníkovi), provedení úpravy ocenění expozice, veškeré faktory si můžete dohledat v [16].

Většina bank jako default klienta vyhodnocuje skutečnost, kdy se klient dostane do prodlení se splacením některého ze svých závazků o více než 90 dní, při čemž již po 30 dnech

zpravidla jedná s klientem aktivně a vybízí ho ke splnění a zaplacení. Při vyhodnocování portfolia je proto vhodné uvažovat oba tyto typy defaultu, neboť již 30denní prodlení pro banku znamená další administrativu a tedy náklady, které finanční instituce musí být schopna pokrýt.

2.3.2 Rating

V případě ratingu hovoříme o metodě vyhodnocení klienta a jeho bonity na základě individuálního přístupu. Tato metoda je využívána pro významné korporátní zákazníky. Rating lze definovat jako nezávislé hodnocení, jehož cílem je zjistit, a to na základě komplexního rozboru veškerých známých rizik hodnoceného subjektu, jak je tento subjekt schopen a ochoten dostát včas a v plné výši všem svým splatným závazkům [10]. Výsledkem ratingu je známka, která vyjadřuje míru rizika pro věřitele, respektive bonitu dlužníka a jeho schopnost dostát závazkům z úvěru.

V rámci ratingu jsou pak sledovány jednotlivé charakteristiky žádajícího subjektu jako ukazatele rozvahy, ukazatele výkazu zisku a ztrát a toky cash flow, kvalita managmentu, budoucí potenciál podniku, platební morálka a charakteristika odvětví žadatele. Vyhodnocení tak probíhá nejen s kvantifikovanými veličinami, ale také kvalitativními.

2.3.3 Scoring

Jedná se o metodu vyhodnocení bonity klienta v rámci vyhodnocení portfoliového rizika. Tato metoda je založena na historických datech o klientech obdobných charakteristik a chování (behaviorální, historické) a statistických metodách.

Data, která jsou při vyhodnocování žadatele o konkrétní úvěr posuzována, jsou data poskytnutá klientem poskytovateli (u soukromé osoby to může být např. výše příjmu, zaměstnání, věk, jiné finanční závazky, vzdělání apod.) a dále data o samotném produktu (délka kontraktu, výše úvěru, úroková míra apod.). Při vyhodnocování scoring automaticky vygeneruje klientovi hodnotu, která má značit jeho úvěruschopnost, tato hodnota je založená na posuzovaných datech v porovnání s průběhem úvěrů klientů s podobnými charakteristikami, které má již poskytovatel úvěru ve svém portfoliu. Scoring tak

dokáže prostřednictvím odhadu na základě historických dat rozdělit klienty na rizikové (defaultní) a bonitní.

Celé vyhodnocování bonity respektive defaultnosti klienta probíhá jako testování hypotéz. Testujeme hypotézu H_0, že u klienta nastane default, přesné vymezení defaultu je vždy na poskytovateli úvěru, pro naše účely defaultem rozumějme situaci, že klientova alespoň jedna splátka jakéhokoliv závazku vůči věřiteli je více jak 90 (30) dní po splatnosti, proti alternětivě H_1, že klient je bonitní a tedy žádná splátka jeho závazku vůči věřiteli není 90 (30) dní po splatnosti. Při vytváření modelu si musíme uvědomit, že mohou nastat dvě situace, které jsou pro poskytovatele úvěru nežádoucí.

- Za prvé na základě informací vyhodnotí žadatele jako bonitního, přestože on úvěr nesplatí, tedy neodhadne u něj default, ač skutečně nastane, tj. bude zamítnuta hypotéza H_0, ač nastane - **chyba I. druhu**.
- Pro poskytovatele úvěru je ale nežádoucí i situace, kdy vyhodnotí klienta jako defaultního, ač on by úvěr byl schopný splácet, tj. nezamítneme hypotézu H_0, ač neplatí - **chyba II. druhu**[6].

Na užívání scoringu lze vysledovat následující přednosti:

- snížení času nutného k vyhodnocení - funguje u poskytovatele úvěru jako zcela automatizovaný prostředek,
- zajištění rovných podmínek pro všechny se stejnými charakteristikami (nehrozí riziko obvinění z diskriminace),
- snížení nákladů banky.

I tato metoda ovšem přináší své nevýhody. Model by měl být neustále ověřovaný, při rozšíření užívání scoringu na jiné produkty je nutné posoudit, zda produkt bude určen pro stejný charakter klientů jako jsou ti, dle jejichž dat je model vytvořen. V neposlední řadě může dojít také k chybám plynoucím z technických chyb vzniklých při samotné tvorbě modelu.

[6]U poskytovatelů úvěrů, kteří vyhodnocují klienta jen na základě scoringu a nemají k němu žádnou doplňující metodu, nelze získat informaci o chybě druhého druhu, neboť když klientovi úvěr neposkytnou, nemají jak zjistit, zda by tento klient úvěr splácel. Jen prostřednictvím scoringu má většina poskytovatelů úvěru řešen přístup k soukromé osobě (spotřebiteli).

Rating a scoring bývá využíván pro různé situace, každá z metod má vedle svých výhod i slabiny, srovnání ratingu a scoringu na základě nejčastěji sledovaných parametrů je provedeno v tabulce 2.1.

parametr	**rating**	**scoring**
užití na klienta	*korporátní*	*retailový*
doba zpracování	*řád dní*	*řád minut a hodin*
výsledek	*známka dle vah jednotlivých kriterií*	*hodnota součtu za jednotlivá kritéria*
způsob vyhodnocení	*finanční analýza*	*automatizovaný proces*

Tabulka 2.1: Srovnání ratingu a scoringu

2.3.4 Důsledky pro banky a ostatní poskytovatele úvěrů

Zavedení metody credit scoringu při vyhodnocování úvěruschopnosti klientů s sebou přináší navýšení konkurence mezi poskytovateli úvěrů. Dříve banky především malým podnikatelským subjektům půjčovali své peníze jen na základě dobrých referencí přímo o tomto konkrétním subjektu, které byly schopny získat a posoudit jen pokud se nacházely ve stejné lokalitě se žadatelem, vyhodnocení nového klienta pro ně navíc znamenalo časově náročný úkon. Díky credit scoringu jsou banky schopny zpracovat žádost o úvěr prostřednictvím jejich automatizovaných a centralizovaných systémů.

Díky schopnosti posoudit klientovu bonitu rychleji se úvěrovým institucím navýšilo i jejich portfolio, aniž by se navýšila také míra neplatičů. Banky tak začaly být schopné poskytovat úvěr širší skupině subjektů, čímž došlo i k navýšení dostupnosti úvěru pro klienty. Na základě vyhodnocení credit scoringu je navíc banka schopna nabídnout dobrým klientům úvěr za zvýhodněných podmínek (snížení úroku, poplatků apod.).

Kapitola 3

Tvorba a vyhodnocení modelu scoringu

V této kapitole bude prezentován základní přístup k přípravě modelu scoringu. Od prvotního sběru dat až po metody, které se nabízí k jeho zpracování. Nejčastěji je k vyhodnocení scoringu v praxi používána metoda logitové regrese, proto její popis tvoří hlavní část této kapitoly a je taktéž následně užita i v aplikační části.

Cílem této kapitoly je taky poukázat na hlavní společné, ale naopak i odlišné rysy logitové a lineární regrese tak, aby bylo zřejmé, proč se k vyhodnocení scoringu využívá regrese logitové na úkor lineární regrese, která je všemi známá a snadno interpretovatelná.

Celá kapitola je vypracována na základě literatury [1] až [8] a [16] až [18]. Obecný přístup k logitové regresi lze dohledat a použít Šedivé [17], pro více podrobností k logitové regresi lze použít knihu Hosmera a Lemershowa [4]. K aplikační praxi logitové regrese ve scoringu byla vedle výše uvedených navíc využita příručka SAS pro zpracování scoringového modelu [19].

3.1 Přístup ke tvorbě modelu

Scoring v praxi slouží k ohodnocení úvěrového rizika, tj. k posouzení bonity klienta a jeho schopnosti úvěr splatit. Podstatou scoringové funkce je kvantifikovat tuto dlužníkovu

vlastnost. Scoringová funkce bývá vytvářena nad množinou všech možných charakteristik o klientovi, ale taktéž o produktu, o který klient žádá.[7]

Definovaným vlastnostem (kvantitativním i kvalitativním) klienta a produktu je přidělena číselná hodnota (zpravidla body). Každý klient žádající o úvěr poté na základě vyhodnocení touto funkcí získá skóre, které hovoří o jeho schopnosti splatit úvěr dle smluvních podmínek. Některé bankovní instituce skóre definují odhadem pravděpodobnosti, s jakou daný klient úvěr splatí, jiné zavádí bodovou stupnici, podle které posuzují klientovu bonitu.

Podle skóre se poté bankovní instituce rozhodují, zda žádajícímu klientovi půjčku poskytnou a případně za jakých podmínek. Žadateli s vyšším skóre může být nabídnuta nižší úroková míra, naopak klientovi s nižším skóre může být úroková míra navýšena či požadováno dodatečné zajištění, pojištění schopnosti splácet apod. tak, aby banka eliminovala své riziko z klientova defaultu.

Ještě dříve než banka přistoupí k sestavení nového modelu credit scoringu musí si stanovit jeho účel. Některá banka může chtít navýšit svoje portfolio a tedy stát se přístupnější pro žadatele úvěru (i za cenu snížení "kvality"jejích klientů), jiná naopak chce snížit delikvenci ve svém portfoliu a jiná zase může chtít navýšit rozmanitost své klientely či prosadit určitý produkt.

Významný parametr, který je ve scoringu často zmiňován a který je nutné stanovit právě podle účelu scoringu, je tzv. "cutoff", jedná se o hranici, na které se láme, kdy je klient vyhodnocen jako bonitní a kdy je u něj naopak predikován default.

Při sestavování modelu credit scoringu bývá zpravidla používán následující postup:

- získání vhodných vstupních dat a jejich úprava,
- vytvoření modelu skórovací funkce na základě statistických metod,
- interpretace výsledků poskytnutých modelem,
- vyhodnocení modelu.

[7] Více o získávání dat a jejich zpracování v následující kapitole "Zpracování dat".

3.2 Zpracování dat

Data používaná ke zpracování skóringové funkce mohou pocházet jak z vlastních historických dat dané finanční instituce, tak z externích zdrojů [8]. Při výběru dat je nutné zvolit vhodnou délku období, ze kterého budou data pocházet tak, aby bylo zajištěno stálé chování klientely v daném období, neovlivněné vznikem rozdílných podmínek. Zároveň je nutné sledovat a případně zohlednit metodiku sběru dat a její případný vývoj.

Sesbíraná data je nutné podrobit důkladné analýze. Nejprve je nutné pochopit pozadí sběru a tvorby dat. V datech se můžeme setkat s logickou chybou [9] či s chybou systematickou [10].

Mezi daty můžeme objevit různé druhy vysvětlujících proměnných, data kvalitativní povahy:

- nominální - popisují jen, zda prvek má danou vlastnost nebo nikoliv (např. vzdělání, pohlaví),
- ordinální - určují, zda daný prvek má nějakou vlastnost i jakou intenzitu dané vlastnosti prvek obsahuje, tedy stanovují i relaci uspořádání (např. věk),

která vystupují v modelech logitové regrese zpravidla jako umělé (dummy) proměnné, a data kvantitativní povahy

- spojitá - vyjádřená skutečnou hodnotou naměřené veličiny,
- intervalová - kvantitativní data jsou rozdělena do intervalů podle intenzity zkoumané vlastnosti. Tento přístup je v logitové regresi velmi často používán,
- poměrová data - vyjadřující podíl mezi dvěma veličinami, dávající společně hod-

[8] Využívané bývají tzv. Credit Bureau - úvěrové referenční agentury, které fungují jako sběrna a databáze veřejně dostupných dat či výzkumů i dat poskytovaných soukromými subjekty. V České republice je nejvýznamnější Czech Banking Credit Burreau.

[9] Jedná se o typ chyby, kdy data nabývají nereálné hodnoty - typicky se může objevit záporný čas, chyba v řádech apod. Data s logickou chybou je nutno ze souboru odstranit, je-li odhalena chyba pro celý soubor dat, např. chyba v řádech, lze data opravit na skutečné hodnoty.

[10] Chyba vznikající ze špatné metodologie sběru dat, např. nesrozumitelně položená otázka na klienta, který poté zpravidla zvolí první možnost, či neúplný soubor dat některé z proměnných z důvodu, že hodnoty této proměnné se nezaznamenávaly po celou dobu apod. Proměnné, u nichž chybí velký počet hodnot, je nutno ze zkoumaného souboru odstranit.

notu jedné závislé proměnné, nezáleží u nich na absolutní hodnotě jednotlivých proměnných vystupujících v podílu,

- různé transformace kvantitativních proměnných [11] nebo interakce mezi proměnnými.

Následně je nutné se vhodným způsobem vypořádat s odlehlými pozorováními, které mohou významně ovlivnit výsledky některých modelů. Při modelování scoringu se problém s odlehlými pozorováními zpravidla řeší rozdělením dat do intervalů (kategorií), přičemž každému intervalu se přiřadí nějaká číselná hodnota (kódování, bodování), data pak vystupují v modelu jako umělé (dummy) proměnné. Tím se data s extrémně nízkými či vysokými hodnotami skryjí do otevřených krajních intervalů a získají jemu přiřazenou umělou hodnotu, která již vůči celému pozorování není odlehlou. Pro přiblížení tohoto typu zpracování dat si v tabulce 3.1 uvedeme intervalové rozdělení jedné proměnné používané ve scoringovém modelu pro živnostníky [12].

Skutečná hodnota	**Bodování**
X je neznámé nebo $X < 50$	0
$0 \leq X < 50$	4
$50 \leq X < 150$	8
$150 \leq X < 250$	12
$250 \leq X < 500$	16
$500 \leq X$	20

Tabulka 3.1: Rozdělení spojité proměnné "Zisk"na intervaly

V dalším kroku by mělo být posouzeno, jak se jednotlivé proměnné chovají a jaké jsou mezi nimi vzájemné vztahy. Mezi veličinami mohou existovat různé formy závislosti. Šedivá [17] uvádí následující typy:

- deterministická závislost - pevná závislost, kdy výskytu jednoho jevu nutně odpovídá výskyt druhého jevu,

[11] Z důvodu odhalené závislosti mezi proměnnými, se musí některé proměnné transformovat, aby byla závislost odstraněna, nejtypičtější je logaritmická transformace. Více o závislosti proměnných dále v této kapitole.

[12] Všechny proměnné používané v tomto modelu scoringu budou představeny v následující kapitole a navíc tvoří přílohu A

- stochastická (statistická) závislost - závislost, kdy jedna veličina ovlivňuje druhou (vyskytuje se od silné po slabou).

Při vytváření statistického modelu zpravidla vycházíme z předpokladu, že jsou vstupní proměnné nezávislé, tzn. vzájemně se neovlivňují. Šedivá [17] uvádí definici nezávislých veličin následovně:
Řekneme, že X a Y jsou nezávislé $\Leftrightarrow F(x,y) = F_1(x) \cdot F_2(y)$,

kde $F(x,y)$ je sdružená distribuční funkce nezávislých náhodných veličin X a Y,
$F_1(x)$ je distribuční funkce nezávislé náhodné veličiny x,
$F_2(y)$ je distribuční funkce nezávislé náhodné veličiny y.

K popisu vzájemných vztahů mezi veličinami dále používáme koeficient kovariance (popř. korelace), namísto o závislosti potom hovoříme o korelovanosti veličin [13].
Řekneme, že X a Y jsou nekorelované $\Leftrightarrow cov(X,Y) = 0$ [14],

kde $cov(X,Y)$ vyjadřuje kovarianci náhodných veličin a je dána předpisem:

$$cov(X,Y) = E([X - E(X)] \cdot [Y - E(Y)]) \tag{3.1}$$

kde $E(X)$ značí střední hodnotu veličiny X,
$E(Y)$ střední hodnotu veličiny Y.

Dle Hlinicy [11] korelace vystihují lineární typy závislostí, proto jsou vhodné k modelování pouze tehdy, pokud jsou "přibližně"lineární. Reif [1] definuje korelační koeficient $\rho = \rho(X,Y)$, který vyjadřuje míru statistické lineární závislosti [15] veličin X, Y vztahem:

$$\rho = \frac{cov(X,Y)}{\sqrt{D(X) \cdot D(Y)}} \tag{3.2}$$

kde $D(X)$ je kladný rozptyl veličiny X,
$D(Y)$ značí kladný rozptyl veličiny Y,
$cov(X,Y)$ vyjadřuje kovarianci náhodných veličin.

Pro nezávislost a korelovanost platí následující vztahy:
Jestliže jsou X a Y nezávislé $\Rightarrow$ X a Y nekorelované.

[13] Vlastnosti korelovanosti a závislosti nesmíme zaměňovat

[14] Zatímco výše uvedená definice nezávislosti platí obecně, prostřednictvím zde definované kovariance popisujeme jen korelovanost lineárního typu

[15] Pro odhad nelineární závislosti lze použít Spearmanův koeficient, viz. [18].

Jestliže jsou X a Y nekorelované nemusí být X a Y nezávislé!
Jestliže jsou X a Y korelované $\Rightarrow$ X a Y závislé.

V případě náhodného výběru z dvourozměrného rozdělení se k vyjádření korelace využívá výběrový korelační koeficient:

$$r = r_{XY} = \frac{S_{XY}}{\sqrt{S(X)^2 \cdot S(Y)^2}} = \frac{S_{XY}}{S(X) \cdot S(Y)} \tag{3.3}$$

kde $S(X)$,$S(Y)$ jsou výběrové směrodatné odchylky proměnných X a Y,
S_{XY} je výběrová kovariance.

Výběrová kovariance je definovaná:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}) \cdot (Y_i - \bar{Y}) \tag{3.4}$$

kde $i = 1, 2, ...n$,
$\bar{X}$ je průměr proměnné X,
$\bar{Y}$ průměr proměnné Y [1].

Pro posouzení závislosti mezi páry více proměnných vytvoříme korelační matici $\boldsymbol{R_{XX}}$, kde na místech matice a_{ij} jsou korelační koeficienty mezi *i-tou* a *j-tou* nezávisle proměnnou.

Následně musí být provedeno testování korelačních koeficientů, to se provádí za předpokladu, že rozděleni obou proměnných je normální a jejich vztah je přibližně lineární. Testuje se hypotéza o nulové hodnotě korelačního koeficientu základního souboru, tedy $H_0 : \rho_{xy} = 0$, tzn. že korelace je nulová. Alternativní hypotéza $H_1 : \rho_{xy}$ je nenulový. Test hypotézy se provadi pomocí testového kritéria:

$$t = \frac{\mid r \mid}{\sqrt{1 - r^2}} \cdot \sqrt{n-2} \tag{3.5}$$

Padne-li vypočtená hodnota testovacího kritéria do kritického oboru, $|t| > t_{\alpha/2(n-2)}$ [16]. zamítáme hypotézu H_0 na zvolené hladině významnosti α. Tím zamítneme, že jsou daná data nezávislá, otestovat stejným testem sílu závislosti již není možné, neboť nejsou splněny základní předpoklady pro použití testu.

Testujeme-li multikolinearitu pro větší počet nezávisle proměnných, jejichž koeficienty korelace jsou v korelační matici $\boldsymbol{R_{XX}}$, není vhodné testovat každý korelační koeficient

[16] $t_{\alpha/2(n-2)}$ odpovídá příslušnému kvantilu Studentova rozdělení

samostatně výše uvedeným testem, neboť narůstá riziko chyby I. druhu, proto se používá souhrnný test pro ověření multikolinearity např. Farrarovo-Glauberův, který uvádí Cipra [18], ten pak počítá s celou korelační maticí:

$$-(n-1-\frac{(2k+5)}{6})\cdot\ln|\boldsymbol{R_{XX}}|\geq\chi^2_{1-\alpha}(k\frac{k-1}{2}), \tag{3.6}$$

kde k je počet proměnnných,
n je počet pozorování,
$\boldsymbol{R_{XX}}$ je výběrová korelační matice.

Testována je hypotéza H_0: že všechny korelační koeficienty mezi jednotlivými závislými proměnnými se významně (na hladině významnosti α) neodlišují od nuly. Kritický obor $\chi^2_{1-\alpha}(k(k-1)/2)$ odpovídá příslušnému kvantilu χ-kvadrát rozdělení.

Tímto jsou posouzeny vztahy mezi páry proměnných, po odstranění kolinearity mezi nimi se ovšem v souboru dat, kde vystupuje více nezávislých proměnných může vyskytovat ještě závislost některé z proměnných na lineární kombinaci jiných proměnných. Vícenásobnou korelaci lze pak posoudit prostřednictvím koeficientu mnohonásobné korelace [18].

$$r_{y\boldsymbol{X}}=\sqrt{\boldsymbol{R}_{y\boldsymbol{X}}\boldsymbol{R}^{-1}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{R}'_{y\boldsymbol{X}}} \tag{3.7}$$

Více podrobností k mnohonásobné korelaci, včetně kritických hodnot stanovených k jejímu testování lze dohledat v [18].

Jednotlivé proměnné dále mohou vytvářet shluky, což může také ovlivnit výsledky modelu scoringu a může být ověřeno cluster analýzou, provedení cluster analýzy překračuje rozsah této práce a nebude zde aplikována[17].

Analýza dat je časově velmi náročná, ale je velmi důležitou pro konečné výsledky celého modelu. Po provedení analýzy dat rozdělíme data na testovací a validační část. Testovací část dat slouží k vytvoření modelu, validační poté k ověření jeho vlastností.

[17] Se základním přístupem cluster (shlukové) analýzy se lze seznámit v [17] a v [1], podrobněji lze problematiku cluster analýzy najít např. v knize: P. HEBÁK, J. HUSTOPECKÝ, I. PECÁKOVÁ, M. PRŮŠA, H. ŘEZANKOVÁ, A. SVOBODOVÁ, P. VLACH, it Vícerozměrné statistické metody (3). Praha (2005).

3.3 Statistické možnosti řešení modelu scoringu

K vytvoření modelu skórovací funkce se nabízí několik možností za využití různých statistických nástrojů. Seznam možných statistických nástrojů je uveden níže v tabulce.

Nejčastěji využívaný přístup k analýze dat různé povahy je regrese. Při tvorbě skóringových modelů bývá v praxi nejčastěji používána metoda logitové (logistické) regrese, proto se především na ni zaměříme také v této práci. Jedním z důvodů pro volbu metody logitové regrese je absence přísných předpokladů na vstupní proměnné. Logitová regrese se neumí vypořádat jen s chybějícími hodnotami proměnných a dále vyžaduje vyřešení multikolinearity mezi proměnnými[18]. V následujících kapitolách budou vedle metodiky tvorby modelu logitové regrese ukázány taktéž důvody pro volbu této metody.

Statistická metoda	**Technika**	**Poznámka**
Lineární regrese	*Metoda nejmenších čtverců*	*východiskem některých dalších metod*
Logitová regrese	*Metoda maximální věrohodnosti*	*modifikace lineární regrese*
Probitová regrese	*Metoda maximální věrohodnosti*	*modifikace lineární regrese*
Diskriminační analýza	*Teorie rozhodování*	*rozdělení do skupin*
Rozhodovací stromy	*Rekurzivní segmentovací algoritmy*	*neparametrická metoda*
Neuronové sítě	*Jedno- nebo vícevrstvý perceptron*	*neparametrická metoda*
Analýza přežití	*Cox Proportional Hazard Model*	*na rozdíl od výše uvedených metod řešících otázku "zda nastane default", odpovídá na otázku "kdy nastane default"*

Tabulka 3.2: Seznam statistických metod

[18]S chybějícími hodnotami proměnných, jakožto i s multikolinearitou mezi proměnnými, umí naproti tomu pracovat neparametrické metody, jejich problémem ovšem je jejich nesnadná interpretace jak bankovním pracovníkům, tak klientům.

3.4 Lineárně regresní model

Regresní vztah obecně popisuje změnu, která nastává u závislé vysvětlované proměnné Y při změnách nezávislých proměnných X_1, X_2, ..., X_k. Ač model scoringu bývá nejčastěji (a ani v této práci tomu jinak nebude) popisován pomocí logitové regrese, v této kapitole si nejprve přiblížíme nejjednodušší formu regrese, regresi lineární, jejíž některé myšlenky jsou využívány i v regresi logitové.

Lineární regresi Hušek [2] definuje následujícím způsobem. Nechť máme n pozorování k nezávislých proměnných $X_1, X_2, \ldots, X_k$ a konkrétní hodnoty závisle proměnné Y, potom pro stochastickou lineární závislost mezi vysvětlovanou proměnnou Y a k vysvětlujícími proměnnými X_1, X_2, ..., X_k platí

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u, \tag{3.8}$$

kde β_0 je tzv. úrovňová konstanta, absolutní člen[19], β_j je *j-tý* regresní koeficient (parametr), pro $j = 1, 2, \ldots, k$ a u je náhodná složka[20].

Soustavu n rovnic pro n pozorování lze zapsat maticově:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{3.9}$$

nebo

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{k1} \\ 1 & X_{12} & \ldots & X_{k2} \\ 1 & \cdot & \ldots & \cdot \\ 1 & \cdot & \ldots & \cdot \\ 1 & \cdot & \ldots & \cdot \\ 1 & X_{1n} & \ldots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{bmatrix}. \tag{3.10}$$

Vždy musí platit, že $n > k$,rozdíl $n - k$ potom představuje počet stupňů volnosti daného modelu.

Jestliže pro náhodnou složku platí $E(u) = 0$, pak očekávanou hodnotu Y můžeme vyjádřit:

$$E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3.11}$$

[19] Můžeme definovat zvláštní umělou proměnnou X_0, která nabývá ve všech pozorováních hodnoty 1.
[20] Vlastnosti náhodné složky v lineárně regresním modelu jsou uvedeny dále.

regresní parametry $\beta_1, \beta_2, \ldots, \beta_k$ představují změnu $E(Y)$, která odpovídá jednotkové změně příslušné proměnné X, za podmínky, že ostatní vysvětlující proměnné zůstávají shodné (ceteris paribus). β_0 je absolutní člen, který udává, jaké hodnoty nabývá proměnná Y, jsou-li všechny proměnné X nulové.

Výše popsaná regresní funkce je jen teoretická, neznáme hodnoty regresních parametrů ani náhodnou složku. Na základě empirických údajů - závislé i nezávislé proměnné - odhadneme empirickou (výběrovou) regresní funkci:

$$\hat{Y} = b_0 + b_1 X_1 + \ldots + b_k X_k, \tag{3.12}$$

kde $b_0, b_2, \ldots, b_k$ jsou bodové odhady neznámých parametrů $\beta_0, \beta_2, \ldots, \beta_k$,
$\hat{Y}$ je predikovaná hodnota Y.

Pro vyrovnané hodnoty jejích jednotlivých pozorování Y_i poté platí:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + \ldots + b_k X_{ki}, \tag{3.13}$$

pro $i = 1, 2, \ldots, n$
Rozdíl mezi hodnotami Y_i a $\hat{Y}_i$ se nazývá reziduum e_i: $e_i = Y_i - \hat{Y}_i$, pro $i = 1, 2, \ldots, n$.
Měřitelná rezidua lze chápat jako odhad neznámé složky u_i.

Dle [1] lineárním regresním modelu mají být splněny následující podmínky (předpoklady):

- **A.** $E(\boldsymbol{u}) = 0$, tj. chyby měření nejsou systematické,
- **B.** $E(\boldsymbol{u}\boldsymbol{u'}) = \sigma^2 \boldsymbol{I_n}$, vyjadřuje homoskedacitu a sériovou nezávislost dat, v párech nekorelované,
- **C.** $E(\boldsymbol{X'u}) = 0$, tj. $\boldsymbol{X}$ je nestochastická matice a tedy jediný náhodný prvek v modelu je náhodná složka u,
- **D.** $\boldsymbol{X}$ má plnou hodnost $k + 1$, tedy matice nezávisle proměnných X neobsahuje žádné dva perfektně závislé sloupce proměnných.

Vlastnosti **A. - C.** lze shrnout do jedné silnější podmínky **E.** pro náhodnou složku: $u \sim N_n(0, \sigma^2 \boldsymbol{I})$

Při splnění výše uvedených předpokladů může být odhad parametrů u lineární regrese proveden prostřednictvním metody nejmenších čtverců (MNČ). Při splnění podmínky

normality náhodné složky mohou být provedeny testy významnosti jednotlivých parametrů, jakožto i celého modelu. Metoda nejmenších spočívá v minimalizaci součtu čtverců reziduí, Hušek [2] ji popisuje takto:

$$\boldsymbol{e'e} = \boldsymbol{y} - \boldsymbol{Xb'y} - \boldsymbol{Xb} = \boldsymbol{y'y} - 2\boldsymbol{b'X'y} + \boldsymbol{b'X'Xb}, \tag{3.14}$$

První parciální derivaci podle $\boldsymbol{b}$ položíme rovnou nule:

$$\frac{\partial(\boldsymbol{e'e})}{\partial(\boldsymbol{b'})} = -2\boldsymbol{X'y} + 2\boldsymbol{X'Xb} = 0 \tag{3.15}$$

Úpravou dostaneme soustavu normálních rovnic metody nejmenších čtverců

$$\boldsymbol{X'Xb} = \boldsymbol{X'y}, \tag{3.16}$$

jestliže tedy existuje $\boldsymbol{(X'X)^{-1}}$, tedy je splněna výše vymezená podmínka D., dostaneme bodovou odhadovou funkci b ve tvaru

$$\boldsymbol{b} = \boldsymbol{(X'X)^{-1}X'y}, \tag{3.17}$$

Tento výraz splňuje podmínku dosažení minimalizace rovnice součtu čtverců reziduí a lze dokázat, že $\boldsymbol{b}$ je nestranný lineární odhad parametru $\boldsymbol{\beta}$.

Splnění výše uvedených podmínek však velmi limituje užití metody nejmenších čtverců.

Kvalita modelu se posuzuje prostřednictvím reziduálního součtu čtverců $\boldsymbol{e}$ a

$$RSS = S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y'(I - H)Y}, \tag{3.18}$$

kde $\boldsymbol{H}$ je projekční matice definovaná jako $\boldsymbol{H} = \boldsymbol{X(X'X)^{-1}X'}$.

Jako nestranný odhad parametru σ^2 slouží veličina s^2, která je definovaná jako

$$s^2 = \frac{RSS}{n - k} \tag{3.19}$$

Při modelování lineární regrese nabývá závisle proměnná Y hodnot od $-\infty$ do $+\infty$, zatímco závislé proměnná u odhadu defaultu je pouze binární, nabývá hodnoty 0, kdy klient je bonitní nebo hodnoty 1, u klienta nastane default. Proto musíme přistoupit k úpravě modelu regrese, která proměnnou transformuje na celý interval $(-\infty, +\infty)$.

3.5 Logitový model

Nechť máme závisle proměnnou binárního typu nabývající hodnoty $y_i = 1$, nastane-li u *i-tého* klientu default nebo $y_i = 0$, je-li *i-tý* klient bonitní, potom má tato veličina alternativní rozdělení.

Definujeme $\pi(X_i) = P(y_i = 1|X_i)$ jako podmíněnou pravděpodobnost defaultu *i-tého* klienta za podmínky, že má vektor vysvětlujících proměnných X_i.
A $(1-\pi(X_i)) = P(y_i = 0|X_i)$ jako podmíněnou pravděpodobnost bonitního *i-tého* klienta za podmínky, že má vektor vysvětlujících proměnných X_i.

Pravděpodobnost $\pi(X)$ nahradíme proměnnou *odds*, která bývá označována jako šance jevu a definuje se jako podíl pravděpodobností $\pi(X)$ a $(1 - \pi(X))$:

$$odds(\pi(X)) = \frac{\pi(X)}{1 - \pi(X)} \tag{3.20}$$

Hodnotu této funkce najdeme v intervalu $\langle 0, +\infty)$.

K tomu, abychom se dostali do intervalu $(-\infty, +\infty)$, tedy intervalu, do kterého patří hodnoty odhadnuté regresní rovnicí, provedeme ještě logaritmickou transformaci[21]. Dostaneme proměnnou obecně nazývanou logit.

$$g = logit[\pi(X)] = \ln \frac{\pi(X)}{1 - \pi(X)} \tag{3.21}$$

Vztah mezi logitem a vektorem vysvětlujících proměnných X má lineární charakter.

Tomuto typu regrese se díky provedenému typu transformace říká logitová a zápis regresní rovnice je následující:

$$g = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3.22}$$

a dá se interpretovat, že logaritmus šancí je lineární funkcí vysvětlujících proměnných[22].

Vztah mezi pravděpodobností $\pi(X)$ a vektorem vysvětlujících proměnných má v případě logitové regrese následující tvar:

$$\pi(X) = \frac{e^{(\beta_0+\beta_1 X_1+...+\beta_k X_k)}}{1 + e^{(\beta_0+\beta_1 X_1+...+\beta_k X_k)}} = \frac{e^{(\beta^{\mathbf{T}} X_i)}}{1 + e^{(\beta^{\mathbf{T}} X_i)}} = \frac{1}{1 + e^{-(\beta^{\mathbf{T}} X_i)}} \tag{3.23}$$

[21]Použijeme-li k transformaci namísto exponenciely normální rozdělení, získáme namísto logitové regrese probitovou, která se odlišuje právě jen použitou transformací

[22]Odhadované parametry ani proměnné v modelu logitové regrese neodpovídají parametrům a proměnným nacházejícím se v modelu lineární regrese

Skalární součin $\beta^{\mathbf{T}} X_i$ se v teorii řízení úvěrového rizika nazývá "skóre". Výše uvedená rovnice vznikla zpětnou transformací a má exponenciální charakter, nikoliv lineární.

Nyní, kdy známe základní princip logitové regrese, se můžeme vrátit zpět k předpokladům, které jsou kladeny na lineárně regresní model a jejich změně v případě, používáme-li model logitové regrese. Důvodem, proč jsme logitovou regresi vůbec využili, byla transformace závisle proměnné z intervalu $(-\infty, +\infty)$ na interval $\langle 0, 1\rangle$.

Další předpoklad lineárně regresního modelu je kladen na náhodnou složku a vyžaduje konstantní rozptyl náhodné složky nezávisle na hodnotách nezávisle proměnné. Ve chvíli, máme-li naši závisle proměnnou binárního typu 0 nebo 1, její rozptyl nabývá hodnoty $var = \pi(X) \cdot (1 - \pi(X))$, tzn. pro různé pravděpodobnosti jevu $\pi(X)$, nabývá různé hodnoty rozptylu:

$$\pi(X) = 0,5 \Rightarrow var = 0,5 \cdot 0,5 = 0,25$$
$$\pi(X) = 0,9 \Rightarrow var = 0,9 \cdot 0,1 = 0,09$$
$$\pi(X) = 1,0 \Rightarrow var = 1,0 \cdot 0,0 = 0,00.$$

Testy významnosti jednotlivých parametrů (a tedy i nezávisle proměnných) v lineárně regresním modelu mohou být provedeny jen za splnění předpokladu normality náhodné složky. Tento předpoklad lze u proměnné nabývající jen hodnot 0 nebo 1 jen velmi těžko ověřit, respektive je-li závisle proměnná z alternativního rozdělení, musí i její náhodná složka pocházet z alternativního rozdělení.

Z výše uvedeného plyne, že po provedení logistické transformace závisle proměnné nejsou v modelu splněny předpoklady lineárně regresního modelu, proto nelze lineární regresi ani aplikovat na naše data. Logitová regrese tak, jak je popsaná v této práci, tyto předpoklady na data neklade a pro její použití není nutné ani ověřovat tyto předpoklady.

3.5.1 Odhad parametrů

Pro vytvoření logitové regresní funkce potřebujeme odhadnout parametry $\beta_0, \ldots, \beta_k$. Tyto parametry opět udávají váhu jednotlivých vysvětlujících proměnných $X_1, \ldots, X_k$, ale tentokrát nikoliv na Y, nýbrž na $logit[\pi(X)]$. Zatímco u lineární regrese je odhad parametrů prováděn metodou nejmenších čtverců, která minimalizuje sumu čtverců odchylek očekávané hodnoty závisle proměnné a hodnoty závisle proměnné stanovené pomocí mo-

delu lineární regrese, u logitové regrese neexistuje matematické řešení, které by dokázalo explicitně vyjádřit hodnotu odhadu na základě metody nejmenších čtverců.

Nejčastěji se tak pro odhad parametrů v logitové regresi používá metoda maximální věrohodnosti. Tato metoda maximalizuje pravděpodobnost toho, že nám modelem bude predikována očekávaná hodnota.

Metoda maximální věrohodnosti je založena na konstrukci věrohodnostní funkce, která sjednocuje pravděpodobnosti získané z pozorování nezávislých proměnných pro modelované hodnoty závislé proměnné. Vychází z pravděpodobnosti proměnné Y, která může nabývat jen dvou hodnot 0 nebo 1.

$$P(y_i = a) = \pi(\mathbf{x_i})^a \cdot (1 - \pi(\mathbf{x_i})^{1-a}, pro\ a = 0, 1. \tag{3.24}$$

Za předpokladu, že pozorované hodnoty jsou nezávislé, definujeme věrohodnostní funkci $l(\beta)$ jako

$$l(\beta) = \prod_{i=1}^{n} \pi(\mathbf{x_i})^{y_i} \cdot (1 - \pi(\mathbf{x_i}))^{1-y_i}, \tag{3.25}$$

tj. jako součin podmíněných pravděpodobností pro jednotlivá pozorování.

Pro výpočty se namísto maximalizace funkce $l(\beta)$ definované součinem dle rovnice 3.22 používá maximalizace sumy $\ln[l(\beta)]$, tedy zlogaritmované rovnice 3.22.

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(\mathbf{x_i})] + (1 - y_i) \ln[1 - \pi(\mathbf{x_i})]\}. \tag{3.26}$$

V dalším kroku vytvoříme parciální derivace funkce $L(\beta)$ pro jednotlivé parametry $\beta_0, \ldots, \beta_k$ a položíme rovny nule a získáme tím tzv. věrohodnostní rovnice:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x_i})] = 0 \tag{3.27}$$

a

$$\sum_{j=1}^{k} x_{ij} [y_j - \pi(\mathbf{x_j})] = 0, \tag{3.28}$$

pro j = 0, . . . , k

Tuto nelineární rovnici následně řešíme numericky iterační metodou. Jako řešení dostaneme odhad parametrů β.

Ještě musíme odhadnout chybu takto provedeného odhadu, což provedeme přes matici druhých parciálních derivací funkce $[L(\beta)]$ podle β,

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi(\mathbf{x_i})[1-\pi(\mathbf{x_i})] \tag{3.29}$$

pro j = 0, . . . , k

a

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi(\mathbf{x_i})[1-\pi(\mathbf{x_i})], \tag{3.30}$$

pro $j \neq l$ a j = 0, . . . , k.

Tyto druhé parciální derivace tvoří matici o rozměru $(k+1) \times (k+1)$, tato matice s opačnými znaménkami se nazývá Fisherova informační matice a značí se $\boldsymbol{I(\beta)}$. Inverzní matice k $\boldsymbol{I(\beta)}$ je potom asymptotická variační matice, tedy $\boldsymbol{Var(\beta) = I^{-1}(\beta)}$. Potom $Var(\beta_j)$, tedy rozptyl *j-té* složky vektoru β je *j-tý* diagonální prvek matice $\boldsymbol{Var(\beta)}$.

Jelikož neznáme parametry β_j, musíme dosadit námi vyjádřené odhady parametrů $\hat{\beta}_j$ a dostaneme asymptotický odhad variance parametrů $\hat{Var}(\hat{\beta}_j)$. Odhad směrodatné odchylky se poté stanoví jako:

$$\hat{SE}(\hat{\beta}_j) = \sqrt{\boldsymbol{\hat{Var}(\hat{\beta}_j)}} \tag{3.31}$$

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

kde $\boldsymbol{X}$ je $n \times (k+1)$ rozměrná matice dat nezávisle proměnných,
$\boldsymbol{V}$ je $n \times n$ rozměrná matice, kde na diagonále jsou jednotlivé rozptyly:

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}(x_1)[1-\hat{\pi(x_1}] & 0 & \dots & 0 \\ 0 & \hat{\pi}(x_2)[1-\hat{\pi(x_2}] & \dots & 0 \\ \cdot & 0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 0 & \dots & 0 & \hat{\pi}(x_n)[1-\hat{\pi(x_n}] \end{bmatrix}.$$

Ve chvíli, kdy získáme odhady parametrů β je nezbytné otestovat významnost jednotlivých parametrů a tedy i příslušných nezávisle proměnných na výsledek modelu. Hosmer [4] uvádí jako nejčastěji používaný test pro tyto účely *Waldův test* W, kde je na zvolené hladině významnosti α testována nulová hypotéza H_0: $\beta_j = 0$, proti alternativě

H_1: $\beta_j \neq 0$. Nemůžeme-li na zvolené hladině významnosti α zamítnout testovací kritérium H_0, znamená to, že daný parametr je na zvolené hladině významnosti α nevýznamný pro daný model. Testovací kritérium je pak dáno prostřednictvím Waldovy testovací statistiky:

$$W_j = \frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)}\,, \tag{3.32}$$

tato má za platnosti nulové hypotézy normované normální rozdělení $W_j \backsim N(0,1)$, W_j^2 má zhruba χ^2 rozdělení s jedním stupněm volnosti.
Kvantil normálního rozdělení příslušný dané hladině významnosti α určuje kritické hodnoty, které rozdělují obor možných hodnot testovacího kritéria (v tomto případě Waldovy testovací statistiky) na dvě množiny: obor nezamítnutí hypotézy H_0 a kritický obor, padne-li hodnota testovacího kritéria do kritického oboru, zamítáme na zvolené hladině významnosti α hypotézu H_0.

Interval spolehlivosti pro Waldovu statistiku se určí poté ze vzorce:

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\, \hat{SE}(\hat{\beta}_j), \tag{3.33}$$

kde $z_{1-\frac{\alpha}{2}}$ je příslušný kvantil normálního rozdělení pro zvolenou hladinu významnosti.

Jestliže se ve vymezeném intervalu nachází 0, nelze zamítnout hypotézu o nevýznamnosti parametru pro model na zvolené hladině významnosti α.

Analogicky Waldovu testu se v lineární regresi dělá t-test o významnosti parametru pro model na stanovené hladině významnosti, t-test, jakožto i podmínky a způsob jeho aplikace lze dohledat např. v [1].

Při softwarovém zpracování testování hypotéz se namísto stanovení kritických hodnot využívá tzv. $p-$ *hodnota* testu, $p-$ *hodnota* testu představuje zjednodušeně pravděpodobnost, že výsledek testovacího kritéria za platnosti hypotézy H_0 bude náležet do kritického oboru. Čím menší p-hodnota, tím nepravděpodobněji by takového výsledku (testovací kritérium nabývá hodnoty z kritického oboru) za předpokladu platnosti H_0 bylo dosaženo. Výhodou $p-$ *hodnoty* je, že není nutné znát konkrétní volbu hladiny významnosti α a jí příslušné kritické hodnoty, samotný ukazatel $p-$*hodnoty* obsahuje dostatečnou informaci aniž by k výpočtu byla nutno použít zvolenou hladinu významnosti α.

3.5.2 Testování a vyhodnocení modelu

Pro testování hypotézy o významnosti celého modelu se používají testy vybudované nad věrohodnostní funkcí a jejími odhady. Mezi ně se řadí test "odchylek"(deviance) a poměrový test věrohodnosti [4]. Tyto testy slouží k porovnání dvou modelů vůči sobě.

$$D = 2[L_s - L(\hat{\beta})] = -2 \ln \frac{l(\hat{\beta})}{l_s}, \tag{3.34}$$

kde l_s je věrohodnostní funkce saturovaného modelu,
tj. modelu, který obsahuje tolik parametrů, kolik je dat v modelu,
L_s je $\ln(l_s)$.

Z definice saturovaného modelu plyne i hodnota l_s v případě, kdy závisle proměnná y_i nabývá jen dvou hodnot (0 nebo 1):

$$l_s = \prod_{i=1}^{n} y_i^{y_i} \cdot (1 - y_i)^{1-y_i} = 1, \tag{3.35}$$

a proto $D = -2 \ln l(\hat{\beta})$.

Výše uvedený test odchylek má v logitové regresi stejnou roli jako v lineární regresi reziduální součet čtverců.

Stejnou transformaci věrohodnostní funkce využívá i další test, proto si znovu připomeňme, jak získáme její odhad: $\ln(l) = \sum_{i=1}^{n} \ln[\hat{\pi}(x_i)]$, touto transformací se z intervalu $\langle 0, 1 \rangle$ dostáváme do intervalu $\langle 0, \infty)$.Při využívání metody maximální věrohodnosti hledáme co největší l, při testování bývá hodnota l zaměněna hodnotou $-2 \ln(l)$, kterou tedy vyžadujeme naopak co nejmenší. Tato statistika pak bývá použita k porovnání dvou modelů mezi sebou, při čemž hledáme model s maximální statistikou l.

Chceme-li opět ukázat na nějakou analogii s lineárním regresním modelem, tak ve chvíli, kdy se v něm zajímáme o koeficient determinace R^2 (více o koeficientu determinace lze najít v [1] nebo v [18]), zajímá nás v logitovém modelu statistika l.

V poměrovém testu věrohodnosti pak porovnáváme,

$$-2 \ln \frac{l_0}{l_1} = -2(L_0 - L_1), \tag{3.36}$$

kde L_1 je hodnota věrohodnostní funkce pro plný model,
L_0 je hodnota věrohodnostní funkce pro model mající o vybranou proměnnou méně.

Tato transformace věrohodnostní funkce může být aproximována statistikou mající χ^2 rozdělení.

Na statistice l jsou postaveny i další kritéria k posouzení kvality logitového modelu. Mezi ně patří např. Akaikeho kritérium AIC:

$$AIC = -2\ln(l) + 2(k+s), \tag{3.37}$$

kde s je počet hodnot závislé vysvětlované proměnné ponížený o 1, tedy u scoringu, kde je binární hodnota 1 (nastane default) nebo 0 (bonitní klient), je $s = 1$,
k je poté počet vysvětlujících proměnných v modelu.

Schwarzovo kritérium SC je taktéž vybudováno na věrohodnostní funkci a její hodnotě,

$$SC = -2\ln(l) + (k+s)\ln(n), \tag{3.38}$$

kde s je počet hodnot závislé vysvětlované proměnné ponížený o 1,
k je počet vysvětlujících proměnných v modelu,
n je počet pozorování.

Výše zmíněná kritéria slouží k vzájemnému porovnání modelů mezi sebou, nižší hodnoty indikují i vhodnější model pro naše data.

Další možnost, jak bývá logitový model hodnocen, je využití měr asociace, které vyjadřují sílu vazby mezi predikovanými a skutečnými hodnotami modelu. Tyto asociační míry párují hodnotu odhadu s hodnotou skutečnou, sledují páry s následujícími vlastnostmi:

- n_c - počet shodných párů,
- n_d - počet opačných párů,
- n_t - počet ostatních párů, $n_t = t - n_c - n_d$,
- t - celkový počet párů, $t = n_t + n_c + n_d$,
- n - počet pozorování.

Přirozeně platí, že model je tím kvalitnější, čím více obsahuje shodnýc párů n_c. Nejčastěji užívané asociační kritérium je c, pro kterou platí, že čím vyšší hodnoty nabývá, tím kvalitnější model je.

$$c = \frac{n_c + 0,5(t - n_c - n_d)}{t} = \frac{n_c + 0,5n_t}{t}\,. \tag{3.39}$$

Mezi další patří např. Somerovo D či Kendallovo Tau

$$Somerovo\ D = \frac{n_c - n_d}{t}\,, \tag{3.40}$$

Toto kritérium může nabývat hodnoty z intervalu $\langle 0, 1\rangle$, přičemž platí, že čím vyšší hodnoty nabývá, tím je model hodnocen jako kvalitnější.

$$Kendallovo\ \tau = \frac{n_c - n_d}{0,5}\ \cdot n \cdot (n-1). \tag{3.41}$$

Kendallovo τ může nabývat hodnot z celého oboru reálných čísel. Je-li počet opačných párů v modelu vyšší než počet shodných párů, τ nabývá záporné hodnoty, opět tedy platí, že hledáme model, pro který je τ nejvyšší. V absolutním měřítku není ani toto kritérium aplikovatelné a slouží jen k vzájemnému porovnání mezi více modely.

Máme-li sestaven model scoringu je nutno se zamyslet nad vhodnou hodnotou "cutoff". Všimneme-li si procesu stanovování logitové regrese "cutoff"je zde možné stanovit pro stejný model trojím způsobem:

- prostřednictvím podmíněné pravděpodobnosti $\pi(X)$, tento přístup bývá nejčastěji využíván ve statistice a následně ve statistickém softwaru, kde bývá navíc defaultně nastavena hodnota 50% [23],
- prostřednictvím šance $odds(\pi(X))$,
- prostřednictvím hodnoty logitu, tento přístup bývá často používán ve scoringu vůči bankovním pracovníkům i vůči klientům, neboť je snadno interpretován [24].

3.6 Výběr vhodných proměnných

Při používání logitové regrese musíme vždy vyhodnotit data, co do hlediska chybějících hodnot pro některé proměnné či logické chyby při sběru dat, jakožto i multikolinearity

[23] Hodnotu pravděpodobnosti lze většinou samozřejmě skriptem upravit.

[24] Uvědomíme-li si, že vztah logitu s jednotlivými proměnnými je lineární, není vhodné stanovovat hodnotu cutoff pomocí logitu, používáme-li kvantitativní proměnné bez jejich kategorizování a limitování, neboť výkyv jedné nezávislé proměnné nám může významně ovlivnit výstupní proměnnou v absolutní výši a ohodnocení by nebylo postavené na kombinaci všech nezávislých proměnných, jak je zamýšleno. Užití logitu není dobře interpretovatelné ani v případě samých binárních proměnných.

mezi jednotlivými proměnnými. Logitová regrese neumí modelovat scoring nad neúplnými daty (chybějící pozorování) a výsledky získané logitovou regresí jsou silně ovlivněny multikolinearitou mezi proměnnými jakožto i nesourodými daty (nekontinuita či chyba při sběru dat). Základní analýzu dat je nutné tedy udělat i při zpracování současného modelu scoringu či u modelu scoringu, kde nemáme možnost volby proměnných.

Ve chvíli, kdy nemáme k dispozici dosavadní model scoringu nebo nemáme pevně stanoveno (dáno často interními předpisy některých finančních institucí), které proměnné musíme do modelu zahrnout a které naopak nemůžeme, měli bychom všechny proměnné zkoumat jednotlivě a posuzovat jejich vliv na odhad bonity/defaultu klienta. Závěrem této naší analýzy by měl být výběr vhodných proměnných, které zahrneme do modelu.

3.6.1 Vliv jednotlivých nezávisle proměnných na závisle proměnnou

Výběr proměnných, na kterých chceme vybudovat svůj model credit scoringu, by měl začít analýzou vztahu každé možné vstupní proměnné a výstupní závisle proměnné. K tomuto je určeno několik metod, přičemž platí, že by měly být používány jako doplňkové a vždy by mělo být použito současně více z nich. [25]

Zatímco Hosmer s Lemeshowem [4] uvádí jako metodu pro sledování vlivu jednotlivých nezávisle proměnných na závisle proměnnou zachování lineárního trendu mezi logitem a nezávisle proměnnou. To znamená, každá vstupní proměnná by měla mít samostatně lineární vztah k očekávanému výstupnímu logitu. Pro testování linearity poté vybranou závisle proměnnou rozděluje do kvantilů a stanovuje proporcionalitu defaultu v jednotlivých kvantilech závisle proměnné. S měnící se hodnotou závisle proměnné by pak mělo docházet i k lineární změně výskytu defaultu.

Pro každý kvantil závisle proměnné je pak vytvořena umělá dummy proměnná a vliv daného kvantilu jakožto i rozdělení celé proměnné je pak testováno prostřednictvím testu významnosti parametrů, stejně jako jsou testovány pro různé nezávisle proměnné. Je

[25] S ohledem na typ a rozsah poskytnutých dat pro zpracování aplikační části této práce nelze plně provést analýzu vlivu jednotlivých proměnných na celkový model scoringu. V aplikační práci proto bude vliv jednotlivé proměnné na výstup předveden jen na některé ze zvolených závisle proměnných, aniž by takto byly otestovány veškeré vstupní proměnné modelu.

vytvořen model závislosti defaultu na n kvantilech (atributech) jedné vstupní proměnné, a pomocí Waldova testu je testován význam daného atributu pro model.

Při tvorbě scoringu je testování jednotlivých vstupních proměnných zpracováváno na výše uvedeném principu jen s využitím definovaných koeficientů, které jsou snadno interpretovatelné veřejnosti.

První z metod je postavena na testu poměrem věrohodnosti. Porovnáváme zde věrohodnostní funkci pro model popisující vztah mezi závisle proměnnou a nezávisle proměnnou a konstantou a model obsahující jen závisle proměnnou a konstantu.

$$-2 \ln \frac{L_0}{L_1} = -2(L_0 - L_1), \tag{3.42}$$

kde L_1 je hodnota věrohodnostní funkce model obsahující nezávisle proměnnou,
L_0 je hodnota věrohodnostní funkce pro model obsahující jen konstantu.

Dalším přístupem je stanovení ukazatele **WOE - Weight of Evidence**. Pro stanovení tohoto koeficientu je nutno nezávisle proměnné rozdělit do m kategorií, intervalů a pro každý interval je pak stanoven jeho vliv na závisle proměnnou dle následujícího vzorce:

$$WOE_i = \ln \frac{N_i}{\sum_{i=1}^{m} N_i} / \frac{P_i}{\sum_{i=1}^{m} P_i}, \tag{3.43}$$

kde P je počet výskytu jevu,
N je počet nevýskytu jevu,
i je sledovaný atribut - kategorie dané proměnné,
P_i je počet výskytu jevu v dané kategorii,
N_i je počet nevýskytu jevu v dané kategorii.

Určení hodnoty WOE je nejčastěji používaným pohledem na proměnné vstupující do scoringu, pro lepší pochopení významu tohoto koeficientu může být tento přepsán následujícím způsobem:

$$WOE_i = \ln \frac{N_i}{P_i} - \ln \frac{\sum_{i=1}^{m} N_i}{\sum_{i=1}^{m} P_i}, \tag{3.44}$$

Tento zápis lépe vyjadřuje význam hodnot tohoto ukazatele, je-li hodnota WOE rovna nule, znamená to, že tato kategorie proměnné má stejné *odds* jako celá proměnná. Je-li WOE záporné, znamená to, že častěji default nastává pro klienty, u nichž sledovaná proměnná nabývá hodnoty právě z posuzované kategorie proměnné. Naopak kladné WOE znamená, že klient s proměnnou v dané kategorii je méně rizikový než klient průměrný.

Samotné WOE ovšem neuvažuje proporcionalitu pozorování s jeho jednou hodnotou (daným atributem) vůči celkovému datovému souboru. Pro stanovení vlivu proměnné na celý model je užíváno jiných ukazatelů.

Pro určení relativního významu proměnné na daný model se užívá ukazatel **IV - Information Value**. Hodnota tohoto koeficientu se stanoví

$$IV = \sum_{i=1}^{m} \left(\frac{N_i}{\sum_{i=1}^{m} N_i} - \frac{P_i}{\sum_{i=1}^{m} P_i} \right) \cdot WOE_i, \tag{3.45}$$

kde P je počet výskytu jevu,
N je počet nevýskytu jevu,
i je sledovaný atribut - kategorie dané proměnné,
P_i je počet výskytu jevu v dané kategorii,
N_i je počet nevýskytu jevu v dané kategorii
WOE_i je WOE pro danou kategorii.

Hodnota IV je vždy kladná, proměnné s IV menším než $0,1$ bývají označovány za nevýznamné, zatímco hodnoty nad $0,3$ bývají zařazovány do skórovacích modelů. Interpretovat hodnotu tohoto koeficientu může být velmi obtížné, neboť s ním není asociován žádný statistický test významnosti. Proto je nutné vždy u nevýznamných proměnných dle tohoto koeficientu sledovat, zda daná proměnná, respektive její kategorie nemá významnou vypovídací hodnotu ve spojení s jinou proměnnou. Obecně proto není vhodné vyhodnocovat proměnnou jen na základě provedení tohoto koeficientu.

Dalším přístupem k testování vlivu jednotlivé proměnné na závisle proměnnou je stanovení hodnot **Gini koeficientu**. Tento koeficient je znám a užíván v ekonomické teorii a jeho grafickým zobrazením je Lorenzova křivka. Giniho koeficient dosahuje hodnot z intervalu $\langle 0, 1 \rangle$, pokud je roven 0 pak Lorenzova křivka je přímka na diagonále. Ani Giniho keoficient není užíván k testování statistických hypotéz, ale pouze k vyjádření síly závislosti. V případě logitové regrese může být Giniho koeficient počítán pomocí distribuční funkce *odds*.

Při zkoumání vlivu jednotlivých proměnných na závisle proměnnou si můžeme všimnout, že některé z nich mají podobný vztah k závisle proměnné. Tyto formy závislosti mohou být dále zkoumány shlukovou (cluster) analýzou. Tento typ analýzy vstupních proměnných přesahuje rozsah této práce a nebude v této práci aplikován.

3.6.2 Stepwise analýza

Po provedení analýzy vztahu jednotlivých nezávisle proměnných a závisle proměnné se provádí výběr proměnných do modelu, který již zohledňuje i vztah k ostatním nezávisle proměnným. Jako metoda pro začleňování proměnných do modelu scoringu bývá používána stepwise analýza. Postup stepwise analýzy je blíže popsán v [4].

Stepwise analýza vyhodnocuje vliv jednotlivých proměnných na celkový model postupně v několika krocích, přičemž pro vyhodnocení významu proměnné využíváme ukazatele definované výše v kapitole Testování a vyhodnocení modelu. Pro přidávání proměnných do modelu si musíme stanovit mezní p-hodnotu, do které budeme přidávat proměnné do modelu. Hosmer [4] uvádí jako vhodnou hodnotu 10 %. V prvním kroku poté vystavíme model logitové regrese jen jako vztah závislé proměnné na konstantě. Zaznamenáme si jeho věrohodnostní poměr L_0. Do modelu jen s konstantou přidáváme postupně veškeré nezávislé proměnné a porovnáváme hodnotu L_j. Po prostřídání všech proměnných jen s konstantou, vybereme tu proměnnou s nejnižší hodnotou L_j při splnění podmínky stanovené p-hodnoty a přidáme ji do modelu.

Do takto vytvořené rovnice opět přidáváme postupně všechny zbývající proměnné a porovnáváme hodnoty L_j, do modelu pak přidáme opět proměnnou, jejíž model dosahuje nejnižší hodnoty L_j při splnění podmínky stanovené p-hodnoty. Postup opakujeme do té doby, dokud můžeme do modelu přidat proměnné mající menší než stanovenou hraniční p-hodnotou.

Kapitola 4

Poskytovatel dat a cíl práce

Poskytovatel úvěru, který se podílel na specifikaci zadání této práce a poskytl data pro její zpracování, nechce být z důvodu citlivosti údajů jmenován.

Tento poskytovatel úvěru vstoupil na český trh v devatesátých letech 20. století. Jeho hlavním přínosem měla být podpora prodeje automobilů všech značek koncernu řízeného a vlastněného zakladatelem společnosti. V průběhu deseti let se stala společnost jednou z pěti nejvýznamnějších captive společností působící na českém trhu v oblasti účelového financování formou leasingu, účelového úvěru či splátkového prodeje a poskytování služeb souvisejících s automobily.

Při příchodu na český trh měla společnost nulové portfolio a ke každému klientovi mohla přistupovat individuálně. S narůstajícím portfoliem bylo ovšem nutné začít svoje úvěrové riziko aktivně řídit a co nejvíce procesů automatizovat. V tuto chvíli došlo k vytvoření specializovaného úvěrového oddělení (Credit Risk Managmentu CRM) a implementování procesů pro vyhodnocování úvěrového rizika. Toto úvěrové oddělení musí dodržovat nejen evropskou regulaci úvěrového rizika (nejběžněji Basel II), ale musí se držet mnohem přísnějšími koncernovými pravidly.

Vedle individuálního posuzování klienta formou ratingu, tak byl implementován i první model scoringu. Tento byl vytvořen centrálně pro celou skupinu evropských captive společností bez zohlednění individuálních rysů jednotlivých zemí, spíše po vzoru bankovních společností bez zohlednění specifických ukazatelů účelového financování. Cut-off tohoto modelu tak byl nastaven velmi přísně a většina klientů stejně musela být

dále vyhodnocována zaměstnanci úvěrového oddělení. Hlavní přínos tohoto měl spočívat v získání stejného souboru vlatních dat o klientech. Po prvních dvou letech mohl být poprvé vyhodnocen a upraven.

4.1 Konkretizace cíle práce

Model scoringu, jehož data jsou použita pro aplikační část této práce, je třetím scoringovým modelem využívaným společností, který už byl budován i na vlastních datech společnosti z českého trhu. Přístup koncernu ke scoringu je ovšem stále centralizovaný a koncern se snaží v celém regionu model vybudovat obdobně na základě stejných politik a taktéž požadavků na vstupní proměnné. Hlavní důvod k tomuto přístupu je především řízení souhrnného úvěrového rizika všech evropských společností a možnost interpretovat a hodnotit scoring souhrnně pro celou Evropu.

Koncernový model scoringu je proto vytvářen nad vstupním datovým souborem ze skupiny evropských zemí a souhrnně pro všechny země jsou vybrány vhodné posuzované vstupní parametry, ať už o klientovi nebo o finančních produktech. Takto navržený model je pak testován na datech z konkrétního státu a na základě diskuse o ekonomickém významu modelu jsou případně vylučovány proměnné, které nevyhovují podmínkám daného státu, tyto ovšem zpravidla nebývaly nahrazeny žádnou inidividualizovanou proměnnou specifickou pro daný trh. Tato rozhodnutí byla v rovině "statistického modelování"zpravidla odůvodněna malým vzorkem dat z jednotlivého státu a na "bázi manažerské"potřebou hodnotit všechny trhy společně. I tato společnost začíná od centralizovaného modelu ustupovat a při zachování jistých firemních politik přistupuje i k vytváření modelů scoringu pro jednotlivé státy. Data, která jsou použita v této práci ovšem pochází ještě z období, kdy model scoringu byl vytvářen přísně centrálně.

Modely scoringu jsou proto nastaveny s přísným cutoff a významná část zákazníků je na základě scoringu zamítnuta a označena za defaultní. Je zde pak definovaná rozsáhlá "šedá zóna", kterou scoring zamítl, ale je zde stále možný alternativní způsob posouzení klienta pracovníky CRM.

Cílem této práce má být pak právě posouzení modelu scoringu jen z pohledu českých dat, aniž bychom byli při vytváření modelu limitováni interními směrnicemi společnosti.

Výsledný model pak nemá sloužit k přímé aplikaci, ale má pracovníkům CRM poskytnout alternativu k vyhodnocení jimi používaných modelů scoringů, která získávají od svých nadřízených z centrály.

Společnost by chtěla samostatně zpracovaná vyhodnocení modelu scoringu používat k vyjednávání o možnostech eliminace šedé zóny, tj. zóny, v níž zákazníci nejsou označeni "tvrdě"za defaultní, ale je zde dána vždy ještě možnost přehodnocení závěru scoringu pracovníkem CRM a více rozhodovat jen automaticky, pokud by bylo možné nad daty pustit individualizovaný model.

Tato společnost rozděluje svoje klienty z hlediska vyhodnocování jejich úvěruschopnosti do tří skupin:

- soukromá osoba - spotřebitel,
- podnikatel - živnostník a
- podnikatel - obchodní společnost.

Pro každý segment svých klientů poté využívá jinou skórovací funkci. Každá funkce obsahuje jiné závislé proměnné o vlastnostech klientů a nabízených produktech. Díky zaměření na účelové financování, lze do všech scoringových modelů zahrnout i specifické parametry těchto produktů, typicky zálohu a zůstatkovou hodnotu.

Nejvýrazněji se s problémem "šedé zóny"společnost potýká u segmentu živnostník. U tohoto typu zákazníka lze již jen těžko získat nějaká další adekvátní data vhodná pro individuální posouzení [26]. Vyhodnocení živnostníků tak zabere pracovníkům CRM velké množství času aniž by zde bylo možné sledovat významnou přidanou hodnotu, nejčastěji se nakonec stejně přiklání k tomu, v jaké části šedé zóny, se ten který žádající subjekt nachází, tedy jen uměle ponižují cutoff u scoringu tohoto segmentu, přičemž jejich hlavní náplní práce by mělo být posuzování především úvěruschopnosti klientů, jejichž úvěrový rámec bude mít významný podíl v celém portfoliu společnosti. Celá práce by měla být proto zaměřena především na tuto funkci.

Model scoringu živnostníka bude nejen zkoumán, co do své úspěšnosti odhadů a vhodnosti volby vstupních proměnných, ale bude pro něj navržen i nový model. Pro model scoringu

[26] U obchodních společností pracovníci CRM naproti tomu dokážou zohlednit obchodní plán společnosti, její renome apod., takovéto údaje ovšem u živnostníků nelze předpokládat

živnostník bude taktéž provedena analýza vztahu jednotlivých vstupních proměnných a výstupu. Ač v zadání práce bylo vytvoření vlastního modelu při zapojení i jiných proměnných než které jsou využívány modelem, společnost v dobu, ze které jsou poskytnuta tato data, neuchovávala o klientech jiná data než která byla vyhodnocována v rámci scoringu. Jediné proměnné, které by mohly být navíc zahrnuty do modelu jsou některé údaje o produktu, o který klient žádal, rozšířit model o další data o produktu ale není zájmem poskytovatele dat.

Zběžně budou analyzovány i druhé dva modely scoringu společnosti. Díky svému zaměření společnost ve svém portfoliu bohužel má jen nepatrné množství spotřebitelů, pro které se scoring používá v bankovním sektoru nejčastěji a u kterých se ve scoringové funkci objevuje více proměnných týkajících se přímo klienta[27]. Z důvodu malého vzorku dat nebudou mít závěry dosažené v této části práce výraznou predikční schopnost.

Ač hlavní složku portfolia dané společnosti tvoří financování pro velké obchodní společnosti, nemá společnost dostatečné množství dat pro vyhodnocení scoringu ani těchto klientů, neboť většina zákazníků z této skupiny je vyhodnocována jako korporátní zákazník prostřednictvím ratingu, neboť jejich angažmá představuje pro naši společnost významný podíl na portfoliu. Ale i subjekty, které spadnou do rozsahu angažmá schvalovaného v rámci scoringu, nemají zpravidla jen jednu smlouvu. Proto default jedné obchodní společnosti často znamená default více smluv, přičemž defaut definovaný jako závisle proměnná je default na smlouvě. Zlepšovat model scoringu pro tento segment tak není ani mezi prioritami společnosti. Pro vyhodnocení skórovací funkce pro obchodní společnosti, tak opět chybí dostatečné množství dat a výsledky mohou být považovány pouze za orientační.

4.2 Poskytnutá data

Data zpracovaná v této práci byla sesbírána během jednoho roku poskytování financování klientům na základě jejich žádostí. Doba, po kterou mohl nastat default u každého klienta,

[27]U spotřebitele lze nejvíce využít hodnocení typických vlastností skupiny osob k vyhodnocení konkrétního žadatele o úvěr, neexistuje zde již příliš jiných individualizovaných vlastností, které by ovlivnily chování jednotlivce. Nevýhodou daného poskytovatele úvěru v tomto segmentu je, že nemá přístup do centrální bankovní databáze, kde jsou údaje o chování klientů a jejich defaultech sdílena

byla stanovena na dva roky ode dne, kdy smlouva vstoupila v platnost. Každá smlouva tak byla pravidelně sledována a vyhodnocována co do vymezeného defaultu po dobu dvou let od uzavření.

Společnost za default klienta, který se snaží minimalizovat, považuje situaci definovanou Baselem II, kdy klient je 90 dní po splatnosti některého ze svých závazků z dané smlouvy - default90. Taktéž ji ovšem zajímá stav, kdy klient je 30 dní po splatnosti - default30, neboť ve společnosti je nastaven interní postup a s klienty 30 dní po splatnosti je zahájeno aktivní administrativní řízení. Default90 je podmnožinou jevu default30, u každého klienta, u kterého nastal default90 musel dříve nastat i default30. Společnost se snaží své úvěrové riziko vyhodnocovat prostřednictvím vztahu mezi vstupními proměnnými a defaultem30 tak, aby minimalizovala své veškeré rizikonáklady. S ohledem na to, že default90 se v pozorovaných datech nevyskytuje příliš často a je vždy podmíněn nastáním defaultu30, budeme se i my v této práci řídit tímto přístupem.

Nastavení různé míry rizika by pak mělo být provedeno prostřednictvím různého cutoff. Model scoringu bude mít opět nastaveny dvě hodnoty cutoff:
1. hodnotu, jejímž dosažením bude klient automaticky prostřednictvím scoringu vyhodnocen jako bonitní,
2. hodnotu, při jejímž dosažení bude moci být klient ještě vyhodnocen pracovníkem CRM. Klienti nedosahující ani této hodnoty nemohou již být k poskytnutí úvěru schváleni vůbec, tedy ani individuálně pracovníky CRM.
Při nastavení hodnoty cutoff by mělo být poté vyhodnoceno, zda se mění také proporcionalita jevu default90 v jednotlivých segmentech klientů. V práci bude vyhodnocena úspěšnost modelu pro různé hodnoty nastaveného cutoff, samotná volba cutoff musí bý nechána na manažerském rozhodnutí ve společnosti, neboť dle hodnoty stanovené cutoff se může zúžit případně rozšířit tzv. "šedá zóna".

Veškerá kvantitativní data poskytnutá pro tuto práci jsou již společností rozdělena do intervalů a každému intervalu je přiřazeno vlastní bodové ohodnocení. Přístup bodování ve scoringu je velmi častý a pro vytvoření správného modelu scoringu je nutné i jeho vhodné zvolení. O tento bod vytváření scoringové funkce budeme ochuzeni, neboť společnost z důvodu citlivosti údajů přesná data neposkytla. Musíme vycházet z předpokladu, že data jsou na bázi zpracování jednotlivé proměnné upravována už pro jednotlivé trhy zvlášť, a proto bodové ohodnocení vybrané pro model scoringu by mělo skutečně odpovídat.

Pro účely vyhodnocování našeho modelu budeme tyto proměnné považovat za kvantitativní spojitého typu. U vytvořených kategorií proměnných lze ještě používat druhý přístup k proměnným a to vytváření umělých proměnných. Každá umělá proměnná vyjadřuje jen fakt, zda daný klient nabývá či nenabývá pro danou proměnnou hodnoty z daného intervalu, bez zohlednění, jakou hodnotu, ten, který interval má.

Některé proměnné se vyskytují ve všech modelech. První skupina těchto proměnných se vztahuje k vlastnostem poptávaného úvěrového produktu, mezi ně řadíme:

- Záloha - proměnná typická pro účelové financování, vyjadřuje jak velký díl ceny předmětu, na jehož pořízení si klient žádá o úvěr, zaplatil klient předem, v absolutní výši pak může nabývat hodnot od 0 do (necelých) 100 %,
- Trvání smlouvy TS - doba, po kterou chce klient daný úvěr splácet,
- ObjektRisiko ObR - poskytovatelem úvěru vytvořená umělá proměnná, která hodnotí různé segmenty financovaných produktů, při jejím vytváření byl zohledněn např. pokles ceny v daném segmentu, poptávka po daném segmentu apod.,
- Stáří objektu SO - stáří předmětu, na který klient žádá financování, pro nový předmět v absolutní výši (nikoliv v našem kódování) nabývá tato proměnná hodnoty 0.

Další proměnná společná pro všechny modely vypovídá o dosavadním chování klienta:

- Platební morálka - proměnná hodnotící, jak klient plnil své závazky z dříve uzavřených smluv u poskytovatele, zda včas, či zda některý závazek plnil až po splatnosti, případně jak dlouho po splatnosti. Platební morálka vypovídá v podstatě o defaultnosti klienta v jeho dřívějších závazcích.

Ostatní proměnné jsou již specifické pro každý model a budou definovány u jednotlivých modelů zvlášť.

V následujících tabulkách najdeme seznam všech proměnných vyskytujících se v jednotlivých modelech scoringu spolu s intervaly hodnot, kterých mohou nabývat (po provedeném kódování jednotlivých proměnných společností, toto kódování proběhlo na základě vyhodnocení jednotlivých proměnných v celokoncernovém vzorku dat). Veškerá data lze dohledat v elektronické podobě na CD přiloženém k této práci.

Nejprve se podíváme na seznam proměnných vyskytujících se ve vyhodnocovaném modelu scoringu "Živnostník". Veškerá data k této funkci scoringu jsou na přiloženém CD, ve složce enterpreneur, souboru enterpreneurs.xlsx. Nad rámec proměnných společných pro všechny modely scoringu zde najdeme specifické proměnné:

- Zisk (v tisících Kč za rok) - hospodářský výsledek (zisk/ztráta), kterého živnostník dosáhl za předchocí účetní období,
- Obrat (v tisících Kč za rok) - obrat, kterého živnostník dosáhl v předchozím účetním období,
- Délka podníkání DP - doba, po kterou daný žadatel provozuje svoji živnost na základě příslušných oprávnění, dobu vzniku živnostenského oprávnění lze dohledat a ověřit v živnostenském rejstříku, pro nového živnostníka nabývá tato proměnná v absolutní výši (nemusí tomu tak být v našem kódování) hodnoty 0.

Celkový souhrn proměnných vyskytujících se v modelu scoringu "Živnostník"je níže v tabulce 4.1.

Proměnná	**Bodový rozsah**
Zisk	$0 - 20$
Obrat	$0 - 25$
Délka podnikání DP	$0 - 20$
Platební morálka PM	$0 - 30$
Záloha	$0 - 40$
Trvání smlouvy TS	$3 - 15$
ObjektRisiko ObR	$0 - 20$
Stáří objektu SO	$4 - 20$

Tabulka 4.1: Seznam proměnných scoring "Živnostník"

Nyní se zaměříme na model scoringu "Soukromá osoba"[28], veškerá data jsou v elektronické podobě na přiloženém CD, ve složce consumer, soubor $consumers.xlsx$. V tomto modelu vystupují vedle společných proměnných specifické proměnné:

[28]Pod pojmem soukromá osoba je rozuměn spotřebitel, tento pojem je tak v této práci taktéž používán pro označení soukromé osoby.

- Délka pracovního poměru DPP - doba, po kterou daný žadatel, pracuje na stálý pracovní poměr u současného zaměstnavatele, pro osoby nezaměstnané či nepracující na stálý pracovní poměr, stejně jako pro studenty či důchodce nabývá tato proměnná hodnoty 0,
- Pracovní zařazení PZ - uměle vytvořená proměnná popisující typ pracovní činnosti např. manažer, kancelářská činnost, dělník, kterou žadatel vykonává, tato proměnná by neměla hodnotit jen možný dosahovaný příjem, nýbrž i riziko ztráty zaměstnání,
- Věk - věk žadatele o úvěr (financování),
- Vzdělání V - opět umělá proměnná vyjadřující stupeň dosaženého vzdělání klienta, získávání údajů o této proměnné je v praxi obtížné, reálně je tak často automaticky doplňována hodnota příslušející nejnižšímu vzdělání,
- Roční splátky/Roční příjem - kvantitativní poměrová proměnná mezi splátkami žádaného produktu a příjmem žadatele, v absolutních hodnotách proměnné by mělo platit, že proměnná nabývá hodnoty menší než 1 a pro poskytovatele úvěru by mělo být logicky žádané, aby byla co nejnižší.

Seznam všech proměnných vyskytujících se ve scoringu "Soukromá osoba"a intervaly hodnot jejich kódování jsou shrnuty níže v tabulce 4.2. Data o platební morálce však byla sbírána a uchovávána nesystematicky a musela být ze souboru vyloučena.

Proměnná	**Bodový rozsah**
Délka pracovního poměru DPP	$0-10$
Pracovní zařazení PZ	$0-10$
Věk	$2-15$
Vzdělání V	$0-10$
Platební morálka PM	$0-30$
Roční splátky/Roční příjem RP	$0-30$
Záloha	$0-40$
Trvání smlouvy TS	$3-15$
ObjektRisiko ObR	$0-20$
Stáří objektu SO	$4-20$

Tabulka 4.2: Seznam proměnných scoring "Soukromá osoba"

Naposledy se podíváme na seznam proměnných vyskytujících se ve scoringu "Obchodní společnost"[29], veškerá data jsou v elektronické podobě na přiloženém CD, ve složce company, soubor *companies.xlsx*. V modelu "Společnost"vystupují vedle společných proměnných následující specifické proměnné:

- Vlastní kapitál/Bilanční suma VKB - proměnná hodnotící poměr mezi ekonomickými ukazateli společnosti za předchozí účetní období,
- Obrat (v tisících Kč za rok) - obrat, kterého obchodní společnost dosáhla v předchozím účetním období,
- EcoRatio - uměle vytvořená proměnná hodnotící další ekonomické ukazatele společnosti za předchozí účetní období, při jejím vytváření je např. zohledněn zisk společnosti,
- Stáří společnosti DP - doba, po kterou společnost funguje na trhu, datum založení společnosti lze dohledat a ověřit v obchodním rejstříku.

Seznam proměnných vyskytujících se ve scoringu "Společnost"včetně intervalů hodnot jejich kódového ohodnocení je shrnut níže v tabulce 4.3. Také zde nešlo použít dat proměnné Platební morálka.

Proměnná	**Bodový rozsah**
Vlastní kapitál/Bilanční suma VKB	$0-15$
Obrat	$0-10$
EcoRatio	$0-20$
Stáří společnosti DP	$0-20$
Platební morálka PM	$0-30$
Záloha	$0-40$
Trvání smlouvy TS	$3-15$
ObjektRisiko ObR	$0-20$
Stáří objektu SO	$4-20$

Tabulka 4.3: Seznam proměnných scoring "Společnost"

[29]Pojem obchodní společnost může být v této práci nahrazen kratším pojmem "společnost"nebo veřejností hojně užívaným pojmem "firma".

Kapitola 5

Aplikace logitového modelu na reálná data

V této kapitole bude vyhodnocena scoringová funkce používaná určité období reálně v praxi daným poskytovatelem úvěru.

Ke zpracování aplikační části bude použit následující software:

- MS excel - soubor k základnímu zpracování dat a provedení jednoduchých úkonů,
- gretl[30] - k vytvoření modelu scoringu, testování významnosti jednotlivých parametrů a vyhodnocení celého modelu prostřednictvím ukazatelů a testů nastíněných v kapitole 3 a
- matlab - k verifikaci modelu na validačních datech.

Veškeré modely budou zpracovány s nastavenou základní hodnotou cutoff prostřednictvím pravděpodobnosti 0,5, to znamená, že klient je vyhodnocen jako defaultní, je-li u něj pravděpodobnost jevu "default"větší než 50 %. Pro scoring "Živnostník"budou užity i jiné hodnoty cutoff a zhodnocen jejich vliv na úspěšnost modelu.

[30]Gretl je volně dostupný statistický software, vyvinutý v rámci projektu GNU, dostupný z gretl.sourceforge.net. Na stejné internetové adrese lze i dohledat více informací k tomuto programu. Základní manuál, jak aplikovat gretl na výpočty použité v této práci tvoří přílohu C.

5.1 Získání dat

Pro vytvoření modelu scoringu je v prvé řadě nezbytné porozumět systematice získávání dat a ohodnocení. Data byla získána během jednoho roku nabízení účelového financování. Díky způsobu vyhodnocování klientů, kdy klienti zamítnutí na základě scoringové funkce jakožto nedostatečně bonitní, jsou ještě v druhém kole posuzováni pracovníky CRM (Credit Risk Managmentu), kteří je mnohdy ještě schválí [31], máme i informaci o chybě druhého druhu. Soubor dat, na kterém je vyhodnocován model scoringu, je tak tvořen všemi klienty, kteří byli vyhodnocováni prostřednictvím scoringu a kterým byl poskytnut úvěr, ať již na základě jen samotné funkce scoringu nebo na základě pozdějšího rozhodnutí pracovníka CRM. Skutečnost, zda se klient projeví jako bonitní nebo zda u něj nastane default, byla sledována po dva roky od uzavření smlouvy.

Při zpracování modelu scoringu ani jeho vyhodnocení nelze uvažovat klienty, kterým nebyl úvěr poskytnut, neboť data o nich nejsou již nadále uchovávána a s ohledem na to, že s nimi nebyla uzavřena žádná úvěrová smlouva nelze ani hodnotit jejich schopnost plnit své závazky, tedy jejich bonitnost a defaultnost. Za základní soubor tedy považujeme pozorování klientů, kterým byl poskytnut úvěr. Klienti, kterým byl poskytnut úvěr jen na základě scoringu, jsou považováni za bonitní klienty, naopak klienti, kterým byl poskytnut úvěr až na základě dalšího posouzení pracovníky CRM jsou považováni za klienty, kteří byli scoringem vyhodnoceni jako nebonitní (defaultní). Chybou I. druhu v rámci našeho základního souboru rozumíme situaci, kdy klient byl scoringem vyhodnocen jako bonitní, ač u něj nastal default. Chybou II. druhu rozumíme situaci, kdy klient byl scoringem vyhodnocen jako defaultní (úvěr mu byl pak poskytnut na základě rozhodnutí pracovníků CRM) a on byl schopen svůj úvěr splácet bez výskytu defaultu.

Při získávání dat byla zjištěna nekonzistentnost ve sběru dat Platební morálka, kdy nejprve byla brána nejvyšší hodnota platební morálky (počet dní, kolik je klient po splatnosti ze svých dosavadních smluv) a později (u méně smluv) již průměrná platební morálka za poslední půl roku. Celá proměnná Platební morálka byla pro model scoringu Živnostník upravena dle prvotního přístupu.

[31] Zpravidla rozhodnout o nutnosti dodatečného zajištění třetí osobou.

5.2 Model s jednou nezávisle proměnnou - základní koncepce logitového modelu

Pro vysvětlení základní myšlenky logitové regrese si uděláme jednoduchý příklad, kdy default klienta by závisel jen na jedné nezávislé proměnné. Vstupní data vybereme ze scoringové funkce živnostník. Jako závisle proměnnou zvolíme default30, klient je více jak 30 dní po splatnosti, případ, kdy nastane default nazveme D a kdy nenastane B (jako bonitní klient). Za nezávisle proměnnou zvolíme zisk, rozdělený do kategorií, viz. tabulka 3.1. Jednotlivé intervaly označíme jen $kat1, \ldots, kat6$[32], každému klientovi přiřadíme jednu kategorii, ve které se vyskytuje jeho zisk (podrobnosti v tabulce 5.1).

Default(30)	**Kat1**	**Kat2**	**Kat3**	**Kat4**	**Kat5**	**Kat6**	**Celkem**
D	20	5	14	4	9	13	65
B	168	8	30	35	47	97	385
Celkem	188	13	44	39	56	110	450

Tabulka 5.1: Kontingenční tabulka podle defaultnosti a zisku klienta

V případě dvou proměnných lze logitovou regresi řešit prostřednictvím kontingenční tabulky a dle rozdělení hodnot v ní dopočítat následující hodnoty.

Pravděpodobnost, že u klienta nastane default (D):

$$P(D) = \frac{65}{450} = 0,144.$$

Pravděpodobnost, že u klienta nastane default (D), jestliže má zisk Kat1, lze dopočítat jednoduše jako podmíněnou pravděpodobnost jevu "nastane default"za podmínky, že klient má zisk Kat1:

$$P(D|Kat1) = \frac{20}{188} = 0,106.$$

Pravděpodobnost, že klient bude bonitní (B), jestliže má zisk Kat1:

$$P(B|Kat1) = \frac{168}{188} = 0,894.$$

[32]Tento příklad je jen názorný, proto nejsou důležité přesné hodnoty proměnných v kategorii

Pravděpodobnost je dále možné vyjádřit jako šanci (odds). Šance, že u klienta nastane default (D):

$$odds(D) = \frac{65}{385} = 0,169.$$

Šance, že u klienta nastane default (D), jestliže má zisk Kat1:

$$odds(D|Kat1) = \frac{20}{168} = 0,119.$$

Šance, že klient bude bonitní (B), jestliže má zisk Kat1:

$$odds(B|Kat1) = \frac{168}{20} = 8,400.$$

Šance a pravděpodobnost se dá mezi sebou snadno převádět prostřednictvím rovnice 3.20.

$$odds(D|Kat1) = \frac{P(D|Kat1)}{1 - P(D|Kat1)} = \frac{0,106}{1 - 0,106} = 0,119$$

a opačně

$$P(D|Kat1) = \frac{odds(D|Kat1)}{1 + odds(D|Kat1)} = \frac{0,119}{1 + 0,119} = 0,106.$$

Logaritmováním šancí poté dostaneme lineární vztah mezi logaritmem šancí závislé proměnné a nezávislými proměnnými. Pro ukázku získání zápisu této rovnice si naši nezávisle proměnnou ještě zjednodušíme pouze na binární. Klienty se ziskem z $kat1 - kat3$ označíme za klienty s nízkým ziskem (L) a klienty se ziskem z $kat4 - kat6$ označíme za klienty s vysokým ziskem (H), četnosti, pravděpodobnosti a šance takto definovaných proměnných jsou shrnuty níže v tabulce 5.2.

Default(30)	**Zisk L**	**Zisk H**	**Celkem**
D	39	26	65
B	206	179	385
Celkem	245	205	450
$P(D)$	$0,159$	$0,127$	$0,144$
$P(B)$	$0,841$	$0,873$	$0,856$
$odds(D)$	$0,189$	$0,145$	$0,169$
$odds(B)$	$5,282$	$6,885$	$5,923$

Tabulka 5.2: Kontingenční tabulka podle defaultnosti a zisku klienta 2

Z tabulky vidíme, že u klienta s nízkým příjmem je pravděpodobnost jeho defaultu $15,9\ \%$, zatímco u klienta s vysokým příjmem je tato pravděpodobnost pouze $12,7\ \%$.

Nyní odvodíme rovnici pro logit závisle proměnné $default30$ na základě nezávisle proměnné $Zisk$.

$$\begin{aligned}\ln[odds(D|Zisk)] &= \ln[odds(D|ZiskL)] + \ln[odds(D|ZiskH) - odds(D|ZiskL)] \cdot Zisk = \\ &= \ln[odds(D|ZiskL)] + \tfrac{\ln[odds(D|ZiskH)]}{\ln[odds(D|ZiskL)]} \cdot Zisk = \\ &= \ln(0,189) + [\ln(0,145) - \ln(0,189)] \cdot Zisk = \\ &= -1,664 - 0,265 \cdot Zisk\end{aligned}$$

Pravděpodobnost pak můžeme zpětně dopočítat dle rovnice 3.23.

$$\pi(X) = \frac{1}{1 + e^{-(-1,664-0,265 \cdot Zisk)}}$$

Můžeme dopočítat hodnoty, které taktéž vidíme v tabulce 5.2 a interpretovat výsledky, že máme-li klienta s nízkým příjmem je pravděpodobnost jeho defaultu $15,9\ \%$, zatímco u klienta s vysokým příjmem je tato pravděpodobnost pouze $12,7\ \%$. [33]

Vidíme, že hodnocení úvěruschopnosti klienta jen na základě jedné proměnné by nepřinášelo postačující výsledky, proto přejdeme k modelu scoringu vybudovaném na celé skupině proměnných.

5.3 Model scoringu Živnostník

Jako první se budeme věnovat užívanému modelu scoringu určenému pro vyhodnocení bonity živnostníka. Získaná data jsme očistili o data, v nichž chybí údaj o některé ze sledovaných nezávisle proměnných. Data o závisle proměnné defaultu, tak jak jsme si ho definovali výše - tedy default30 jako situaci, kdy je klient 30 dní po splatnosti, a default90 potom stav, kdy je klient 90 dní po splatnosti, jsme vyhledali a přidali k souboru s informacemi o klientovi z pravidelných deliquency reportů společnosti. [34]

[33] Defaultní nastavení většiny software pro logitovou regresi a tedy i pro vyhodnocování scoringu je 50 %, při takovém nastavení by byl každý náš klient na základě jen proměnné Zisk označen za bonitního.

[34] Ze souboru nezávisle proměnných jsme vyloučili také proměnnou "Industry", která dle informací ze společnosti byla ve scoringové funkci dříve, ale v tuto dobu je do ní jen automaticky generována 1, která je jako příznak publikována v případě, že u daného klienta má být proveden rating namísto scoringu.

Takto upravený datový soubor obsahuje celkem 769 pozorování. V tomto souboru nastal default90 celkem 36krát a default30 celkem 100krát (default30 obsahuje i veškerá pozorování default90). Přitom 152 klientů z toho bylo schváleno až pracovníky CRM, tj. poté, kdy byli skóringovou funkcí vyhodnoceni jako nebonitní. Z toho jen u 17 z nich skutečně došlo k defaultu90 a 27 potom k defaultu30. Celkové vyhodnocení chyb I. i II. druhu je v tabulkách 5.3 a 5.4 níže. Jak již bylo zmíněno v kapitole 5.1 základní soubor tvoří data o klientech, kterým byl poskytnut úvěr. Chybou I. druhu rozumíme v rámci našeho základního souboru dat situaci, kdy klient byl pouhým scoringem vyhodnocen jako bonitní, ač u něj nastal default. Chybou II. druhu rozumíme poté situaci, kdy klient byl scoringem vyhodnocen jako defaultní (úvěr mu byl později poskytnut na základě rozhodnutí pracovníků CRM) a on byl schopen svůj úvěr splácet bez výskytu defaultu.

Odhad—Skutečnost	**1**	**0**	**Počet odhadů**
1 default30	27	125	152
0 bonitní klient	73	544	617
Počet skutečných jevů	100	669	769

Tabulka 5.3: Úspěšnost stávajícího modelu - default30

Ač je scoringový model budován nad závisle proměnnou default30, bude zde vyhodnocena i úspěšnost predikce defaultu90, jeho shrnutí je v tabulce 5.4.

Odhad—Skutečnost	**1**	**0**	**Počet odhadů**
1 default90	17	135	152
0 bonitní klient	19	598	617
Počet skutečných jevů	36	733	769

Tabulka 5.4: Úspěšnost stávajícího modelu - default90

Celková úspěšnost scoringového modelu:

Typ defaultu	**Úspěšnost**
$Default30$	$74,25\%$
$Default90$	$79,97\%$

Tabulka 5.5: Úspěšnost stávajícího modelu

U současné skórovací funkce nemáme k dispozici parametry β, které byly použity při vyhodnocování scoringu, známe ale data o vstupních proměnných rozdělených do intervalů, cutoff dělící klienty na bonitní jen na základě scoringu a klienty, u kterých je nutné ještě posouzení pracovníky CRM [35] a skutečném defaultu. V následujícím kroku posoudíme vhodnost zvolených vstupních proměnných na odhadlém modelu scoringové funkce vypracovaném nad stejnými vstupními proměnnými.

5.3.1 Základní analýza vstupních proměnných

Ze všech dat si vybereme prvních 450 (cca 60 %) pro vytvoření našeho modelu a zbývající data si necháme pro otestování našich závěrů. Validace úspěšnosti jednotlivých modelů navržených v rámci scoringu Živnostník bude vyhodnocena v subkapitole 5.3.6 shrnující scoring Živnostník. Validační soubor dat obsahuje celkem 319 pozorování. Veškeré nezávislé proměnné vystupující v našem modelu jsou kategoriálního typu (rozdělené do intervalů) a jejich kódování je dané společností, která nám data poskytla. Soubor všech proměnných a jejich kódové ohodnocení je v příloze A[36].

Na základě Pearsonova koeficientu korelace jsme určili stupeň korelovanosti mezi proměnnými, při čemž nejvyšší hodnotu nabyl Pearsonův koeficient v případě proměnných Zisk a Obrat a dosahoval hodnoty $0,56$[37], korelovanost těchto dvou proměnných byla následně potvrzena testem s testovacím kritériem dle vzorce 3.5. Korelační matice je uvedena níže v tabulce, testy pro párové korelační koeficienty, jsou poté součástí přílohy na CD ve složce enterpreneur, soubor enterpreneurs.xlsx.

[35] Toto nám určuje množství scoringem správně a špatně vyhodnocených případů.

[36] U rozdělení do kategorií a jejich číselného kódování dále zkontrolujeme, zda splňují podmínku lineárního trendu mezi nezávislou proměnnou a logitem.

[37] Korelovanost mezi těmito proměnnými se jeví logická i z ekonomického úhlu pohledu na tato data a finanční instituce předpokládají korelovanost mezi jednotlivými finančními ukazateli živnostníka i obchodních společností, přesto by nechtěli eliminovat používání finančních ukazatelů ve svém modelu scoringu.

$\boldsymbol{R}$	ObR	SO	Zaloha	TS	DP	Obrat	Zisk	PM
ObR	1,000	0,319	−0,250	−0,047	−0,045	0,062	0,061	0,112
SO	0,319	1,000	−0,170	−0,068	−0,017	0,037	0,058	0,090
Zaloha	−0,250	−0,170	1,000	0,209	−0,027	−0,443	−0,369	−0,292
TS	−0,047	−0,068	0,209	1,000	0,004	−0,074	−0,147	−0,044
DP	−0,045	−0,017	−0,027	0,004	1,000	0,174	0,006	0,054
Obrat	0,062	0,037	−0,443	−0,074	0,174	1,000	0,557	0,190
Zisk	0,061	0,058	−0,369	−0,147	0,006	0,557	1,000	0,129
PM	0,112	0,090	−0,292	−0,044	0,054	0,190	0,129	1,000

Tabulka 5.6: Korelační matice

Testem multikolinearity na základě Farrarova-Glauberova testu definovaného předpisem 3.6 byla na hladině významnosti 5 % zamítnuta hypotéza H_0, že se párové koeficienty významně neodlišují od nuly. Hodnota testovacího kritéria 743,37 je větší než tabelovaná hodnota 95 % kvantilu χ^2 rozdělení (přibližně 17).

Reif [1] uvádí nutnost sílu kolinearity nějak měřit a vymezuje Variance Inflation Factor:

$$VIF_j = \frac{1}{1 - R_j^2}\,,\ kde\ j = 1, \ldots,\ k, \tag{5.1}$$

přičemž platí, že koeficienty VIF_j odpovídají diagonálním prvkům inverzní matice k matici $\boldsymbol{R_{XX}}$.

Kolinearita	ObR	SO	Zaloha	TS	DP	Obrat	Zisk	PM
VIF	1,17	1,13	1,49	1,06	1,05	1,69	1,52	1,10

Tabulka 5.7: Míra kolinearity proměnných Živnostník

Taktéž Reif [1] a Cipra [17] uvádí, že projevem silné multikolinearity je hodnota vyšší než 10, některými autory dokonce uváděno 100. Na základě našich koeficientů VIF není tedy nutné kvůli zjištěné kolinearitě některé proměnné nebo celý model transformovat. U hodnocení výsledků bychom si přesto měli všimnout velikosti chyby odhadů, neboť mezi daty může existovat i jiná forma závislosti než párová, která právě nejvíce ovlivní odhad parametrů a velikost jejich chyb.

Korelační matici proměnných stejně jako hodnoty VIF můžeme získat taktéž spuštěním skriptu logitENT.inp v gretlu, který je uložen na přiloženém CD, ve složce enterpreneur.

5.3.2 Odhad parametrů

Odhad parametrů v logitové regresi je prováděn metodou maximální věrohodosti. Pro vypracování tohoto odhadu použijeme software gretl. Skript pro vytvoření celého modelu je přiložen na CD, které je součástí této práce, složka enterpreneur, soubor logitENT.inp, výstup tohoto skriptu je poté uložen v textovém souboru enterpreneurMODEL.txt. Skript si lze taktéž přímo spustit v software gretl, stručný návod, jak skripty v gretlu používat, je obsahem přílohy C a taktéž elektronicky uložen na CD, soubor Gretl_manual.docx.

Odhad parametrů je pro srovnání proveden i v matlabu - zdrojový kód tvoří přílohu B[38].

Metoda maximální věrohodnosti je vykonávána iteračně, také software gretl využívá maximalizaci sumy přirozených logaritmů věrohodnostní funkce (rovnice 3.26) namísto maximalizace součinu věrohodnostní funkce (rovnice 3.25). Po ukončení iterací jsme získali hodnotu součinu logaritmu věrohodnostní funce: $\ln[l(\beta)] = -152,030$.

Jako výsledek iterační metody dostaneme odhad parametrů β_j. V gretlu můžeme skript doplnit, aby nám zároveň odhadl i chybu těchto odhadů, respektive jejich směrodatnou odchylku (SE rovnice 3.31). V tabulce 5.8 máme uvedeny veškeré odhady parametrů β_j včetně odhadu jejich chyb. V tabulce 5.8 jsou vedle odhadů parametrů β_j a odhadu jejich chyb uvedeny ještě hodnoty Waldovy statistiky pro jednotlivé parametry dopočítané na základě vzorce 3.32 a p-hodnoty příslušející hodnotám těchto statistik, s těmito dvěma hodnotami budeme pracovat dále v této práci při vyhodnocení významnosti jednotlivých parametrů.

[38]Elektronická podoba skriptu je na přiloženém CD, složka Matlab.

Proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$Konstanta$	$3,083$	$1,114$	$2,768$	$0,06$
$ObjektRiziko(ObR)$	$0,025$	$0,055$	$0,464$	$0,643$
$StariObjektu(SO)$	$-0,042$	$0,041$	$-1,013$	$0,311$
$Zaloha$ (Z)	$-0,083$	$0,016$	$-5,085$	$< 0,001$
$TrvaniSmlouvy(TS)$	$-0,195$	$0,056$	$-3,451$	$< 0,001$
$DelkaPodnikani(DP)$	$-0,075$	$0,027$	$-2,731$	$0,006$
$Obrat$	$-0,029$	$0,024$	$-1,201$	$0,230$
$Zisk$	$-0,053$	$0,027$	$-1,957$	$0,050$
$PlatebniMoralka(PM)$	$-0,002$	$0,011$	$-0,190$	$0,850$

Tabulka 5.8: Odhad parametru β modelu 1 Živnostník

Na základě těchto bodových odhadů parametrů můžeme zapsat rovnici logitové regrese při použití všech závislých proměnných.

$$\begin{aligned} logit[\pi(x)] \quad &= 3,083 + 0,025 \cdot ObR - 0,042 \cdot SO - 0,083 \cdot Z - \\ &-0,195 \cdot TS - 0,075 \cdot DP - 0,029 \cdot Obrat + 0,053 \cdot Zisk - 0,002 \cdot PM \end{aligned}$$

Podmíněné pravděpodobnosti (nastání defaultu) dopočteme dle vzorce:

$$\pi(\hat{D}30) = \frac{1}{1 + e^{-logit[\pi(x)]}}$$

Podmíněná pravděpodobnost jevu, že bude klient bonitní, je doplňkem k této pravděpodobnosti. Zkusíme si do výše uvedeného vzorce dosadit hodnoty jednoho z klientů žádajícího o úvěr:

$$\begin{aligned} logit[\pi(x)] \quad &= 3,083 + 0,025 \cdot 9 - 0,042 \cdot 10 - 0,083 \cdot 25 - 0,195 \cdot 12 - \\ &-0,075 \cdot 20 - 0,029 \cdot 0 + 0,053 \cdot 0 - 0,002 \cdot 0 = \mathbf{-3,006}. \end{aligned}$$

Pravděpodobnost nastání defaultu u daného klienta dopočteme potom:

$$\pi(\hat{D}30) = \frac{1}{1 + e^{3,006}} = \mathbf{0,047}.$$

Jako rozhodná pravděpodobnost se v našem modelu používá hodnota $0,5$, jakmile pozorování má $\pi(\hat{D}30) > 0,5$ je u klienta predikován default [39]. Klient, jehož data jsme

[39] S ohledem na to, že se celou dobu vyskytujeme v $\langle 0,\ 1\rangle$, jsou výsledky velmi citlivé na zaokrouhlování, proto pro validaci našich výsledků je nutno pracovat s přesnými hodnotami odhadnutých parametrů β. Validační algoritmus je proto pro naše potřeby připraven v matlabu.

zkusili dosadit do rovnice, by byl tak na základě scoringu vyhodnocen jako bonitní neboť pravděpodobnost nastání defaultu u něj nedosahuje hodnoty větší než 50 %.

Pravděpodobnosti (cutoff), od které chceme klienta považovat skutečně za defaultního lze poté upravit podle účelu a kvality odhadů daného modelu. Zohlednění a nastavení různého cutoff bude provedeno až pro upravený model scoringu, respektive bude vyhodnocena úspěšnost modelu pro různé hodnoty nastaveného cutoff, samotná volba cutoff musí bý nechána na manažerském rozhodnutí ve společnosti, neboť dle hodnoty stanovené cutoff se může zúžit případně rozšířit tzv. "šedá zóna".

Odhad šance pro celý model dopočteme pomocí odhadnutých parametrů následovně:

$$odds(\hat{\pi}(D30)) = e^{-(logit[\pi(x)])}$$

Pro jednotlivé proměnné je pak šance vymezena následovně, např. pro zisk:

$$odds(\pi(\hat{D30}|Zisk)) = e^{-0{,}053 \cdot Zisk}$$

Celkově bylo správně předpovězeno 388 jevů z 450, tj. 86,2 %. Z toho bylo 9krát správně předpovězeno, že nastane default a 379krát byl správně odhadnut bonitní klient. K chybě II. druhu (předpovím default, ač ten skutečně nenastane) došlo jen v 6 případech. Chyby I. druhu jsme se na základě tohoto modelu dopustili 56krát (do portfolia jsme pustili 56 defaultních zákazníků). V tabulce 5.9 je uvedena úspěšnost modelu:

Skutečnost—Odhad	**1**	**0**	**Počet odhadů**
1 default	9	6	15
0 bonitní klient	56	379	435
Počet skutečných jevů	65	385	450

Tabulka 5.9: Úspěšnost modelu 1 Živnosntík - default30

Čistě na základě tohoto modelu scoringu bychom úvěr poskytli 435 žadatelům a úvěr zamítli jen 15 žadatelům. Z těchto výsledků lze usuzovat, že hranice cutoff nastavená na hodnotu pravděpodobnosti 0,5 je příliš vysoká pro vymezení defaultu klienta a pokud chceme cutoff používat pro vymezení šedé zóny a některé klienty i nadále postupovat k posouzení pracovníkům CRM měli bychom hranici cutoff, pod kterou je klient vyhodnocován jen na základě scoringu, snížit. Vliv jiného nastavení hranice cutoff na úspěšnost modelu bude zkoumán v celkovém vyhodnocení modelu "Živnostník".

Pro vyhodnocení kvality modelu jsme využili míry vybudované na věrohodnostní funkci, dle vzorce 3.37 bylo dopočteno Akaikeho kritérium a dle vzorce 3.38 poté Schwarzovo kritérium.

$$AIC = -2\ln(l) + 2(k+s) = -2(-152,030) + 2(1+8) = 322,06,$$

$$SC = -2\ln(l) + (k+s)\ln(n) = -2(-152,030) + 2(1+8)\ln(450) = 359,04.$$

Tato kritéria však nemají samostatně významnou vypovídací hodnotu, ale slouží jen pro srovnání modelu s jiným navrženým modelem. Tyto hodnoty tak využijeme dále v této práci ke srovnání s námi navrženým modelem. Čím nižší hodnoty kritérium nabývá, tím je model hodnocen jako kvalitnější.

5.3.3 Význam jednotlivých parametrů

Poté, co máme odhady parametrů a tedy návrh logitové regrese, měli bychom ještě otestovat význam jednotlivých vstupních proměnných. V tabulce 5.8 si můžeme vedle odhadů parametrů a odhadu chyby těchto parametrů všimnout i dalších dvou sloupců proměnných. Ve třetím sloupci jsou dopočítány na základě vzorce 3.32 hodnoty Waldových statistik, tedy hodnoty testovacího kritéria při testování významnosti jednotlivých parametrů. V posledním sloupci je stanovena p-hodnota, vztahující se k Waldovu testu, na které je vyhodnocována významnost vybraného parametru.

Výpočet Waldovy statistiky např. pro proměnnou Zisk výpočet probíhá následovně:

$$W_{Zisk} = \frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)} = \frac{0,053}{0,027} = -1,96.$$

$$W^2_{Zisk} = (-1,96)^2 = 3,84.$$

Při využívání softwarového testování hypotéz, testujeme hypotézu $H_0 : \beta_j = 0$ proti alternativě $H_1 : \beta_j \neq 0$, se namísto vymezení kritického oboru na zvolené hladině významnosti α pro testovací kritérium využívá p-hodnota daného testu, ta zjednodušeně představuje pravděpodobnost, že hodnota testovacího kritéria za platnosti H_0 se bude nacházet v kritickém oboru. Čím menší p-hodnota, tím nepravděpodobněji by takového výsledku (testovací kritérium nabývá hodnoty z kritického oboru) za předpokladu platnosti H_0 bylo dosaženo.

Stejně jako stanovujeme hladinu významnosti α pro testovací kritérium musíme stanovit mezní α pro p-hodnotu. Doporučovaná hodnota je 10 %. V tabulce 5.8 vidíme, že některé proměnné mají vyšší p-hodnotu Waldova testu, tzn. lze vyvodit závěr, že dané parametry jsou pro regresi nevýznamné (nelze zamítnout hypotézu H_0 o jejich nulovosti).

Odstraňovat proměnné budeme iteračně po jedné, nejprve odstraníme tu s nejvyšší p-hodnotou. Proměnné odebíráme postupně, dokud p-hodnoty uvšech proměnných nejsou nižší než 10 % (doporučení v [5]). Zjednodušený zápis postupu je níže v tabulce 5.10. Veškeré výpočty jsou provedeny opět v software gretl, skript je uložen na CD ve složce enterpreneur logitENTomit.inp, výstup skriptu je pak v téže složce v souboru logitEN-Tomit.txt.

V tabulce je vždy uveden odhad parametrů proměnné, která má být odstraněna, včetně hodnoty příslušné Waldovy statistiky a p-hodnoty. Tento odhad je vždy proveden v novém modelu po odstranění předchozí nevýznamné proměnné, odhady první odstraňované proměnné, tak pocházejí z výše navrženého modelu.

V tabulce jsou dále uvedena kritéria míry vybudovaná na věrohodnostních funkcích modelů, sloužící k porovnání modelů mezi sebou, Aikeho a Schwarzovo kritérium. Čím nižší hodnoty nabývají tato kritéria tím je model hodnocen jako kvalitnější. Po odebrání proměnné s nejvyšší p-hodnotou nám vznikne nový model, u kterého vždy zkontrolujeme testem postaveném na věrohodnostních funkcích, AIC nebo SC, zda se kvalita nového modelu skutečně zvýšila.

Za celé čtyři iterace snižování počtu proměnných se snížila hodnota logaritmu věrohodnosti o 1,364 (negativní vliv), původní hodnota $\ln[l(\beta_0)] = -152,030$, tak byla snížena na $\ln[l(\beta_1)] = -153,394$. Hodnoty všech tří zkoumaných asociačních měr se naopak pravidelně snižovaly, ve srování s logaritmem věrohodnosti výrazněji (pozitivní vliv).

Vyhodnocení modelu se všemi proměnnými				
Test poměrem věrohodnosti			AIC	$322,060$
			SC	$359,043$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
Platební morálka	$-0,002$	$0,011$	$-0,190$	$0,850$
Vyhodnocení modelu po odstranění proměnné Platební morálka				
Test poměrem věrohodnosti			AIC	$320,095$
			SC	$352,969$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$ObjektRiziko(ObRisk)$	$0,025$	$0,055$	$0,464$	$0,643$
Vyhodnocení modelu po odstranění proměnné ObjektRiziko				
Test poměrem věrohodnosti			AIC	$318,299$
			SC	$347,063$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$StariObjektu(SO)$	$-0,036$	$0,038$	$-0,959$	$0,338$
Vyhodnocení modelu po odstranění proměnné StariObjektu				
Test poměrem věrohodnosti			AIC	$317,302$
			SC	$341,958$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$Obrat$	$-0,029$	$0,025$	$-1,183$	$0,2368$
Vyhodnocení modelu po odstranění proměnné Obrat				
Test poměrem věrohodnosti			AIC	$316,788$
			SC	$337,334$
Ostatní proměnné mají p-hodnotu menší než 0,10				

Tabulka 5.10: Postupné odstraňování nevýznamných β_j - Živnostník

Celková úspěšnost modelu hodnocená jakožto procento správně předpovězených hodnot dosahuje v tuto chvíli 390 predikovaných jevů (na počátku 388), což přestavuje 86,7 % pozorování. V novém modelu byly eliminovány dva výskyty chyby II. druhu, v tuto chvíli modelem předpovíme 4krát default, ač ten nakonec skutečně nenastane a 56krát se dopustíme chyby I. druhu, kdy vyhodnotíme klienta jako bonitního, ač on při splácení selže.

Odhad parametrů spolu s odhadem jejich chyb(SE), Waldových statistik a p-hodnot je zapsán v následující tabulce 5.11. V tabulce vidíme, že p-hodnoty pro tyto parametry zbývajících proměnných dosahují maximálně výše $0,01$, z čehož lze vyvodit závěr, že zbývající proměnné jsou významné pro model.

Proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$Konstanta$	$2,715$	$0,769$	$3,531$	$< 0,001$
$Zaloha(Z)$	$-0,074$	$0,014$	$-5,505$	$< 0,001$
$TrvaniSmlouvy(TS)$	$-0,196$	$0,057$	$-3,469$	$0,001$
$DelkaPodnikani(DP)$	$-0,084$	$0,026$	$-3,2471$	$< 0,001$
$Zisk$	$-0,061$	$0,024$	$-2,561$	$0,010$

Tabulka 5.11: Model 2 Živnostník - odhad β

Finální model po odstranění všech nevýznamných proměnných lze zapsat do regresní rovnici pro vyjádření logitu:

$$logit[\pi(x)] \quad = 2,715 - 0,074 \cdot Z - 0,196 \cdot TS - 0,084 \cdot DP - 0,061 \cdot Zisk$$

Vidíme, že nejvýrazněji se změnil vliv zisku na logit, který zcela obrátil orientaci své závislosti (to mohlo být způsobeno vztahy závislosti mezi veličinami). V novém modelu je pravděpodobnost nastání defaultu negativně závislá na všech vstupních proměnných, to znamená, zvyšujeme-li hodnotu jakékoli vstupní proměnné, snižujeme tím zároveň predikovanou pravděpodobnost, že u klienta nastane default.

Podmíněné pravděpodobnosti (nastání defaultu) dopočteme dle vzorce:

$$\pi(\hat{D}30) = \frac{1}{1 + e^{-(2,715-0,074 \cdot Z-0,196 \cdot TS-0,084 \cdot DP-0,061 \cdot Zisk)}}$$

Při znalosti kategoriálních proměnných a jejich hodnot lze tak na základě výše uvedených rovnic modelu dopočítat maximální logit a jemu příslušející podmíněnou pravděpodobnost defaultu a minimální logit spolu s příslušející pravděpodobností. Logit pro minimální hodnoty vstupních proměnných dosahuje výše

$$logit[\pi(x)] \quad = 2,715 - 0,074 \cdot 0 - 0,196 \cdot 3 - 0,084 \cdot 0 - 0,061 \cdot 0 = 2,127,$$

což odpovídá pravděpodobnosti nastání defaultu ve výši:

$$pi(max) = \frac{1}{1 + e^{-2,127}} = 0,893.$$

Pro maximální hodnoty vstupních proměnných pak dostáváme hodnoty:

$$logit[\pi(x)] \quad = 2,715 - 0,074 \cdot 40 - 0,196 \cdot 15 - 0,084 \cdot 20 - 0,061 \cdot 20 = -6,085,$$

což odpovídá pravděpodobnosti nastání defaultu výše:

$$pi(min) = \frac{1}{1 + e^{6,085}} = 0,002.$$

Zvolíme-li cutoff (hranici, od které vyhodnotíme klienta jako defaultního) větší než 0,893, pak budou všichni žadatelé o úvěr vyhodnoceni jako bonitní, neboť pro žádného z žadatelů nemůže být na základě daných vstupních proměnných predikována vyšší pravděpodobnost defaultu. Zvolíme-li naopak hranici cutoff menší než 0,002 budou všichni klienti vyhodnoceni jako defaultní, eliminujeme tím chybu I. druhu a model je tak 100% úspěšný z pohledu odhadu defaultních klientů, ale my nikomu nesposkytneme úvěr.

Na základě nezamítnutí hypotézy H_0 o nulovosti některých parametrů β_j lze původní vyhodnocovaný model považovat za nekvalitní z hlediska používání nevýznamných proměnných. Důvody pro jejich používání jsou ovšem dány nejen korporátními směrnicemi, ale v některých případech i obchodním pohledem na model scoringu. V praxi je například nepřijatelné vyloučit proměnnou Obrat z modelu, ač byla na zvolené hladině významnosti ze zkoumaného modelu vyloučena.

Při vyhodnocování závěru bychom měli navíc vzpomenout možný vliv multikolinearity mezi vstupními proměnnými, která byla na základě koeficientu VIF vyhodnocena jako nevýznamná, ve vstupních datech ovšem může navíc existovat korelovanost mezi lineární kombinací několika vstupních proměnných a jinou vstupní proměnnou. Jedním z hlavních důsledků korelovanosti mezi daty je právě nepřesnost v odhadech parametrů β_j a některé proměnné se mohou v důsledku kolinearity tvářit jako nevýznamné, ač tomu tak ve skutečnosti není. Podíváme-li se ale na konečnou verzi rovnice modelu, vidíme, že se oproti původní rovnici snížily i odhady chyb parametrů, z čehož lze usuzovat na zlepšení celkové kvality modelu i odhadu jejich parametrů. Naše závěry o správnosti vyloučení

některých proměnných z důvodu nevýznamnosti jejich parametrů se pokusíme potvrdit v následující kapitole.

5.3.4 Návrh a tvorba nového modelu

V tomto kroku bude navrhován nový model scoringu jen ze vstupních proměnných, které jsou již používány v současném modelu scoringu společnosti. U všech proměnných bude navíc stále užíváno i zvolené kódování jednotlivých proměnných a jejich rozdělení do intervalů. Z důvodu citlivosti údajů vstupujících do modelu scoringu nebyla veškerá nekódovaná data poskytovatelem dat zpřístupněna. Základní analýza takovýchto vstupních proměnných regresního modelu byla provedena již v předchozí kapitole při vyhodnocení používaného modelu. Díky opatřením na ochranu dat nelze provést úplné zhodnocení vlivu jednotlivých proměnných na výstup, přesto bude každé proměnné věnována pozornost i jednotlivě a bude posouzena vhodnost zvoleného "kódování"pro danou veličinu a její začlenění do modelu.

Samotná volba proměnných, které budou zahrnuty do modelu, pak bude provedena prostřednictvím stepwise analýzy, postup blíže popsán v [4], základní princip stepwise analýzy vymezen v kapitole 3.6.2. Pro stepwise analýzu si připravíme veškerá vhodná vstupní data, které bychom pro regresi chtěli použít, tj. v našem případě veškeré proměnné, které v současném scoringu vystupují.

V prvním kroku poté vystavíme model logitové regrese jen na konstantě. Zaznamenáme si jeho věrohodnostní poměr L_0. Do modelu jen s konstantou přidáváme postupně veškeré nezávislé proměnné a porovnáváme hodnotu L_j. Pro přidávání proměnných do modelu si musíme stanovit mezní p-hodnotu, do které budeme přidávat proměnné do modelu. Hosmer [4] uvádí jako vhodnou hodnotu 10 % [40]. Po prostřídání všech proměnných jen s konstantou, vybereme tu proměnnou s nejnižší hodnotou L_j při splnění podmínky stanovené p-hodnoty a přidáme ji do modelu. Do takto vytvořené rovnice opět přidáváme postupně všechny zbývající proměnné a porovnáváme hodnoty L_j, postup opakujeme do té doby, dokud máme proměnné s menší než stanovenou hraniční p-hodnotou.

Provedená stepwise analýza je zjednodušeně zobrazena v tabulce 5.12, v tabulce máme

[40] V případě, nemáme-li vstupní proměnné s významným vztahem k výstupu na základě provedené analýzy jednotlivých proměnných, lze poté užívat vyšší p-hodnoty.

zobrazenou vždy hodnotu L_j pro model po přidání dané proměnné a p-hodnotu příslušející parametru takto přidané proměnné. Celá stepwise analýza je provedena v gretlu a lze ji spustit prostřednictvím skriptu logitENTstepwise.inp, který je uložen na přiloženém CD, ve složce enterpreneur. Nemáme-li program gretl k dispozici a neplánujeme-li jeho instalaci, lze si na přiloženém CD v téže složce otevřít výstup z daného skriptu v textovém formátu v souboru logitENTstepwise.txt .

Stepwise	L_j	*ObR*	*SO*	*Zaloha*	*TS*	*DP*	*Obrat*	*Zisk*	*PM*
const	$-185,83$	$-181,96$	$-185,29$	$\mathbf{-168,90}$	$-175,49$	$-182,29$	$-185,81$	$-183,83$	$-183,83$
$p(X_1)$		$0,008$	$0,328$	$<\mathbf{0.001}$	$<0,001$	$0,006$	$0,888$	$0,851$	$0,034$
Zaloha	$\mathbf{-168,90}$	$-168,44$	$-168,78$		$-162,58$	$-164,50$	$-164,97$	$-166,68$	$-168,90$
$p(X_2)$		$0,345$	$0,629$	<0.001	$\mathbf{0.001}$	$0,002$	$0,008$	$0,038$	$0,943$
TS	$-162,58$	$-161,91$	$-162,47$			$-157,92$	$-158,41$	$-158,49$	$-162,58$
$p(X_3)$		$0,232$	$0,633$	$<\mathbf{0.001}$	$0,004$	$0,002$	$0,07$	$0,005$	$0,975$
DP	$-157,92$	$-157,59$	$-157,77$				$-155,29$	$-153,39$	$-157,92$
$p(X_4)$		$0,395$	$0,588$	$<\mathbf{0.001}$	$0,004$	$0,002$	$0,029$	$0,001$	$0,976$
Zisk	$-153,39$	$-153,36$	$-152,99$				$-152,65$		$-153,36$
$p(X_5)$		$0,789$	$0,372$	$<\mathbf{0.001}$	$<\mathbf{0.001}$	$<\mathbf{0.001}$	$0,234$	$\mathbf{0.001}$	$0,801$

Tabulka 5.12: Stepwise analýza

Proměnné, které na základě stepwise analýzy vybereme do modelu, jsou:
Záloha, TrváníSmlouvy, DélkaPodnikání a Zisk.

Vidíme, že se jedná o stejné proměnné, které nám v modelu zůstaly i v předchozí kapitole, kde byla v podstatě stepwise analýza praktikována opačně. Při splnění základního předpokladu o nezávislosti proměnných bychom měli při aplikaci stepwise analýzy oběma směry získat stejné výsledky. Na základě těchto výsledků lze konstatovat, že počáteční odhalená kolinearita mezi proměnnými významně neovlivňuje výsledky odhadů parametrů β_j.

Dostáváme tak stejnou rovnici logitové regrese jako v předchozí kapitole.

$$logit[\pi(x)] = 2,715 - 0,074 \cdot Z - 0,196 \cdot TS - 0,084 \cdot DP - 0,061 \cdot Zisk$$

A podmíněnou pravděpodobnost nastání defaultu poté dopočteme dle vzorce:

$$\pi(\hat{D}30) = \frac{1}{1 + e^{-(2,715-0,074 \cdot Z-0,196 \cdot TS-0,084 \cdot DP-0,061 \cdot Zisk)}}$$

Úspěšnost modelu lze poté popsat i prostřednictvím správně předpovězených jevů, celkem modelem předpovíme správně 390 z 450 případů, tj. model je úspěšný v $86,7$ % případů.

Vyhodnocení chyb I. i II. druhu je provedeno v tabulce 5.13. Čistě na základě provedení scoringu bychom úvěr poskytli 437 žadatelům z 450, přičemž u 56 z nich bychom se tak dopustili chyby I. druhu, neboť u klienta byl později indikován default. Default by na základě tohoto scoringu nezískalo 13 žadatelů, přičemž u 9 z nich došlo ke správné predikci defaultu, zatímco u 4 jsme se dopustili chyby II. druhu a ve skutečnosti bonitného klienta scoringem označili za defaultního. Oproti modelu obsahujícímu všechny proměnné se tak zvýšila celková predikční schopnost modelu a snížil se zároveň výskyt chyby II. druhu, kdy pro dvě pozorování, u nichž jsme se původně dopustili chyby II. druhu, byla nyní správně zamítnuta hypotéza H_0 (default klienta).

Odhad—Skutečnost	**1**	**0**	**Počet odhadů**
1 default	9	4	13
0 bonitní klient	56	381	437
Počet skutečných jevů	65	385	450

Tabulka 5.13: Úspěšnost modelu 2 Živnostník - default30

5.3.5 Vyhodnocení nezávislých proměnných a jejich úpravy

V dalším kroku vyhodnocení modelu bychom se měli vrátit k jednotlivým vstupním proměnným a zkontrolovat jejich vztah vůči logitu. S ohledem na to, že vztah logitu a vektoru vstupních proměnných je vyjádřen lineární rovnicí, měl by být i vztah každé jednotlivé nezávislé proměnné a logitu přibližně lineární.

S ohledem na to, že máme k dispozici jen kategoriální data, která už jsou nějakým způsobem ohodnocena, bez možnosti tato data navázat na konkrétní reálné hodnoty proměnných vztahujících se ke každé žádosi o úvěr, můžeme v této fázi jen posoudit, zda rozdělení do kategorií je zvoleno vhodně a navrhnout změny pro celou kategorii, nikoliv vytvořit kategorie a jejich váhy znovu.

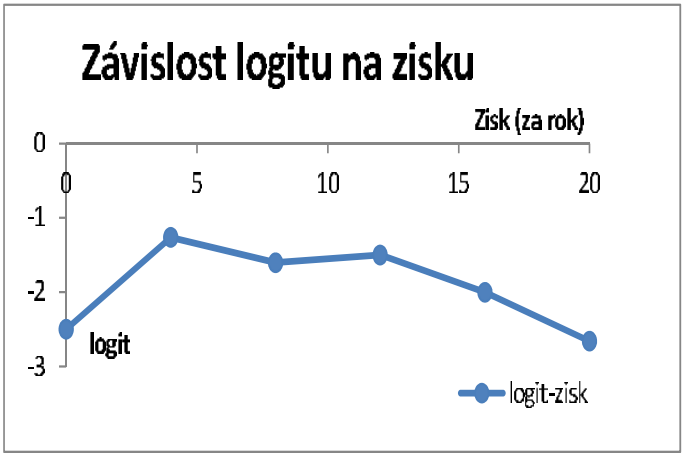
Nejprve se podíváme na četnosti jednotlivých proměnných v jejich vytvořených kategoriích. Proměnné by měly dosahovat přibližně stejného množství v každé kategorii, k vytváření intervalů je používáno rozdělení na kvantily [4]. Poté zkontrolujeme vztah mezi proměnnou a logitem. Celé vyhodnocení jednotlivých proměnných je provedeno v souboru bivariatesENT.xlsx, uloženém ve složce enterpreneur na přiloženém CD. Kon-

tingenční tabulky zobrazující četnosti proměnných v jednotlivých kategoriích jsou vytvořeny i v gretlu skriptem logitENT.inp uloženém tamtéž.

V následujících tabulkách jsou uvedeny četnosti proměnných i hodnota logitu pro příslušnou kategorii, graficky je pak vyobrazeno zobrazení závoslosti logitu na jednotlivých vstupních proměnných.

Zisk		
Bodová hodnota	Četnost	Hodnota logitu
0	188	$-2,502$
4	13	$-1,264$
8	44	$-1,604$
12	39	$-1,500$
16	56	$-2,004$
20	110	$-2,660$

Tabulka 5.14: Vztah zisku a logitu

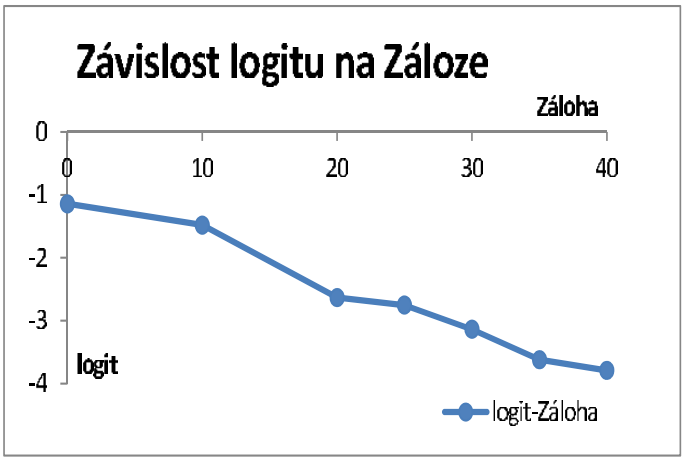


Obrázek 5.1: Vztah mezi logitem a ziskem

Vidíme, že četnosti, kterých dosahuje proměnná Zisk v jednotlivých kategoriích nejsou stejné, bylo by vhodné tak první a poslední kategorii rozdělit na dvě menší. Navíc vztah mezi logitem a ziskem je nelineární pro celý definiční obor kategoriálních vstupních proměnných Zisk. Při odstranění tohoto nedostatku nám jde graficky o to, abychom dostaly veškeré odhadlé hodnoty logitu na jednu přímku vyjadřující závislost logitu na vstupní proměnné Zisk.

Záloha		
Bodová hodnota	Četnost	Hodnota logitu
0	146	$-1,140$
10	42	$-1,479$
20	54	$-2,633$
25	96	$-2,753$
30	49	$-3,137$
35	40	$-3,620$
40	23	$-3,790$

Tabulka 5.15: Vztah zálohy a logitu

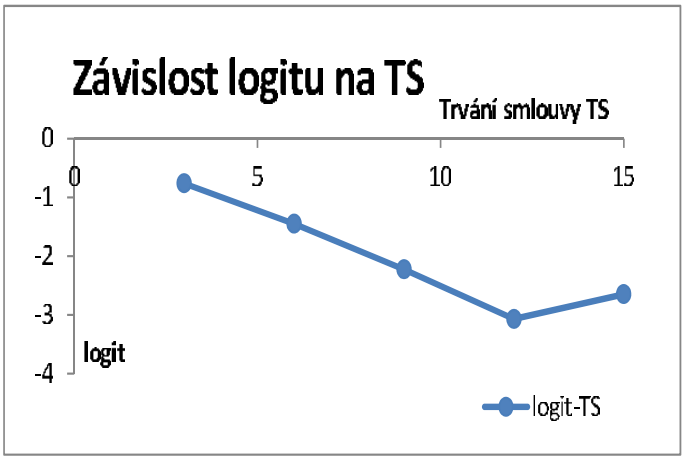


Obrázek 5.2: Vztah mezi logitem a zálohou

U proměnné Záloha naopak vidíme, jak závislost logitu na této vstupní proměnné se řídí takřka lineárně. S narůstající hodnotou kódového ohodnocení Zálohy klesá hodnota logitu a tedy pravděpodobnost, že u daného klienta bude predikován default. S nerovnoměrnými četnostmi u této proměnné nejde provést nápravu, neboť nejvyšší četnost je dosahována u hodnoty 0, která i u skutečných hodnot proměnné Záloha znamená, že klient žádá o úvěr na 100 % ceny předmětu.

Trvání smlouvy		
Bodová hodnota	Četnost	Hodnota logitu
15	23	$-2,650$
12	185	$-3,068$
9	65	$-2,227$
6	164	$-1,448$
3	13	$-0,764$

Tabulka 5.16: Vztah trvání smlouvy a logitu

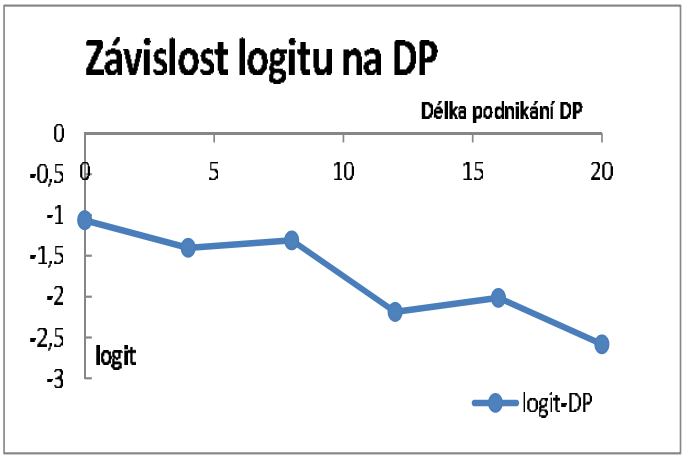


Obrázek 5.3: Vztah mezi logitem a trváním smlouvy

Trvání smlouvy má problém s linearitou jen ve své poslední kategorii. S nerovnoměrnou četností u této proměnné také nejde příliš mnoho udělat. Vysoká četnost pro hodnotu 6, která ve skutečnosti odpovídá délce financování od 4 do 5 let, je dána legislativními daňovými podmínkami finančního leasingu, který jakožto typ účelového financování tvoří významnou množinu smluv poskytovatele úvěru a zejména poté pro segment Živnostník.

Doba podnikání		
Bodová hodnota	Četnost	Hodnota logitu
0	10	$-1,065$
4	15	$-1,404$
8	32	$-1,312$
12	75	$-2,188$
16	68	$-2,014$
20	250	$-2,584$

Tabulka 5.17: Vztah doby podnikání a logitu



Obrázek 5.4: Vztah mezi logitem a dobou podnikání

U proměnné Doba podnikání je možné zcela lineárního trendu dosáhnout jen pospojováním kategorií, které mají nízkou četnost. Vše zobrazeno v tabulce 5.15. Více je možné dohledat na přiloženém CD, ve složce enterpreneur, souboru bivariatesENT.xlsx.

Doba podnikání		
Bodová hodnota	Četnost	Hodnota logitu
2	25	$-1,268$
10	107	$-1,926$
16	68	$-2,014$
20	250	$-2,584$

Tabulka 5.18: Upravení proměnné doba podnikání

S ostatními vstupními daty nelze na úrovni kategoriálních proměnných problém nelinearity vyřešit, bylo by vhodné provést znovu rozdělení do kategorií, případně stanovit frakční polynom a proměnné vhodně transformovat.

Výše uvedený přístup je obecně užíván při vyhodnocování modelu logitové regrese, u modelu scoringu pak byly nadefinovány pro tyto účely navíc koeficienty WOE a IV, které jsou snadno interpretovatelné managementu společnosti. Jejich význam však není možné nikterak statisticky testovat, neboť s nimi není asociován žádný statistický test. Jejich výpočet je pro srovnání proveden na přiloženém CD, ve složce enterpreneur, souboru bivariatesENT.xlsx.

5.3.6 Validace modelu Živnostník

Původní scoring dosahoval úspěšnosti $74,25$ % [41], kdy se podařilo předpovědět z celého souboru 769 pozorování, 571 správně a k chybě v odhadu došlo u 198 pozorování. Bližší specifikace chyb, ke kterým při vyhodnocování vyšlo, je provedena v úvodu kapitoly 5.3.

Model vybudovaný na souboru testovacích dat a zahrnující veškeré proměnné dosahoval úspěšnosti $86,2$ %, kdy správně předpověděl 388 z 450 případů. Oproti výchozí situaci tak došlo k navýšení úspěšnosti predikce o více jak 10 %. Model zahrnující jen významné proměnné pak dosahoval úspěšnosti $86,7$ %, kdy bylo modelem správně předpovězeno 390 z 450 případů a došlo tak opět k navýšení predikční schopnosti modelu. Tento model svoji predikční schopnost navýšil na úkor snížení výskytu chyb II. druhu, kdy pro dvě pozorování v modelu byla nově správně zamítnuta hypotéza H_0, že u klienta nastane default.

Z pohledu vyhodnocení bonity, pak byl úvěr poskytnut celkem 437 žadatelům, z toho 56 defaultním, tedy špatně. Bonita klienta byla správně předpovězena v 381 z 437 případů, tedy s úspěšností $86,2$ %. Za defaultní bylo na základě tohoto modelu považováno 13 klientů, z toho 9 správně a 4krát jsme se dopustili chyby II. druhu.

[41] Na tomto místě bychom měli ještě vzpomenout, že datový soubor, který je předmětem aplikační části této práce, obsahuje jen data o klientech, kteří skutečně úvěr dostali, jako chybu II. druhu pak hodnotíme situaci, kdy úvěr nebyl poskytnut na základě scoringu, ale až na základě rozhodnutí pracovníky CRM. Klienti, kteří byli zamítnuti scoringem a následně i pracovníky CRM nejsou tedy vůbec zahrnuti do předmětu zkoumání.

Tatáž funkce byla ověřena na validačních datech, datový soubor obsahoval celkem 319 pozorování. Validace modelu byla provedena v matlabu, neboť bylo nutné uchovat přesné hodnoty odhadu koeficientů β, jakékoliv zaokrouhlování by u odhadu pravděpodobnosti, která se pohybuje v intervalu $\langle 0, 1 \rangle$ mohlo výrazně ovlivnit dosažené výsledky. Základní skript používaný v matlabu pro validaci dat je uveden v příloze B této práce, elektronicky je navíc uložen na přiloženém CD ve složce Matlab, souboru validaceSCenterpreneur.m.

Na validačních datech pak bylo správně předpovězeno 282 pozorování a 37 pozorování bylo předpovězeno chybně, což představuje $88,40\%$ úspěšnost modelu na validačních datech. Na validačních datech tedy byla ověřena predikční schopnost modelu. Stejné úspěšnosti na validačních datech dosáhl i model obsahující všechny proměnné, které byly v původním modelu.

5.4 Vyhodnocení modelu Soukromá osoba

Všechny podklady k modelu scoringu Soukromá osoba jsou na přiloženém CD ve složce consumer.

Datový soubor obsahuje celkem 158 pozorování. V tomto souboru nastal default90 celkem 5 krát a default30 celkem 14 krát, 48 klientů z toho bylo schváleno až pracovníky CRM, tj. poté, kdy byli skórovací funkcí vyhodnoceni jako nebonitní. Z toho u 4 z nich skutečně došlo k defaultu90 a 11 potom k defaultu30. S ohledem na to, že celý model stavíme na defaultu30, uděláme si zhodnocení právě jen pro default30. Celkové vyhodnocení chyb I. i II. druhu je v tabulce 5.19:

Odhad—Skutečnost	**1**	**0**	**Počet odhadů**
1 default	11	37	48
0 bonitní klient	3	107	110
Počet skutečných jevů	14	144	158

Tabulka 5.19: Úspěšnost modelu Spotřebitel

Celková úspěšnost používaného scoringu dosahuje $74,68$ %, správně bylo předpovězeno 118 ze 158 případů. Chyba II. druhu se zde vyskytuje 37krát, kdy u klienta nebyla

zamítnuta hypotéza H_0 (default klienta), ač ta neplatila. Chyba I. druhu se zde vyskytuje 3krát, kdy byla naopak zamítnuta hypotéza H_0, ač default nastal.

V modelu scoringu pro soukromé osoby (spotřebitele) společnost používá celkem 10 proměnných, bohužel proměnné platební morálka nebyla u tohoto segmentu řádně ukládána a navíc i u něj byla měněna metodika získávání této hodnoty, jelikož není možné zpětně dohledat všechny hodnoty získané stejným procesem, je nutné tuto proměnnou vyloučit rovnou ze souboru. Ve scoringu se tak setkáme s proměnnými s bodovým rozsahem uvedeným v tabulce 5.17. Bližší popis jednotlivých proměnných je proveden v kapitole 4, kde jsou vymezeny jak obecné proměnné vyskytující se ve všech modelech scoringu, tak proměnné specifické pro model scoringu spotřebitele. Veškerá data k této funkci jsou dostupná na přiloženém CD, ve složce consumer, souboru consumers.xlsx.

Proměnná	**Rozsah hodnot**
Délka pracovního poměru DPP	$0-10$
Pracovní zařazení PZ	$0-10$
Věk	$2-15$
Vzdělání VZ	$0-10$
Roční splátky/Roční Příjem RSRP	$0-30$
Záloha Z	$0-40$
Objekt Riziko ObRis	$0-20$
Stáří objektu SO	$4-20$
Trvání smlouvy TS	$3-15$

Tabulka 5.20: Nezávislé proměnné - scoring Spotřebitel

Poté se opět musíme věnovat popisu vztahů mezi proměnnými. Mezi jednotlivými proměnnými byla rozpoznána korelace, která byla ověřena testem multikolinearity. Na základě Farrarova-Glauberova testu definovaného předpisem 3.6 byla na hladině významnosti 5 % zamítnuta hypotéza H_0, že se párové koeficienty významně neodlišují od nuly. Hodnota testovacího kritéria = 116,8 je větší než tabelovaná hodnota 95 % kvantilu χ^2 rozdělení (přibližně 23,3). Síla kolinearity byla změřena prostřednictvím Variance Inflation Factor (VIF), hodnoty této míry pro jednotlivé proměnné jsou uloženy v tabulce 5.21.

Kolinearita	ObR	SO	Zaloha	TS	Vek	VZ	PZ	DPP	RSRP
VIF	$1,08$	$1,19$	$1,37$	$1,15$	$1,06$	$1,13$	$1,07$	$1,17$	$1,45$

Tabulka 5.21: Míra kolinearity proměnných Spotřebitel

Jelikož každá míra kolinearity je opět nižší než 10 [1, 18] nebudeme provádět žádnou transformaci proměnných, jen při hodnocení výsledků nesmíme zapomenout na možný vliv korelací mezi proměnnými.

Nyní můžeme odhadnout parametry β_j. Odhad parametrů je opět proveden v software gretl, kde je stanoven zároveň odhad chyb parametrů, spočten Waldův test a příslušná p-hodnota. Skript pro výpočet, stejně jako výstup skriptu lze nalézt na přiloženém CD, složka consumer. V případě soukromé osoby je proveden odhad parametrů prvního modelu stejně jako odstranění jednotlivých proměnných a odhad parametrů nového modelu postupně v jednom skriptu, soubor LOGITconsumer.inp, výstup tohoto skriptu je poté v textovém dokumentu LOGITconsumer.txt.

Proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$Konstanta$	$7,701$	$3,709$	$2,076$	$0,038$
$ObjektRiziko(ObR)$	$-0,026$	$0,075$	$-0,341$	$0,733$
$StariObjektu(SO)$	$-0,260$	$0,132$	$-1,966$	$0,049$
$Zaloha$ (Z)	$-0,094$	$0,028$	$-3,321$	$0,001$
$TrvaniSmlouvy(TS)$	$-0,247$	$0,153$	$-1,617$	$0,106$
Vek	$-0,042$	$0,120$	$-0,350$	$0,726$
$Vzdelani(VZ)$	$-0,069$	$0,164$	$-0,423$	$0,672$
PZ	$-0,355$	$0,148$	$-2,270$	$0,023$
DPP	$0,045$	$0,120$	$0,375$	$0,707$
$RSRP$	$-0,099$	$0,046$	$-2,131$	$0,033$

Tabulka 5.22: Odhad parametru β modelu 1 Spotřebitel

Po skončení iteračního procesu dosahuje součet logaritmu věrohodnosti hodnoty $L = -21,850$.

Hodnoty Aikeho a Schwarzovo kritérií k posouzení kvality modelu jsou:
$AIC = 63,699$ a $SC = 89,751$.

Hodnoty těchto kritérií jsou pak používány pro porovnání s jinými modely nad stejným souborem vstupních dat, přičemž model je hodnocen jako kvalitnější, pokud dosahuje nižší hodnoty kritéria.

Ze všech proměnných dostaneme tak rovnici scoringu:

$$\begin{aligned} logit[\pi(D30)] \quad &= 7,701 - 0,026 \cdot ObR - 0,260 \cdot SO - 0,094 \cdot Z - 0,247 \cdot TS - \\ &-0,042 \cdot Vek - 0,069 \cdot VZ - 0,335 \cdot PZ + 0,045 \cdot DPP - 0,099 \cdot RSRP \end{aligned}$$

Podmíněné pravděpodobnosti (nastání defaultu) dopočteme dle vzorce:

$$\hat{\pi(D30)} = \frac{1}{1 + e^{-(logit[\pi(D30)])}}$$

Podíváme-li se ovšem na hodnoty odhadlých chyb parametrů β, vidíme, že nabývají významných hodnot a také $p - hodnota$ testů pro jednotlivé proměnné je vyšší než doporučovaná 10 %. Proto z našeho modelu opět odstraníme nevýznamné proměnné, jejich odstranění budeme provádět postupně od proměnných s nejvyšší p-hodnotou. Zjednodušený zápis tohoto postupu je uveden v tabulce 5.23. Podrobněji lze proces odstraňování proměnných nahlédnout v gretlu po spuštění skriptu LOGITconsumer.inp, který je na přiloženém CD ve složce consumer. Ve stejné složce je i výpis z daného skriptu uložený v textovém souboru LOGITconsumer.txt.

Vyhodnocení modelu se všemi proměnnými				
Test poměrem věrohodnosti			AIC	$63,699$
			SC	$89,751$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$ObjektRiziko(ObRisk)$	$-0,026$	$0,075$	$-0,341$	$0,733$
Vyhodnocení modelu po odstranění proměnné ObRisk				
Test poměrem věrohodnosti			AIC	$61,780$
			SC	$85,226$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
Délka pracovního poměru (DPP)	$0,040$	$0,125$	$0,319$	$0,750$
Vyhodnocení modelu po odstranění proměnné DPP				
Test poměrem věrohodnosti			AIC	$59,892$
			SC	$80,734$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
Vzdělání (VZ)	$-0,052$	$0,169$	$-0,306$	$0,759$
Vyhodnocení modelu po odstranění proměnné Vzdělání(VZ)				
Test poměrem věrohodnosti			AIC	$58,013$
			SC	$76,249$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
Věk	$-0,048$	$0,130$	$-0,371$	$0,711$
Vyhodnocení modelu po odstranění proměnné Věk				
Test poměrem věrohodnosti			AIC	$56,153$
			SC	$71,784$
Ostatní proměnné mají p-hodnotu menší než 0,10				

Tabulka 5.23: Postupné odstraňování nevýznamných β_j - Spotřebitel

Z modelu byly postupně odstraněny proměnné (do zvolené p-hodnoty) ObjektRiziko (ObRisk), Délka pracovního poměru (DPP), Vzdělání (VZ), Věk. Model po odstranění nevýznamných proměnných tak obsahuje jen proměnné:

Stáří objektu (SO), Záloha (Z), Trvání smlouvy (TS), Pracovní zařazení (PZ), Roční splátky/Roční Příjem (RSRP).

Rovnici logitové regrese scoringu zahrnující jen významné proměnné pak můžeme pomocí

těchto proměnných a odhadů jejich parametrů zapsat ve tvaru:

$$\begin{aligned} logit[\pi(D30)] \quad &= 6,928 - 0,261 \cdot SO - 0,096 \cdot Z - 0,245 \cdot TS - \\ &-0,305 \cdot PZ - 0,090 \cdot RSRP \end{aligned}$$

Vidíme, že závislost mezi logitem a všemi nezávislými proměnnými je negativní, tedy logit se snižuje s narůstáním jednotlivých vstupních proměnných. Největší vliv na změnu logitu má pak změna v proměnné Pracovní zařazení, u této proměnné také došlo k největší změně v odhadlém parametru oproti původnímu modelu (o 0,03).

Podmíněné pravděpodobnost (nastání defaultu), kterou lze dopočítat dle dále uvedeného vzorce, pak také klesá s rostoucími vstupními proměnnými:

$$\hat{\pi(D30)} = \frac{1}{1 + e^{-(logit[\pi(D30)])}}$$

Také u tohoto modelu můžeme při znalosti kategoriálních proměnných a jejich hodnot na základě výše uvedených rovnic modelu dopočítat maximální logit a jemu příslušející podmíněnou pravděpodobnost defaultu a minimální logit spolu s příslušející pravděpodobností. Logit pro minimální hodnoty vstupních proměnných nabývá hodnoty

$$\begin{aligned} logit[\pi(max)] \quad &= 6,928 - 0,261 \cdot 4 - 0,096 \cdot 0 - 0,245 \cdot 3 - \\ &-0,305 \cdot 0 - 0,090 \cdot 0 = 5,149, \end{aligned}$$

což odpovídá pravděpodobnosti nastání defaultu ve výši:

$$pi(max) = \frac{1}{1 + e^{-5,149}} = 0,994.$$

Pro maximální hodnoty vstupních proměnných pak dostáváme hodnoty:

$$\begin{aligned} logit[\pi(min)] \quad &= 6,928 - 0,261 \cdot 20 - 0,096 \cdot 40 - 0,245 \cdot 15 - \\ &-0,305 \cdot 10 - 0,090 \cdot 30 = -11,557, \end{aligned}$$

což odpovídá pravděpodobnosti nastání defaultu výše:

$$pi(min) = \frac{1}{1 + e^{11,557}} = 0.$$

S ohledem na to, že u tohoto modelu jsou mezní pravděpodobnosti velmi blízké 1 a opačně 0, není zde moc prostoru pro volení cutoff (hranici, od které vyhodnotíme klienta

jako defaultního) nad hranicí pravděpodobnosti pro minimální hodnoty jednotlivých proměnných, tak aby všichni žadatelé o úvěr byli vyhodnoceni jako bonitní, neboť pro žádného z žadatelů nemůže být na základě daných vstupních proměnných predikována vyšší pravděpodobnost defaultu. Opačně zde ani nemá smysl volit hranici cutoff 0, aby byli všichni klienti vyhodnoceni jako defaultní. Tímto krokem se eliminuje chyba I. druhu a model je tak 100% úspěšný z pohledu odhadu defaultních klientů, ale poskytovatel úvěru nikomu ani neposkytne úvěr.

Úspěšnost modelu po odstranění nevýznamných proměnných a vyhodnocení jednotlivých druhů chyb je shrnuta v následující tabulce 5.24:

Odhad—Skutečnost	**1**	**0**	**Počet odhadů**
1 default	2	1	3
0 bonitní klient	7	90	97
Počet skutečných jevů	9	91	100

Tabulka 5.24: Úspěšnost modelu Spotřebitel - default 30 na testovacích datech

Predikční schopnost našeho modelu je tedy 92,0 %, správně bylo předpovězeno 92 ze 100 případů v testovacím souboru dat. S ohledem na šíři vzorku dat však nelze předpokládat přílišnou predikční schopnost na jiných datech. Chyba II. druhu se zde vyskytuje 1krát, kdy u klienta nebyla zamítnuta hypotéza H_0 (default klienta), ač ta neplatila. Chyba I. druhu se zde vyskytuje 7krát, kdy byla naopak zamítnuta hypotéza H_0, ač platila a default nastal. Stejné úspěšnosti i rozložení chyb prvního a druhého druhu pak dosahoval model i se zahrnutými všemi proměnnými.

Zatímco v původně používaném modelu docházelo častěji k chybě II. druhu (37krát) oproti chybě I. druhu (3krát). V modelu navrhovaném touto funkcí častěji dochází k chybám I. druhu(7krát) oproti chybě II. druhu (1krát). Cutoff našeho modelu je tak nastaven mnohem mírněji pro klienty a do našeho portfolia se dostanou i defaultní klienti. Model užívaný původně byl nastaven přísně a do portfolia společnosti pouštěl méně defaultních klientů, ovšem také na úkor klientů, kteří se projevili jako bonitní (chyba II. druhu).

Navržený model tak otestujeme ještě na validačních datech. Pro validaci tohoto modelu nám zbyl již jen soubor o rozsahu 58 pozorování. Validaci modelu provedeme v matlabu,

skript pro validaci je uložen na přiloženém CD, složka Matlab, soubor consumersSCvalidace.m. Původní model na validačních datech dosahuje úspěšnosti 93,1 %, kdy je správně předpovězeno 54 z 58 případů, stejné úspěšnosti na validačních datech dosahuje i model po odstranění nevýznamných proměnných.

5.5 Vyhodnocení modelu Obchodní společnosti

Všechny podklady k modelu scoringu Obchodní společnosti jsou na přiloženém CD ve složce company.

Datový soubor obsahuje celkem 433 pozorování. V této skupině zákazníků nastal default90 celkem 12krát a default30 celkem 31krát, 186 klientů z toho bylo schváleno až pracovníky CRM, tj. poté, kdy byli skórovací funkcí vyhodnoceni jako nebonitní. Z toho jen u 8 z nich skutečně došlo k defaultu90 a 21 potom k defaultu30. S ohledem na to, že celý model stavíme na defaultu30, uděláme si zhodnocení právě jen pro default30. Celkové vyhodnocení chyb I. i II. druhu je v tabulkách níže:

Odhad—Skutečnost	**1**	**0**	**Počet odhadů**
1 default	21	165	186
0 bonitní klient	10	237	247
Počet skutečných jevů	31	402	433

Tabulka 5.25: Úspěšnost modelu Společnost

Úspěšnost používaného scoringu je tedy $59,58$ %, kdy správně bylo předpovězeno 258 ze 433 případů. Chyba II. druhu se zde vyskytuje 165krát, kdy u klienta nebyla zamítnuta hypotéza H_0 (default klienta), ač ta neplatila. Chyba I. druhu se zde vyskytuje 10krát, kdy byla naopak zamítnuta hypotéza H_0, ač platila a default nastal.

V modelu scoringu pro obchodní společnosti finanční instituce používá celkem 9 nezávisle proměnných, bohužel ani u tohoto segmentu nebyla data k proměnné platební morálka řádně ukládána, přičemž stejně platí, že zde byla měněna metodika získávání této hodnoty a hodnoty získané stejným procesem není možné nikde zpětně dohledat ani dopočítat, proto je nutné tuto proměnnou vyloučit rovnou ze souboru. Vyhodnocovat tak budeme

model scoringu s proměnnými, jimž společnost stanovila bodový rozsah uvedený v tabulce 5.21. Veškerá data k této funkci jsou dostupná na přiloženém CD, ve složce company, souboru companies.xlsx.

Proměnná	**Bodový rozsah**
Vlastní kapitál/Bilanční suma VKB	$0-15$
Obrat	$0-10$
EcoRatio	$0-20$
Stáří společnosti DP	$0-20$
Záloha	$0-40$
Trvání smlouvy TS	$3-15$
ObjektRisiko ObR	$0-20$
Stáří objektu SO	$4-20$

Tabulka 5.26: Nezávislé proměnné - scoring Obchodní společnost

Poté se opět musíme věnovat popisu vztahů mezi proměnnými. Mezi jednotlivými proměnnými byla rozpoznána korelace, která byla ověřena testem multikolinearity. Na základě Farrarova-Glauberova testu definovaného předpisem 3.6 byla na hladině významnosti 5 % zamítnuta hypotéza H_0, že se párové koeficienty významně neodlišují od nuly. Hodnota testovacího kritéria = 924,823 je větší než tabelovaná hodnota 95 $\%kvantilu$ χ^2 rozdělení (přibližně 16,7). Síla kolinearity byla změřena prostřednictvím Variance Inflation Factor:

Kolinearita	ObR	SO	Zaloha	TS	DP	Obrat	VKB	EcoRat
VIF	$1,28$	$1,23$	$1,47$	$1,14$	$1,10$	$1,70$	$2,77$	$3,34$

Tabulka 5.27: Míra kolinearity proměnných Obchodní společnost

Jelikož míra kolinearity je opět nižší než 10 nebudeme provádět žádnou transformaci proměnných, ale ani zde bychom při hodnocení výsledků neměli zapomínat na možný vliv korelací mezi proměnnými a jejich lineárními kombinacemi.

Nyní můžeme odhadnout parametry β_j jednotlivých proměnných funkce scoringu obsahující veškeré vstupní proměnné. Odhad parametrů je opět proveden v software gretl, kde je stanoven zároveň odhad chyb parametrů (SE), spočten Waldův test a p-hodnota

příslušející odhadům jednotlivých parametrů. Skript pro výpočet soubor LOGITcompany.inp, stejně jako výstup skriptu v textovém formátu LOGITcompany.txt lze nalézt na přiloženém CD, složka company. Ze všech našich proměnných dostaneme tak rovnici scoringu:

$$\begin{aligned} logit[\pi(D30)] \quad &= -1,232 - 0,014 \cdot ObR - 0,019 \cdot SO - \ 0,037 \cdot Z - \\ &-0,229 \cdot TS - 0,047 \cdot DP + 1,380 \cdot Obrat + \ 0,173 \cdot VKB - \\ &-0,143 \cdot EcoRat \end{aligned}$$

Podmíněné pravděpodobnosti (nastání defaultu) dopočteme dle vzorce:

$$\hat{\pi(D30)} = \frac{1}{1 + e^{-(logit[\pi(x)])}}$$

Hodnoty kritérií k posouzení kvality tohoto modelu jsou:
$L = -51,569 \qquad AIC = 121,138 \qquad SC = 153,656$

Také v tomto modelu je $p-hodnota$ testů pro jednotlivé proměnné několikrát vyšší než doporučovaných 10 %, odhady chyb jednotlivých proměnných jsou navíc často výrazně vyšší než samotný odhadlý parametr. Proto z našeho modelu opět odstraníme nevýznamné proměnné, jejich odstranění opět budeme provádět postupně od proměnných s nejvyšší p-hodnotou.

Vyhodnocení modelu se všemi proměnnými				
Test poměrem věrohodnosti			AIC	$121,138$
			SC	$153,656$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
$ObjektRiziko(ObRisk)$	$-0,014$	$0,079$	$-0,178$	$0,859$
Vyhodnocení modelu po odstranění proměnné ObRisk				
Test poměrem věrohodnosti			AIC	$119,160$
			SC	$148,065$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
Stáří objektu(SO)	$-0,024$	$0,068$	$-0,353$	$0,724$
Vyhodnocení modelu po odstranění proměnné SO				
Test poměrem věrohodnosti			AIC	$117,293$
			SC	$142,584$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
Stáří společnosti (DP)	$-0,044$	$0,038$	$-1,152$	$0,249$
Vyhodnocení modelu po odstranění proměnné DP				
Test poměrem věrohodnosti			AIC	$116,721$
			SC	$138,340$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
Záloha Z	$-0,033$	$0,025$	$-1,330$	$0,184$
Vyhodnocení modelu po odstranění proměnné Záloha				
Test poměrem věrohodnosti			AIC	$116,439$
			SC	$134,504$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
VK Bilance VKB	$0,149$	$0,097$	$1,526$	$0,127$
Vyhodnocení modelu po odstranění proměnné VKB				
Test poměrem věrohodnosti			AIC	$116,302$
			SC	$130,754$
Odstraněná proměnná	$Odhad\beta_j$	$Chybaodhadu$	$Wald$	$p-hodnota$
EcoRatio	$-0,031$	$0,045$	$-0,707$	$0,480$
Vyhodnocení modelu po odstranění proměnné EcoRatio				
Test poměrem věrohodnosti			AIC	$114,765$
			SC	$125,605$
Ostatní proměnné mají p-hodnotu menší než 0,10				

Tabulka 5.28: Postupné odstraňování nevýznamných β_j - Společnost

Námi vytvořená funkce scoringu by tak vedle proměnné obsahovala jen dvě nezávisle proměnné a byl by pro ní platný následující předpis funkce:

$$logit[\pi(D30)] \quad = -2,652 - 0,285 \cdot TS + 1,446 \cdot Obrat$$

U tohoto modelu je tedy nutno se také zamyslet nad ekonomickým významem daného modelu a přemýšlet, zda není vhodné zachovat v modelu více proměnných, ač predikční schopnost námi vytvořeného modelu je stejná jako modelu se všemi vstupními proměnnými. Z předpisu funkce navíc vidíme ekonomicky nelogickou pozitivní závislost výše obratu a logitu, tj. dle námi vytvořené funkce je více pravděpodobné, že nastane default u subjektu, který má vyšší obrat.

Po projití celého výpočtu eliminace proměnných stanovíme jako $p - hodnoty$ testů vylučovaných proměnných 20 % , v modelu tak zůstanou vedle proměnné Trvání Smlouvy a Obrat zachovány i proměnné Záloha, VK Bilance a EcoRatio a předpis modelu logitové regrese vypadá následovně:

$$\begin{aligned} logit[\pi(D30)] \quad =& -1,837 - 0,033 \cdot Z - 0,258 \cdot TS - \\ & -0,142 \cdot EcoRatio + 1,380 \cdot Obrat + 0,166 \cdot VKB \end{aligned}$$

Vidíme opět pozitivní závislost mezi defaultem a obratem společnosti, což se nejeví příliš ekonomicky logické. Nelogický odhad defaultu společnosti ve vztahu k jeho obratu byl způsoben defaultem společností s vysokým obratem, které měly ve zkoumaném portfoliu více smluv, kde tak nastal default. S ohledem na šíři našeho vzorku a poměru defaultu v něm, ovlivní krach významného klienta celý model.

Kvůli nestejnému trendu jednotlivých proměnných je nejednoznačný i výpočet maximální pravděpodobnosti nastání defaultu pro nejnižší hodnotu logitu. Možnosti různých kombinací proměnných s negaivním trendem a proměnných s pozitivním trendem limitují i sledování závislosti pravděpodobnosti na hodnotách vstupních proměnných.

Díky vlivu jednoho subjektu na výsledky našeho modelu, nelze tento považovat za model s predikční schopností, ač vyhodnotíme-li úspěšnost modelu na testovacích datech po odstranění nevýznamných proměnných (na hranici $p - hodnoty$ 20 %), dostáváme se na hodnotu 93,4 % odhadnutých chování, shrnutí je v následující tabulce 5.29:

Odhad—Skutečnost	**1**	**0**	**Počet odhadů**
1 default	2	0	2
0 bonitní klient	18	254	272
Počet skutečných jevů	20	254	274

Tabulka 5.29: Úspěšnost modelu Společnost - na testovacích datech

Predikční schopnost našeho modelu je tedy 93,4 %, správně bylo předpovězeno 256 z 274 případů v souboru dat. Chyba II. druhu se zde vyskytuje 0krát, Všichni klienti zamítnuti scoringem, tak byli skutečně defaultní, scoring nezamítl nikoho, u koho poté default skutečně nenastal, tj. u klienta nebyla zamítnuta hypotéza H_0 (default klienta), ač ta neplatila. Chyba I. druhu se zde vyskytuje 18krát, kdy byla naopak zamítnuta hypotéza H_0, ač platila a default nastal.

Zatímco v původně používaném modelu docházelo častěji k chybě II. druhu (165krát) oproti chybě I. druhu (10krát). V modelu navrhovaném touto funkcí byla chyba II. druhu zcela eliminována. Cutoff našeho modelu je s ohledem na vstupní proměnné nastaven mírně pro klienty a do portfolia se dostanou i defaultní klienti. Z důvodu nízké defaultnosti ve sledovaném souboru dat, tak scoring, který úvěr poskytne téměř každému, dosahuje vysoké celkové predikční úspěšnosti. Model užívaný původně byl nastaven přísně a do portfolia společnosti pouštěl méně defaultních klientů, ovšem také na úkor klientů, kteří se projevili jako bonitní (chyba II. druhu).

Pro validaci modelu Společnosti pak využijeme opět software Matlab, kde spustíme skript companiesSCvalidace uložený na přiloženém CD, ve složce Matlab. Validační soubor dat dosahuje rozsahu 159 pozorování a jsou na něm otestovány všechny tři modely využívané k popisu scoringu Společnosti v této kapitole.

Původní model se všemi vstupními proměnnými na validačních datech dosahuje úspěšnosti 93,1 %, kdy je správně předpovězeno 148 ze 159 případů, stejné úspěšnosti na validačních datech dosahuje i model, u něhož jsme pro odstranění nevýznamných proměnných použili hraniční $p - hodnotu$ 20 %. Model, kde jsme proměnné odstraňovali jako nevýznamné až do doby, dokud se v něm nevyskytovaly jen takové, jejichž $p-hodnota$ byla nižší než 10 % pak dosahuje nepatrně nižší úspěšnosti, kdy dokáže předpovědět 147 ze 159 případů správně, tedy dosahuje úspěšnosti 92,45 %.

5.6 Shrnutí aplikační části

Na tomto místě bychom měli zhodnotit jednotlivé scoringové funkce, které byly v této kapitole zpracovány a navrhnout případně další postup vedoucí ke zlepšení jejich predikční schopnosti.

Scoring segmentu Živnostník dosahoval původní úspěšnosti 74,25 %. Při hodnocení tohoto segmentu jsme původně používali funkci obsahující 8 proměnných, z toho 5 proměnných popisujících vlastnosti poptávaného produktu a 3 popisující vlastnosti žádajícího klienta. Funkce scoringu, kterou jsme nad těmito daty vytvořili prostřednictvím logitové regrese, pak vypadala následovně:

$$\begin{aligned} logit[\pi(x)] \quad &= 3,083 + 0,025 \cdot ObR - 0,042 \cdot SO - 0,083 \cdot Z - \\ &\quad -0,195 \cdot TS - 0,075 \cdot DP - 0,029 \cdot Obrat + 0,053 \cdot Zisk - 0,002 \cdot PM \end{aligned}$$

Vidíme, že v tomto modelu všechny proměnné nenabývají stejného trendu, proměnná ObjektRisiko a proměnná Zisk má stejný trend se závisle proměnnou logit. U proměnné ObjektRisiko, která je vytvořena uměle, může být tento stav logicky odůvodnitelný, u proměnné Zisk, jejíž kódové hodnocení roste s rostoucím ročním Ziskem živnostníka již tento výsledek ovšem logický příliš není, znamená totiž, že čím vyššího Zisku Živnostník dosahuje, tím spíše u něj nastane default a je tak považován za méně úvěruschopného.

Některé proměnné však byly na základě testu významnosti jejich parametrů (testujících hypotézu H_0, že parametr je nulový proti alternativě, že hodnota parametru je různá od nuly) označeny za nevýznamné a z modelu odstraněny, parametry byly odstraňovány, dokud jejich p-hodnota byla vyšší jak $0,1$. Dostali jsme tak nový model scoringu Živnostník, který již obsahoval jen 4 nezávislé vstupní proměnné a jehož rovnice měla tvar:

$$logit[\pi(x)] \quad = 2,715 - 0,074 \cdot Z - 0,196 \cdot TS - 0,084 \cdot DP - 0,061 \cdot Zisk$$

Vidíme, že v tomto modelu scoringu již mají veškeré vstupní proměnné stejný trend, závislost logitu na nich je negativní, čím vyšší hodnoty dosahují jednotlivé proměnné, tím nižší hodnoty dosahuje logit a tedy také tím menší je odhadovaná pravděpodobnost defaultu klienta. V tomto modelu byly zachovány dvě proměnné popisující vlastnost produktu - záloha a délka trvání smlouvy a dvě proměnné hovořící o vlastnostech žadatele klienta - zisk a doba podnikání. V praxi je tento model scoringu neakceptovatelný právě

pro nízký počet proměnných, které obsahuje, zejména poté o vlastnostech klienta, poskytovatel úvěru si nedokáže představit model scoringu bez využití proměnné Obrat.

Navrhovaný model scoringu pro zvolený cutoff dosahoval úspěšnosti 86,7 %, na validačních datech pak úspěšnosti 88,4 %.

Scoring segmentu Spotřebitel dosahoval původní úspěšnosti 74,68 %. Nutno však připomenout, že datový soubor pro celý segment obsahoval jen 158 pozorování. Původně scoring obsahující 9 proměnných, z toho 5 proměnných popisujících vlastnosti žadatele a 4 popisující vlastnosti nabízeného produktu. Funkce scoringu pro tyto proměnné byla následující:

$$\begin{aligned} logit[\pi(D30)] \quad &= 7,701 - 0,026 \cdot ObR - 0,260 \cdot SO - 0,094 \cdot Z - 0,247 \cdot TS - \\ &-0,042 \cdot Vek - 0,069 \cdot VZ - 0,335 \cdot PZ + 0,045 \cdot DPP - 0,099 \cdot RSRP \end{aligned}$$

Vidíme, že v tomto modelu má jediná proměnná Délka pracovního poměru stejný trend se závisle proměnnou logit. U této proměnné tento vývoj taktéž není logicky očekávaný.

I zde byly proměnné na základě testu nevýznamnosti jejich parametrů označeny za nevýznamné a z modelu odstraněny, parametry byly odstraňovány, dokud jejich p-hodnota byla vyšší jak $0,1$. Nový model scoringu Spotřebitele obsahoval jen 5 nezávislých vstupních proměnných, z nichž dvě hodnotí vlastnosti klienta a tři poté popisují vlastnosti finančního produktu, rovnice tohoto modelu je pak dána předpisem:

$$\begin{aligned} logit[\pi(D30)] \quad &= 6,928 - 0,261 \cdot SO - 0,096 \cdot Z - 0,245 \cdot TS - \\ &-0,305 \cdot PZ - 0,090 \cdot RSRP \end{aligned}$$

Vidíme, že v tomto modelu scoringu již mají veškeré vstupní proměnné stejný trend, závislost logitu na nich je negativní, čím vyšší hodnoty dosahují jednotlivé proměnné, tím nižší hodnoty dosahuje logit a tedy také tím menší je odhadovaná pravděpodobnost defaultu klienta. V praxi by i pro tento model scoringu bylo vyžadované, aby obsahoval větší počet proměnných, zejména poté o vlastnostech klienta.

Navrhovaný model scoringu pro zvolený cutoff dosahoval úspěšnosti 92,0 %, na validačních datech pak úspěšnosti 93,1 %.

Scoring segmentu Společnost dosahoval původní úspěšnosti 59,58 %. Ani datový soubor pro tento segment však nebyl příliš rozsáhlý, další nevýhodou tohoto segmentu byl fakt, že se v něm častěji vyskytovali zákazníci, kteří měli více než jednu smlouvu, jejich

chování (default), pak výrazně ovlivňovalo defaultnost pro celý segment. Původně scoring obsahující 8 proměnných, z toho 4 popisují vlastnosti žadatele a 4 vlastnosti nabízeného produktu. Funkce scoringu pro tyto proměnné byla následující:

$$\begin{aligned} logit[\pi(D30)] \quad &= -1,232 - 0,014 \cdot ObR - 0,019 \cdot SO - \ 0,037 \cdot Z - \\ &-0,229 \cdot TS - 0,047 \cdot DP + 1,380 \cdot Obrat + \ 0,173 \cdot VKB - \\ &-0,143 \cdot EcoRat \end{aligned}$$

Vidíme, že ani v tomto modelu nemají veškeré proměnné stejný trend (tato vlastnost není obecně u modelu logitové regrese vyžadována, ale u kódovaných hodnot proměnných v modelu scoringu se tato vlastnost předpokládá). Opačný než předpokládaný trend se vyskytuje u proměnné Obrat a Vlastní kapitál/Bilance, bilancí je zde myšlena bilanční suma dané společnosti za předchozí účetní období.

I zde byly některé proměnné označeny za nevýznamné a z modelu odstraněny, čímž mohl být postupně změněn i trend jednotlivých proměnných. Parametry byly v tomto případě odstraňovány, dokud jejich p-hodnota byla vyšší jak $0,2$. Při využití standardní výše p-hodnoty $0,1$ byl dosažen model, který obsahoval jen dvě vstupní proměnné. Užívání takového modelu muselo být logicky zamítnuto a odstranění proměnných bylo následně provedeno znovu jen do úrovně p-hodnoty $0,2$. Nový model scoringu Společnost obsahoval jen 5 nezávislých vstupních proměnných, z nichž dvě hodnotí vlastnosti klienta a tři poté popisují vlastnosti finančního produktu, rovnice tohoto modelu je pak dána předpisem:

$$\begin{aligned} logit[\pi(D30)] \quad &= -1,837 - 0,033 \cdot Z - 0,258 \cdot TS - \\ &-0,142 \cdot EcoRatio + 1,380 \cdot Obrat + 0,166 \cdot VKB \end{aligned}$$

Vidíme, že v tomto modelu scoringu se opět vyskytují dvě proměnné, které mají nelogický trend vůči logitu. Logit je pozitivně závislý na proměnných Obrat a Vlastní kapitál/Bilanční suma. Tento stav je nelogický, neboť říká, že čím vyšších hodnot dané proměnné nabývají, tím vyšších hodnot nabývá také logit a na základě něj predikovaná pravděpodobnost defaultu. Tento vztah mezi ekonomickými ukazateli společnosti a defaultem byl způsoben výskytem většího klienta v daném vzorku, jehož default ovlivnil celý segment.

Z výše uvedených logických důvodů, tak navrhovaný model scoringu není vhodné aplikovat, ač pro zvolený cutoff dosahoval úspěšnosti 93,4 %, na validačních datech pak úspěšnosti 93,1 %.

Porovnávali-li bychom jednotlivé modely jen podle dosažené predikční úspěšnosti, vyhodnotili bychom jako nejlepší model scoringu navrhovaný pro Společnosti, neboť dosahuje úspěšnosti větší než 93 %. Tento výsledek však nemůžeme považovat za dostačně vypovídající. V tomto segmentu bylo ve vstupních datech obsaženo nejmenší proporcionální množství defaultu, 31 defaultů ze 433 pozorování představující jen 7,15% výskyt defaultu ve sledovaném souboru. Určili-li bychom, že každý klient bude bonitní, dosahujeme úspěšnosti v predikci také téměř 93 %. Důležité je tedy také sledovat, kolik defaultů bylo na základě našeho modelu predikováno, zde dosahujeme na testovacích datech hodnoty 2 z 20, tedy je správně předpovězeno jen 10 % defaultních klientů.

U scoringu Spotřebitele, kde hodnota scoringu taktéž dosahuje vysoké úspěšnosti, je poté jen necelých 10 % defaultních klientů ve sledovaném souboru dat, tento soubor dat navíc nedosahuje ani patřičného rozsahu. Správně předpovězené tu byly dva defaulty z 9, což vypovídá o vyšším procentu správně předpovězených defaultů než u modelu Společnosti, přesto ani u tohoto modelu nelze předpokládat silné předpovídací schopnosti do budoucna, neboť je celý odhadnut na základě úzkého základního souboru.

Scoring Živnostníka pak obsahuje vyšší výskyt defaultu klienta oproti dříve uvedeným modelům. Celkem se v 769 pozorováních vyskytl default 100krát, v testovacích datech pro vytvoření modelu pak bylo 65 defaultů na 450 pozorování, 14,44% výskyt defaultu ve sledovaných datech. Námi navrhovaný model scoringu pak dosahuje úspěšnosti při stanovené hladině cutoff 0,5 86,7 %. Úspěšnost modelu i v tomto případě výrazně zvyšuje jeho schopnost předpovídat bonitního klienta, oproti jeho schopnosti předpovídat klienta defaultního, ten byl na základě modelu správně predikován jen v 9 případech z 65, 4krát byl naopak default predikován aniž by nastal, tedy jen 4krát došlo k chybě II. druhu a 56krát došlo k chybě I. druhu.

Malý poměr výskytu defaultu ovšem způsobuje, že ani změnou cutoff nedosáhneme výrazně lepších výsledků. Při stanovené hranici cutoff na 0,45 dosahuje model úspěšnosti 87, 56 %, způsobené právě větší úspěšností v predikci defaultu. Snižujeme-li cutoff více (o dalších 0,05) lepších výsledků již nedosáhneme, neboť nám výrazněji vzroste chyba I. druhu než kolik je správně předpovězeno defaultů. Zvyšujeme-li cutoff k 1, od hodnoty 0,7 pozorujeme, že už se celková úspěšnost modelu vůbec nemění, 65krát nastane chyba I. druhu, to znamená, že ani jeden default z testovacího souboru není předpovězen správně a všichni klienti jsou vyhodnoceni jako bonitní. Testování nastavení různé hodnoty cutoff je provedeno v matlabu, skripty s jednotlivými hraničními cutoff jsou uloženy

na přiloženém CD, složka Matlab, jednotlivé soubory pak mají názvy enterprSC01.m až enterprSC09.m dle stanovené hodnoty cutoff.

Vidíme, že ve všech modelech scoringu byly zachovány proměnné Záloha a Trvání smlouvy, tyto ve všech také obecně splňují negativní trend vůči logitu. Tyto dvě proměnné vypovídající o vlastnostech produktu se tak jeví jako proměnné s nejlepší vypovídací hodnotou o defaultnosti klienta bez rozlišení do jakého segmentu ho zařadíme. Podíváme-li se blížeji na tyto proměnné vidíme, že jejich hodnoty nevypovídají jen o produktu, ale skrytě i o vlastnostech klienta, u nichž opravdu dle praxe na segmentu nezáleží. Klient platící vysokou zálohu a tedy mající k dispozici peněžní prostředky je vyhodnocován jako bonitní a stejně tak klient, který si půjčuje jen na krátkou dobu je dle těchto funkcí scoringu hodnocen jako ten, u něhož je default predikován s menší pravděpodobností. U účelového financování se v krátkodobých závazcích mohou skrývat i krátkodobé pronájmy, u nichž subjekt pokrývá jen své krátkodobé potřeby a své závazky, i díky zajištění tohoto typu financování, v praxi plní téměř vždy.

Kvůli nedostupným datům jsme bohužel nemohli navrhnout model pracující i s jinými proměnnými než které byly již ve scoringu využívány. Proto se na tomto místě zamyslíme nad teoretickými možnostmi rozšíření modelů scoringu z obecně dostupných dat.

Nejvíce lze v tomto smyslu rozvinout asi model Spotřebitele, ve kterém lze užívat různé údaje o žadateli dostupné z veřejně přístupných zdrojů. Velmi žádané pro tuto oblast by byl přístup do centrální databáze bank, kde jsou ukládány údaje o předchozí platební morálce žadatele i z jiných závazků. Model spotřebitele u jiných poskytovatelů úvěru zpravidla pracuje s proměnnou další závazky spotřebitele, informace o tomto ukazateli jsou dostupné taktéž v centrální databázi bank. Náš model obsahuje jen poměr mezi splátkami a příjmem žadatele, aniž by zohlednil žadatelovy jiné závazky. Dalším vyhodnocovaným kritériem bývá počet členů domácnosti či vyživovaných osob.

U modelu scoringu Společnosti lze pak doporučit více pracovat s ekonomickými ukazateli vyskytujícími se v účetních výkazech obchodních společností a vytvořit z nich nějaké poměrové či jinak trasformované proměnné (mezi ekonomickými ukazateli se dá předpokládat vysoká závislost, proto je nutné tyto před zahrnutím do modelu scoringu nějak upravit).

Pro model scoringu Živnostník pak lze doporučit přístup někde na pomezí Společnosti a Spotřebitele, více se klanící spíše ke Společnosti, a tedy i u Živnostníka se zaměřit

primárně na jeho ekonomické ukazatele, u tohoto segmentu se jeví také jako vhodné sledovat zaměření činnosti Živnostníka a jeho vztah k zamýšlenému financování, či se zajímat o podnikatelské plány, zakázky, renome apod. tohoto segmentu.

Kapitola 6

Závěr

V diplomové práci byl popsán způsob hodnocení bonity klienta a posouzení jeho úvěruschopnosti z pohledu poskytovatele úvěru, při zpracování této práce tak musel být v obecné rovině kladen důraz na to, jaká data jsou poskytovateli úvěru o žadateli úvěru známá.

Po shrnutí teoretických předpokladů k vyhodnocování úvěrového rizika poskytovatele úvěru zejména se zaměřením na scoring byly vymezeny statistické nástroje, které se poskytovateli úvěru nabízí k aktivnímu řízení jeho úvěrového rizika metodou scoringu. Hlavní pozornost byla poté věnována logitové regresi, tato metoda byla pro lepší představivost porovnávána s lineární regresí.

V aplikační části práce byl poté vyhodnocen model scoringu používaný aktivně na trhu. Celkem byly zhodnoceny tři modely scoringu využívané pro různé segmenty klientů - živnostníka, soukromou osobu a společnost. Hlavní pozornost byla poté upřena na model scoringu Živnostník.

Před samotnou aplikací logitové regrese je nutné nejprve zanalyzovat jednotlivé vstupní proměnné, tento úkol byl v této práci limitován poskytnutím jen kódovaných (bodových) hodnot jednotlivých vstupních proměnných. Veškerá vstupní data a vztahy mezi nimi byla tak posuzována jen na základě kódovaných hodnot. Pro kódované proměnné byla posouzena i nezávislost jednotlivých vstupních proměnných jakožto předpoklad užití logitové regrese. Vliv kódovaných hodnot a jejich vztahu vůči predikovanému defaultu klienta byl vyhodnocen pro scoring Živnostník.

Při zpracování modelu scoringu byly odhadnuty parametry jednotlivých vstupních proměnných včetně odhadnutí chyb těchto odhadů (SE). Následně byla otestována významnost jednotlivých parametrů a proměnné, jejichž parametry byly označeny nevýznamnými, byly odstraněny z modelu. Pro takto nově vytvořený model byly parametry odhadnuty znovu a na závěr byla vyhodnocena kvalita jednotlivých modelů prostřednictvím kritérií vybudovaných na věrohodnostní funkci. Scoring Živnostník byl poté ještě jednou vytvořen pomocí Stepwise analýzy, přičemž za potencionální proměnné byly považovány veškeré proměnné využívané ve scoringu společnosti.

Navrhované modely scoringu dosahovaly vysoké predikční úspěšnosti, kolem 90 %. Je nutné si uvědomit, že toto bylo způsobeno nízkým výskytem defaultu (okolo 10 %) v základním souboru dat, proto i model, který by vyhodnotil všechny klienty jako bonitní by dosahoval takto vysoké úspěšnosti, i když jím dosahované vyýsledky by v praxi nebyly žádoucí, neboť portfolio poskytovatele úvěru by se plnilo defaultními klienty. U modelů je proto nutné sledovat výskyt jednotlivých druhů chyb a korigovat jejich výskyt v závislosti na účelu použití scoringu (opačná tendence chceme-li nárůst portfolia oproti situaci, kdy primárně chceme snížit defaultnost portfolia).

Seznam obrázků

Seznam tabulek

Literatura

[1]Reif, J.. *Metody matematické statistiky.* 1. vyd. Plzeň: Západočeská univerzita, Fakulta aplikovaných věd, 2004. 264 s. ISBN 04-347-76.

[2]Hušek, R. *Ekonometrická analýza.* 1. vyd. Praha: OECONOMICA, 2007. 368 s. ISBN 978-80-245-1300-3.

[3]Hušek, R. *Aplikovaná ekonometrie: teorie a praxe.* 1. vyd. Praha: OECONOMICA, 2009. 346 s. ISBN 978-80-245-1623-3.

[4]Hosmer, D., W., Lemeshow, S. *Applied Logistic Regression.* 2. vyd. New York: John Wiley a Sons, 2000. 373 s. ISBN 0-471-35632-8.

[5]Anděl, J. *Základy matematické statistiky.* 1. vyd. Praha: MATFYZPRESS, 2005. 360 s. ISBN 80-86732-40-1.

[6]Anděl, J. *Statistické metody.* 3. vyd. Praha: MATFYZPRESS, 2003. 300 s. ISBN 80-86732-08-8.

[7]Rao, R., C. *Lineární metody statistické indukce a jejich aplikace.* 1. vyd. Praha: Academia, 1978. 668 s. ISBN 104-21-852.

[8]Hardin, J., Hilbe, J. *Generalized linear models and extensions.* 1. vyd. College Station: Stata Press, 2001. 246 s. ISBN 1-881228-60-6.

[9]Jílek, J. *Finanční rizika.* 1. vyd. Praha: Grada, 2000. 635 s. ISBN 80-7169-579-3.

[10]Vinš, P., Liška, V. *Rating.* 1. vyd. Praha: C. H. Beck, 2005. 109 s. ISBN 80-7179-807-X.

[11]Hlinica, J. *Aplikovaná analýza rizika ve finančním managementu a investičním*

rozhodování. 1. vyd. Praha: Grada Publishing, a.s., 2009. 309 s. ISBN 80-7179-817-X.

[12]Česká spořitelna, a. s. *Basel II final cj.pdf* [online]. 2004, [citováno dne 8. dubna 2011]. Dostupné z: <http://www.csas.cz>.

[13]Basel Committee on Banking Supervision. *Basel II framework* [online]. 2000, [citováno dne 19. března 2011]. Dostupné z: ¡http://www.bis.org¿.

[14]Basel Committee on Banking Supervision. *Basel II implementation* [online]. 2000, [citováno dne 20. března 2011]. Dostupné z: ¡http://www.bis.org¿.

[15]de Andrade, F., W., M, Lyn, T. Structural models in consumer credit. *Science-Direct: European Journal of Operational Research,* February 2007, 183, s. 1569 - 1581

[16]Anderson, R. *Credit Scoring Toolkit - Theory and Practise for Retail Credit Risk Management and Decision Automation.* 1. vyd. New York: Oxford University Press Inc., 2007. 731 s. ISBN ???.

[17]Šedivá, B. *Přednášky z předmětu KMA/MME* [online]. 2008, [citováno dne 20. května 2011]. Dostupné z: ¡http://www.kma.zcu.cz¿.

[18]Cipra, T. *Finanční a pojistné vzorce.* 1. vyd. Praha: GRADA Publishing, a.s., 2006. 376 s. ISBN 80-247-1633-X.

[19]SAS application, . *Application SAS - credit scoring* [online]. 2010, [citováno dne 20. května 2012]. Dostupné z: ¡http://www.cinco.mx¿.

Příloha A

Scoring živnostník

Název proměnné a její skutečné hodnoty	**Kódování**
Zisk v tis. Kč (za předchozí účetní období)	
X je neznámé nebo $X < 50$	0
$0 \leq X < 50$	4
$50 \leq X < 150$	8
$150 \leq X < 250$	12
$250 \leq X < 500$	16
$500 \leq X$	20
Obrat v tis. Kč (za předchozí účetní období)	
X je neznámé	0
$X < 1000$	2
$1000 \leq X < 2500$	5
$2500 \leq X < 6000$	10
$6000 \leq X < 12000$	15
$12000 \leq X < 20000$	20
$20000 \leq X$	25

Tabulka A.1: Proměnné vystupující v scoringu živnostníka - 1. část

Název proměnné a její skutečné hodnoty	**Kódování**
Délka podnikání	
$X < 1$	0
$1 \leq X < 50$	4
$2 \leq X < 150$	8
$4 \leq X < 250$	12
$8 \leq X < 500$	16
$12 \leq X$	20
Platební morálka	
$X \leq 1$	30
$1 < X \leq 4$	27
$4 < X \leq 8$	22
$8 < X \leq 13$	18
$13 < X \leq 21$	12
$21 < X \leq 30$	6
$30 < X$ nebo nový klient	0

Tabulka A.2: Proměnné vystupující v scoringu živnostníka - 2. část

Název proměnné a její skutečné hodnoty	**Kódování**
Záloha %	
$X \leq 5$	0
$5 < X \leq 10$	10
$10 < X \leq 20$	20
$20 < X \leq 30$	25
$30 < X \leq 40$	30
$40 < X \leq 55$	35
$55 < X$	40
Délka trvání smlouvy	
$X \leq 12$	15
$12 < X \leq 36$	12
$36 < X \leq 48$	9
$48 < X \leq 60$	6
$60 < X$	3
ObjektRisiko	
$X = 15$	20
$X = 20$	17
$X = 25$	13
$X = 30$	9
$X = 35$	6
$X = 40$	3
$X = 50$	0
Stáří objektu	
$X = 0$	10
$0 < X \leq 1$	13
$1 < X \leq 3$	20
$3 < X \leq 5$	15
$5 < X \leq 8$	8
$8 < X$	4

Tabulka A.3: Proměnné vystupující v scoringu živnostníka - 3. část

Příloha B

Zdrojový kód matlabu

B.1 Odhad parametrů a vyhodnocení úspěšnosti modelu

```
% Odhad parametrů a vyhodnocení úspěšnosti modelu
% Author: Jana Tíkalová
%Ukázka skriptu matlabu pro jeden zvolený model, chcete-li si daný skript v matlabu pustit,\\
%použijte skript uložený elektronicky v příloze této práce.

clear;clc;
data=importdata('vstup_enterpr.xls');
D30=data.data(:,1);
D90=data.data(:,10);
ObjektRisiko=data.data(:,2);
StariObjektu=data.data(:,3);
Zaloha=data.data(:,4);
TrvaniSmlouvy=data.data(:,5);
DelkaPodnik=data.data(:,6);
Obrat=data.data(:,7);
Zisk=data.data(:,8);
PlatebniMoralka=data.data(:,9);
% Model 1 D90
y = D90;
y(D90>1) = 0;
x=[ObjektRisiko StariObjektu Zaloha TrvaniSmlouvy DelkaPodnik Obrat Zisk PlatebniMoralka];
[b,dev,stats]=glmfit(x,y,'binomial','link','logit');
disp('odhady koeficientu')
disp(b)
sance=exp(b)
rozdil_sance=sance-1
save data1 b;
```

```
% Úspěšnost navrženého modelu
yfit=glmval(b,x,'logit');
y_nonan=y(~isnan(yfit));
yfit_nonan=yfit(~isnan(yfit));
success=zeros(length(yfit_nonan),1);
success(yfit_nonan>0.5)=1; %zde je nastavena pravděpodobnost CUTOFF
decl=zeros(length(yfit_nonan),1);
decl(success==y_nonan)=1;
disp('uspesnost modelu')
disp('spravne')
disp(sum(decl))
disp('spatne')
disp(length(yfit_nonan)-sum(decl))
```

```
% Dosazení do logitové funkce
disp('priklad 1')
logit=[9 10 25 12 20 0 0 0]*b
ppst=glmval(b,[9 10 25 12 20 0 0 0],'logit')
max_ppst_idx=find(yfit==(max(yfit)));
min_ppst_idx=find(yfit==(min(yfit)));
```

B.2 Validace testu

```
clear;clc;

data=importdata('validace_enterp.xls');
D30V=data.data(:,1);
D90V=data.data(:,10);
ObjektRisikoV=data.data(:,2);
StariObjektuV=data.data(:,3);
ZalohaV=data.data(:,4);
TrvaniSmlouvyV=data.data(:,5);
DelkaPodnikV=data.data(:,6);
ObratV=data.data(:,7);
ZiskV=data.data(:,8);
PlatebniMoralkaV=data.data(:,9);
load data1;

% Model 1 D30
yV30 = D30V;
yV30(D30V>1) = 0;
yV90 = D90V;
yV90(D90V>1) = 0;
x1V=[ObjektRisikoV StariObjektuV ZalohaV TrvaniSmlouvyV DelkaPodnikV ObratV ZiskV PlatebniMoralkaV];
x2V=[ZalohaV TrvaniSmlouvyV DelkaPodnikV ZiskV];

yVfit=glmval(b1,x1V,'logit');
disp(yVfit);
yV_non=yV30(~isnan(yVfit));
yVfit_non=yVfit(~isnan(yVfit));
disp(yVfit_non);
success=zeros(length(yVfit_non),1);
success(yVfit_non>0.5)=1;
decl=zeros(length(yVfit_non),1);
decl(success==yV_non)=1;

disp('uspesnost na validacnich datech DEFAULT 30 ppst 0,5 model 1')
disp('spravne')
disp(sum(decl2))
disp('spatne')
disp(length(yVfit_non2)-sum(decl2))
disp('uspesnost \%')
disp(sum(decl)/(sum(decl)+length(yVfit_non)-sum(decl)))
```

Příloha C

Jednoduchý manuál programu Gretl - spuštění skriptu

C.1 Co je Gretl?

Název programu Gretl je zkratkou vycházející z Gnu Regression, Econometrics and Time-series Library. Jedná se o software, který obsahuje základní jednoduše aplikovatelné ekonometrické nástroje. Tento software je volně dostupný a lze ho zdarma stáhnout z internetové adresy: gretl.sourceforge.net. Na této stránce je také možné najít podrobný manuál k celému software.

Gretl je možné využívat k celé řadě analýz dat včetně časových řad. Tento software dokáže při svých výpočtech používat nejen analytické postupy, ale využívá i iterační algoritmy. Výstupy z tohoto programu lze generovat do formátu .txt, .rtf ale i LaTexu.

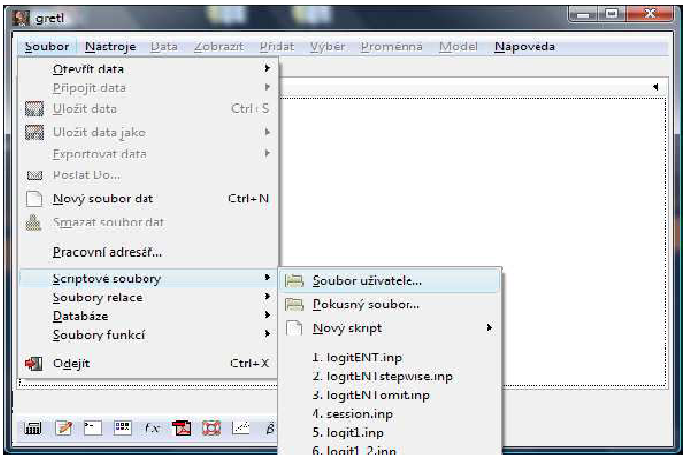
C.2 Instalace a základní ovládání programu

Instalace programu probíhá standardně prostřednictvím průvodce instalací, lze doporučit ponechat veškerá nastavení na přednastavených hodnotách, vyhneme se tak možným případným problémům a nutností doinstalovat některé doplňky dodatečně.

Gretl lze pak ovládat trojím různým způsobem, krom nejsnazšího ovládání prostřednictvím grafického uživatelského rozhraní (GUI), lze gretl ovládat konzolí na psaní kódu a systémovým příkazovým řádkem.

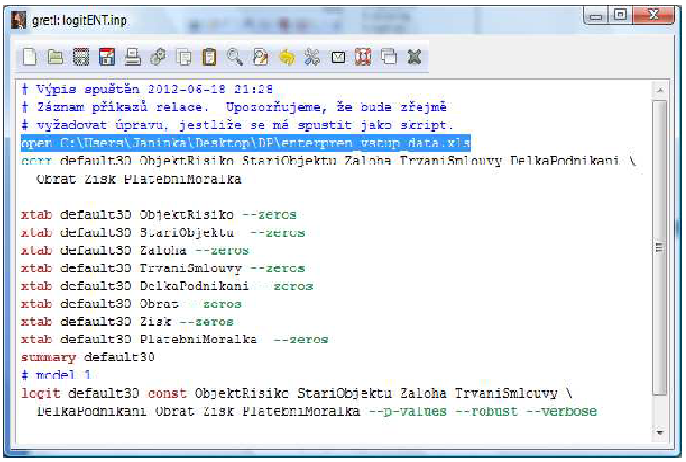
C.3 Spuštění skriptu

Skript spustím prostřednictvím GUI: Soubor → Scriptové soubory → Soubor uživatele... → a na disku vyberu soubor se skriptem, který chci spustit.



Obrázek C.1: Spouštění skriptu

Skriptový soubor sám neimportuje data, která jsou k analýze nutná, ovšem pozor, že cesta k datovému souboru ve formátu .xls je zadaná ve skriptu, proto musíte vždy přepsat první úkol ve skriptu, který otevírá a importuje data. Např. ve skriptu pro vyhodnocení živnostníka je využíván soubor enterpren_vstup_data.xls (open C:\Users\Janinka\Desktop\DP\enterpren_vstup_data.xls), cesta k tomuto souboru však byla zvolena a uložena při vytváření skriptu. Uživatel, který bude používat skript pro soubor, který bude mít uložený jinde, musí tuto část skriptu přepsat na aktuální uložení souboru!



Obrázek C.2: Vzhled skriptu - úprava místa uložení souboru

C.4 Další používání gretlu

Celý software lze samozřejmě používat i přímo k importu souboru a vytvoření modelu. Přes uživatelské prostředí se data importují následovně: Soubor → Otevřít data → Importovat → zvolím formát, ve kterém jsou data uložena např. Excel → a na disku vyberu soubor s daty, která chci používat.

Vytvoření modelu nad statistickým souborem lze poté v uživatelském rozhraní jednoduše volbou Model a proklikat se k modelu, který chci využít. Pro další používání gretlu lze doporučit materiály dostupné na internetových stránkách tvůrce programu, které byly avizovány již v úvodu tohoto stručného manuálu: gretl.sourceforge.net.

Příloha D

Obsah přiloženého CD

K této práci je přiloženo CD, na kterém jsou uložena veškerá vstupní data, zdrojové kódy a výstupy, které nebyly prezentovány přímo v práci.

- Složka 1: Enterpreneur
- Složka 2: Consumer
- Složka 3: Company
- Složka 4: Matlab
- Složka 4: LaTex