ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

PROVOZNĚ EKONOMICKÁ FAKULTA

KATEDRA STATISTIKY

Aplikace credit scoringového modelu v bankovní praxi

Bakalářská práce

Vedoucí práce:
Ing. Tomáš Hlavsa, Ph.D.

Vypracovala:
Helena Dobešová

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Helena Dobešová

Provoz a ekonomika

Název práce

Aplikace credit scoringového modelu v bankovní praxi

Název anglicky

Application of credit scoring model in banking practice

Cíle práce

Cílem bakalářské práce je vytvořit credit scoringový model. Dílčím cílem bude srovnání vybraných modelů a výběr takového, který bude k danému účelu nejvhodnější.

Metodika

Těžiště práce je postaveno na prediktivním modelování. K řešení bude možno využít např. regresní analýzu či rozhodovací stromy.

Doporučený rozsah práce

30 – 40 stran

Klíčová slova

Credit scoring, úvěr, žadatel, prediktivní model, logistická regrese, rozhodovací strom, Gini index, Weight of evidence, rating, behaviorální skóre, aplikační skóre, credit bureau

---

Doporučené zdroje informací

ABBOTT, D. Applied Predictive Analytics. United States of America: John Wiley & Sons, Inc., 2014. ISBN 978-1-118-72796-6

CIPRA, T. Riziko ve financích a pojišťovnictví: Basel III a Solvency II. Praha: Ekopress, s. r. o., 2015. ISBN 978-80-87865-24-8

RYBÁŘ, M., Identifikace úvěrové politiky banky. Univerzita Karlova v Praze, Matematicko – fyzikální fakulta, 2002

RYBÁŘ, M., Regresní modely a jejich výuka. Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, 2014

SIDDIQI, N. Intelligent Credit Scoring. New Jersey: SAS Institute, 2017. ISBN 978-1-119-27915-0.

TUFFÉRY, S. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

Předběžný termín obhajoby

2018/19 LS – PEF

Vedoucí práce

Ing. Tomáš Hlavsa, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 15. 1. 2019

prof. Ing. Libuše Svatošová, CSc.
Vedoucí katedry

Elektronicky schváleno dne 5. 2. 2019

Ing. Martin Pelikán, Ph.D.
Děkan

V Praze dne 12. 02. 2019

Čestné prohlášení

Prohlašuji, že jsem svou bakalářskou práci "Aplikace credit scoringového modelu v bankovní praxi" vypracovala samostatně pod vedením vedoucího bakalářské práce, s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autorka uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušila autorská práva třetích osob.

V Praze dne 15.3.2019

_____

Helena Dobešová

Poděkování

Ráda bych touto cestou poděkovala Ing. Tomáši Hlavsovi, Ph.D. za odborné vedení mé práce, za cenné připomínky, které mi pomohly danou problematiku úspěšně zpracovat, a za celkovou ochotu při vypracování této práce. Dále RNDr. Marianu Rybářovi za profesionální rady a podněty na zlepšení praktické části práce.

Název práce: Aplikace credit scoringového modelu v bankovní praxi

Autor: Helena Dobešová

Katedra: Katedra statistiky

Abstrakt:

Bakalářská práce se zabývá tvorbou credit scoringového modelu, který je aplikovatelný v praxi. Na základě dataminingového procesu byly vytvořeny tři dílčí modely na třech datových sadách, které jsou postaveny na datech v minulosti a předpovídají chování klientů v budoucnu. Jednalo se o model aplikační, kreditní a behaviorální. Dále byly nalezeny signifikantní proměnné, které mají na rozhodnutí o poskytnutí úvěru největší vliv. V metodické části je popsán kompletní postup výstavby prediktivního modelu a celé analýzy. Následně je objasněna nejdůležitější terminologie, kde je kladen důraz na vysvětlení dat, která banky sbírají, a rozdělení klientů. Poté jsou detailně popsány jednotlivé modely, jimiž musí klient v procesu schvalování projít. Pro účely analýzy klientských dat byly využity statistické metody explorační analýzy, prediktivní analýzy a regresní analýzy. Na základě těchto výsledků byly vytvořeny tři scorekarty a tři rovnice logistické regrese. Celý proces byl validován na testovacích datech. Nakonec byly modely demonstrovány na konkrétních klientech a tím potvrzeno jejich praktické využití. Na základě zjištěných informací a výsledků testů byla navržena možná praktická zlepšení celého úvěrového procesu z pohledu banky i klienta.

Klíčová slova:

Credit scoring, úvěr, žadatel, prediktivní model, logistická regrese, rozhodovací strom, Gini index, Weight of evidence, rating, behaviorální skóre, aplikační skóre, credit bureau

Title: Application of credit scoring model in banking practice

Author: Helena Dobešová

Department: Department of Statistics

Abstract:

The bachelor thesis deals with the creation of a credit scoring model, which is applicable in practice. Based on the data mining process, three sub-models have been created on three data sets that are based on data in the past and predict client behavior in the future. It was an application, credit and behavioral model. Furthermore, significant variables were found that have the greatest influence on the decision to grant a loan. The methodological part describes the complete construction of the predictive model and the whole analysis. Subsequently, the most important terminology is clarified, where the emphasis is on explaining the data that banks collect and the distribution of clients. Then the individual models that the client has to undergo in the approval process are described in detail. For the purpose of analyzing client data, statistical methods of exploratory analysis, predictive analysis and regression analysis were used. Based on these results, three scorecards and three logistic regression equations were created. The whole process was validated on test data. Finally, the models were demonstrated on specific clients and confirmed their practical use. Based on the information obtained and the test results, possible practical improvements of the whole credit process from the point of view of both the bank and the client were proposed.

Keywords:

Credit scoring, loan, applicant, predictive model, logistic regression, decision tree, Gini index, Weight of evidence, rating, behavioral score, application score, credit bureau

Obsah

1 Úvod 14

2 Cíl práce 15

3 Metodika 16

3.1 Statistické proměnné 16

3.2 Decision Tree (preselekce klientů) 17

3.3 Diagram postupu práce s daty 18

3.4 Práce s daty, postup prediktivní analýzy 19

3.4.1 Data understanding 19

3.4.2 Eliminace irelevantních proměnných 19

3.4.3 Splitting data 19

3.4.4 Statistic software 19

3.4.5 Data cleaning a missing data 19

3.4.6 Intuitive Behaviour (Weight of Evidence) 20

3.4.7 Single Factor Analysis 22

3.4.7.1 Information Value 22

3.4.7.2 Prediction Power 23

3.4.8 Multikolinearita 26

3.4.9 Long list of variables vs. Short list of variables 27

3.4.10 Multi Factor Analysis 28

3.4.10.1 Logistcká regrese 28

3.4.11 Cross-Validation 31

3.4.12 Satisfactory result, kalibrace 32

3.4.13 Scorekarta 32

3.4.14 Final Score 32

3.4.15 Cluster Analysis 32

3.4.16 Limit of the Loan 32

4 Teoretická východiska 34

4.1 Dlužník a věřitel 34

4.2 DTI, DSTI, DISPO 34

4.3 Úvěr 35

4.4 Druhy úvěrů 35

4.4.1 Rozdělení podle zdroje 35

4.4.2 Rozdělení podle dlužníka 36

4.4.3 Rozdělení podle doby splatnosti 36

4.4.4 Rozdělení podle způsobu zajištění ... 36
4.4.2 Rozdělení podle účelovosti ... 37
4.5 Credit scoring ... 38
4.6 Bankovní data ... 39
4.7 Rozdělení bankovních dat ... 40
4.7.1 Aplikační dataset ... 41
4.7.2 Behaviorální dataset ... 42
4.7.3 Dataset z externích databází ... 42
4.7.3.1 Bankovní Registr Klientských Informací ... 43
4.7.3.2 Nebankovní Registr Klientských informací ... 44
4.7.3.3 SOLUS ... 44
4.8 Rozdělení klientů ... 46
4.8.1 New to Bank klient ... 46
4.8.2 New to Market klient ... 47
4.8.3 Klient banky ... 48
4.9 Úvěrový proces ... 50
4.9.1 Diagram úvěrového procesu ... 52
5 Vlastní práce ... 53
5.1 Decision Tree ... 53
5.1.1 Diagram Decision Tree ... 54
5.2 Aplikační model ... 55
5.2.1 Data understanding ... 55
5.2.2 Eliminace proměnných ... 56
5.2.3 Splitting data ... 56
5.2.4 Statistic software ... 56
5.2.5 Data cleaning ... 57
5.2.6 Intuitive Behaviour – Weight of Evidence ... 59
5.2.7 Single Factor Analysis ... 65
5.2.8 Multikolinearita ... 67
5.2.9 Short list of variables a Long list of variables ... 71
5.2.10 Multi Factor Analysis ... 73
5.2.11 Cross-Validation, cut-off hladina ... 74
5.2.12 ROC křivka ... 78
5.2.13 Satisfactory result ... 80
5.2.14 Výsledná scorekarta aplikačních dat ... 81
5.2.15 Výsledná rovnice logistické regrese ... 83
5.2.16 Finale Score ... 85
5.2.17 Cluster Analysis ... 85
5.2.18 Limit of the Loan ... 85

5.3 Kreditní model 86
5.3.1 Data understanding 86
5.3.2 Splitting data 86
5.3.3 Intuitive Behaviour – Weight of Evidence 86
5.3.4 Long list of variables 88
5.3.5 Výsledky modelu 89
5.3.6 ROC křivka 90
5.3.7 Výsledná scorekarta kreditních dat 91
5.3.8 Výsledná rovnice logistické regrese 92
5.4 Behaviorální model 94
5.4.1 Data understanding 94
5.4.2 Splitting data 94
5.4.3 Intuitive Behaviour – Weight of Evidence 95
5.4.4 Predictive Power 95
5.4.5 Výsledky modelu 96
5.4.6 ROC křivka 97
5.4.7 Scorekarta behaviorálního modelu 98
5.4.8 Výsledná rovnice logistické regrese 100

6 Diskuze a výsledky 105
6.1 Nejlepší typy modelů 105
6.2 Výsledky modelů – praktická zjištění a návrhy 106
6.2.1 Porovnání hodnot Information Value 106
6.2.2 Power indexy 106
6.2.3 WOE vs. nebinovaná proměnná 106
6.2.4 Správné nastavení hladiny cut-off 107
6.2.5 Multikolinearita 107
6.2.6 Výběr výsledného listu proměnných 107
6.2.7 Cross-Validační chyba 107
6.2.8 Interakce 108
6.3 Získávání dat – praktická zjištění a návrhy 108
6.3.1 Aplikační data vs. behaviorální data 108
6.3.2 Zatajené informace 109
6.3.3 Další zdroje informací 109
6.4 Slabiny úvěrového procesu a jejich možná vylepšení 109
6.4.1 Doporučení pro klienta 110
6.4.2 Doporučení pro banku 111
6.5 Ideální klient pro banku 111
6.6 Náhodně vybraní klienti 113

7 Závěr 115

8 Seznam použitých zdrojů a použité literatury 117

9 Přílohy 121

Seznam obrázků, tabulek a grafů

1. Diagram rozdělení statistických proměnných 16
2. Tabulka rozdělení hodnot Information Value 22
3. Graf – ukázka ROC křivky 25
4. Tabulka korelační závislosti 27
5. Tabulka Confusion Matrix 31
6. Obrázek ukázka clusterů 33
7. Obrázek rozdělení úvěrů 37
8. Obrázek ukázka scorekarty 40
9. Obrázek rozdělení bankovních dat 40
10. Obrázek příklad aplikačních dat 41
11. Obrázek příklad behaviorálních dat 42
12. Obrázek příklad dat z externích databází 42
13. Obrázek rozdělení bankovních dat 45
14. Obrázek příklad dělení bankovních klientů 46
15. Obrázek součet skóre NTB klienta 46
16. Obrázek součet skóre NTM klienta 47
17. Obrázek součet skóre BC klienta 48
18. Diagram součtů skóre klientů 49
19. Obrázek rozdělení proměnné y 55
20. Obrázek Box plot před očištěním dat 58
21. Obrázek Box plot po očištění dat 58
22. Grafy – WOE pro proměnnou Age 59
23. Obrázek nastavení hodnot v programu STATISTICA13 60
24. Graf – WOE pro proměnnou Age s 5 % účastí v každém binu 61
25. Grafy – WOE pro proměnnou Age muži a ženy 62
26. Tabulka – WOE hodnoty proměnné Age 63
27. Graf WOE HousingStatus 64
28. Tabulka – WOE hodnoty proměnné HousingStatus 64
29. Tabulka – výsledné hodnoty indexů a Information Value 65
30. Graf – ROC křivka Income 66
31. Graf – ROC křivka WOE_Income 67
32. Tabulka – Correlation Matrix 68

33. Tabulky – výsledné hodnoty χ^2 testů 70
34. Tabulka – Short list of variables 71
35. Tabulka – Long list of variables 72
36. Tabulka – kombinace modelů 73
37. Tabulka – Cross-Validace MODEL03 74
38. Tabulka – Cross-Validace MODEL04 75
39. Tabulka – Cross-Validace MODEL09 75
40. Tabulka – Cross-Validace MODEL10 75
41. Tabulka – Cross-Validace MODEL09, cut-off 0,85 77
42. Tabulka – Cross-Validace MODEL10, cut-off 0,85 77
43. ROC křivka MODEL09 78
44. ROC křivka MODEL10 78
45. Tabulka – Cross-Validace MODEL10, cut-off 0,85 79
46. Graf – Histogram MODEL10, cut-off 0,85 79
47. Tabulka – Cross-Validace MODEL10, cut-off 0,87 80
48. Tabulka – Finální scorekarta aplikačních dat 82
49. Obrázek ukázka clusterů 85
50. Graf WOE proměnná Product_Type 87
51. ROC křivka proměnné Product_Type 87
52. Tabulka – Long list of variables 88
53. Tabulka – Cross-Validace kreditního modelu 89
54. Graf – Cross-Validace kreditního modelu 89
55. ROC křivka kreditního modelu 90
56. Scorecard kreditní data 91
57. Graf – WOE BalMinPast4Q 95
58. Graf – ROC křivka BalMinPast4Q 95
59. Tabulka – Cross-Validace behaviorální model 96
60. Graf – histogram výsledky Cross-Validace behaviorálního modelu 97

Seznam zkratek

AS	Aplikační skóre
AUC	Area Under Curve
BC	Bankovní klient
BRKI	Bankovní registr klientských informací
BS	Behaviorální skóre
CBCB	Czech Banking Credit Bureau
CNCB	Czech Non-Banking Credit Bureau
CS	Kreditní skóre
DISPO	Minimální disponibilní příjem
DSTI	Debt Service to Income
DTI	Debt to Income
IV	Information Value
KPI	Key Performance Indicator
NRKI	Nebankovní registr klientských informací
NTB	Klient New to Bank
NTM	Klient New to Market
ROC	Receiver operating characteristic
SOLUS	Sdružení na Ochranu Leasingu a Úvěrů Spotřebitelů
WOE	Weigt of Evidence

1 Úvod

Všechny instituce, které se zabývají otázkou consumer finance (spotřební financování), mají postavené své rozhodování o přidělení úvěru na rozsáhlé prediktivní analýze. Jedná se o miliardy korun ročně, které jsou rozpůjčovány na základě vyhodnocení předešlé analýzy klientovy bonity.

Dnešní úvěrový trh se dynamicky mění. Umožňuje poskytnutí úvěru na počkání a je rychlý a plný konkurence. Klienti jsou netrpěliví a naučili se přistupovat k pořizování zboží stylem „koupit teď, zaplatit později“. K úvěrovým nabídkám přistupují podobně. Po žádosti o úvěr se dožadují okamžitého výsledku, zda úvěr dostanou či nikoliv. Proto je rychlost, přesnost a jednoduchá interpretace jedna z podmínek správně sestaveného modelu, jinak by došlo kvůli táhlému rozhodování ke ztrátě obchodní příležitosti a klienta.

Každá banka přistupuje k úvěrovému riziku z jiného úhlu pohledu. Všechny se však shodnou na důležitosti správně sestaveného credit scoringového modelu, který bude predikovat spolehlivé výsledky. Při hodnocení rizikovosti klienta neexistuje žádná stoprocentně správná metoda. K perfektní předpovědi se však mohou banky dosti přiblížit. Zároveň by finanční instituce měly poskytovat prostředky za podmínek a za úroky, které pokryjí případné úvěrové riziko a nesolventní klientské jednání.

Banka v žádosti o úvěr sbírá o klientech velké množství dat. Žadatel se ale nedozví signifikantní proměnné, které celý proces schvalování ovlivní a rozhodnou. Finanční instituce nezveřejňují své metodiky ani postupy, aby nebylo možné ze strany klienta ovlivnit celý úvěrový proces ve svůj prospěch. Ani klientští pracovníci v bance nemají podrobný náhled na konkrétní kroky celého procesu a jsou pouze uživateli již hotových modelů.

Vzhledem k aktuálnímu dění, kde je úvěrový trh poměrně silně nasycen, je pro finanční instituce velkou motivací, aby se zabývaly celým procesem do hloubky. Dojde-li například k hospodářské krizi, budou muset banky reflektovat celou situaci, která se také promítne do rozhodování o poskytování úvěrových produktů. Ta se už teď velice zpřísňuje (Pečená, 2010).

Prediktivní analýza se za posledních několik let posunula mílovými kroky kupředu a celý proces rozhodování již nestojí pouze na jedné statistické metodě. Jedná se o mnohem sofistikovanější a propracovanější postup a aplikaci hned několika kombinací statistických metod. Prediktivní analýza a datamining se využívají téměř ve všech odvětvích, kde se pohybuje klient. Od různých nabídek produktů, přes zjištění budoucího chování na trhu, až po marketing.

Tato práce je zaměřena na objasnění celé problematiky úvěrového procesu a na nalezení optimálního credit scoringového modelu, který by zajistil zdravé úvěrování klientů a zároveň maximalizoval zisk při současném řízení rizika a minimalizace ztrát. Celou problematiku prediktivního modelování a objasnění postupu bude doprovázet provazba s praktickým využitím v bankovní praxi, kde bude poukázáno na jednotlivé hrozby celého procesu schvalování a poskytování úvěrů.

2 Cíl práce

Cílem této práce je vytvoření credit scoringového modelu. Dílčím cílem bude vytvoření tří modelů na třech datových sadách, které budou aplikovatelné v praxi a budou mít co možná nejlepší vypovídající charakter. Simultánně se bude práce zabývat slabinami celého úvěrového procesu a modelu, se kterými je možné se potkat v praxi. Dále budou navržena možná vylepšení celého procesu. V neposlední řadě se bude práce zabývat nalezením vhodného nastavení cut-off hladiny, aby banka co nejvíce vytěžila tržní potenciál za současné minimalizace ztrát. V závěru budou představeny proměnné, které by mohly mít největší vliv na predikci modelu. Nedojde však ke kompletnímu odhalení všech proměnných, které banky sbírají, aby nedošlo k vyzrazení know-how.

Dalším cílem je seznámení čtenářů s hloubkou celého procesu a zorientování v problematice. Celý proces bude demonstrován na třech náhodně vybraných klientech s kompletním výsledkem prediktivní analýzy. Ve výsledcích bude demonstrován ideální klient pro banku z každého datasetu.

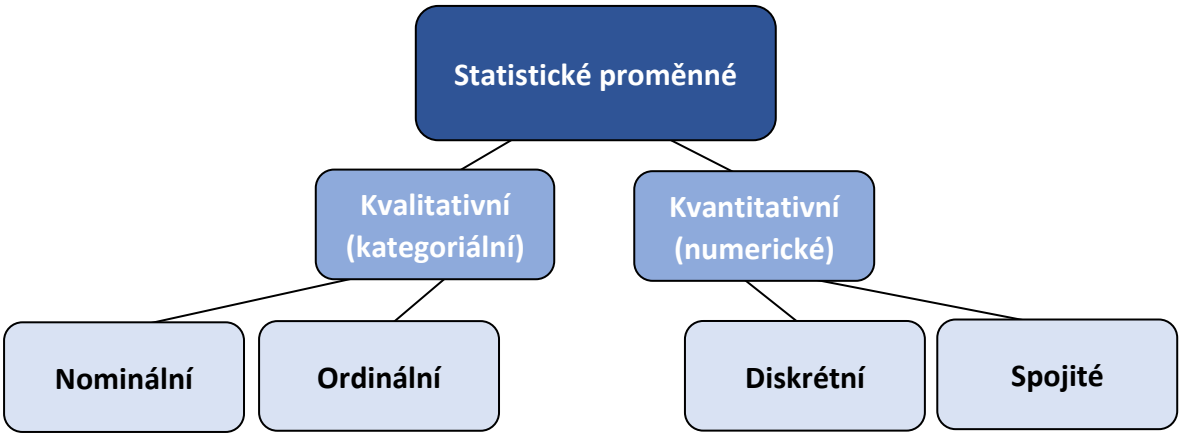
3 Metodika

Celá práce se dělí na dvě hlavní části. První část se zabývá teoretickými východisky. Popisuje celou problematiku a s ní spojenou terminologii, která byla zpracována na základě odborné literatury a podrobného studia dostupných pramenů. Vzhledem k poměrně náročné tematice byla jako zdroj informací volena především literatura zahraniční. Z toho důvodu a z důvodu praktického využívání anglické terminologie jsou v celé práci upřednostěny anglické výrazy oproti českým překladům. Vždy jsou podrobně vysvětleny, aby se čtenář zorientoval ve významu.

Druhá část je zaměřena na vlastní práci, ve které je vyhotoven credit scoringový model a zpracovány tři datové soubory a vyhotoveny tři modely a tři scorekarty. První datový vzorek byl použit na vyhotovení aplikačního modelu, druhý na kreditní model a třetí na model behaviorální.

3.1 Statistické proměnné

Aplikační soubor čítal data od cca 15 000 klientů (50 proměnných), behaviorální data od cca 11 000 klientů (320 proměnných) a kreditní data od cca 8 000 klientů (100 proměnných). Pro správný rozbor statistických dat a pro zvolení vhodného postupu prediktivní analýzy bylo nutné rozdělit jednotlivé proměnné (náhodné veličiny) na příslušné typy. Podle rozdělení se s každou proměnnou jinak pracovalo. Je několik možných dělení jednotlivých proměnných. Pro účely využití v credit scoringovém modelu postačilo dělení na *kvalitativní (kategoriální)* a na *kvantitativní (numerické)*.



1) Diagram rozdělení statistických proměnných. Zdroj: Hindls et al., 2014

Kvalitativní, jak již název napovídá, vyjadřují kvalitu, kategorii či pořadí. Jsou to takové hodnoty, které nelze měřit. V aplikačním datasetu se typicky jednalo o proměnnou *Nationality - národnost*. Dále se dělí na *nominální*, což byla například proměnná *RegionCont - kraj trvalého bydliště* (Středočeský kraj, Kraj Vysočina, Praha) a na *ordinální (pořadové)*, kupříkladu proměnná *Education - vzdělání*, která byla vyjádřena v hodnotách ZŠ, SŠ a VŠ.

Kvantitativní proměnné byly například příjmy – *Income*, počty – *PersonWithNoInc – počet členů domácnosti bez příjmu*, bilanční hodnoty nebo zůstatky. Dále se tyto proměnné dělily na diskrétní a spojité. *Diskrétní* se vyjadřují v celých číslech a mají konečný počet variant. Z kreditních dat to byla například proměnná *počet kreditních karet* a *počet dětí.* Pro příklad proměnné *spojité* lze úvést *Income - příjem, počet odpracovaných měsíců v současném zaměstnání, atp***.** (Hindls et al., 2014)

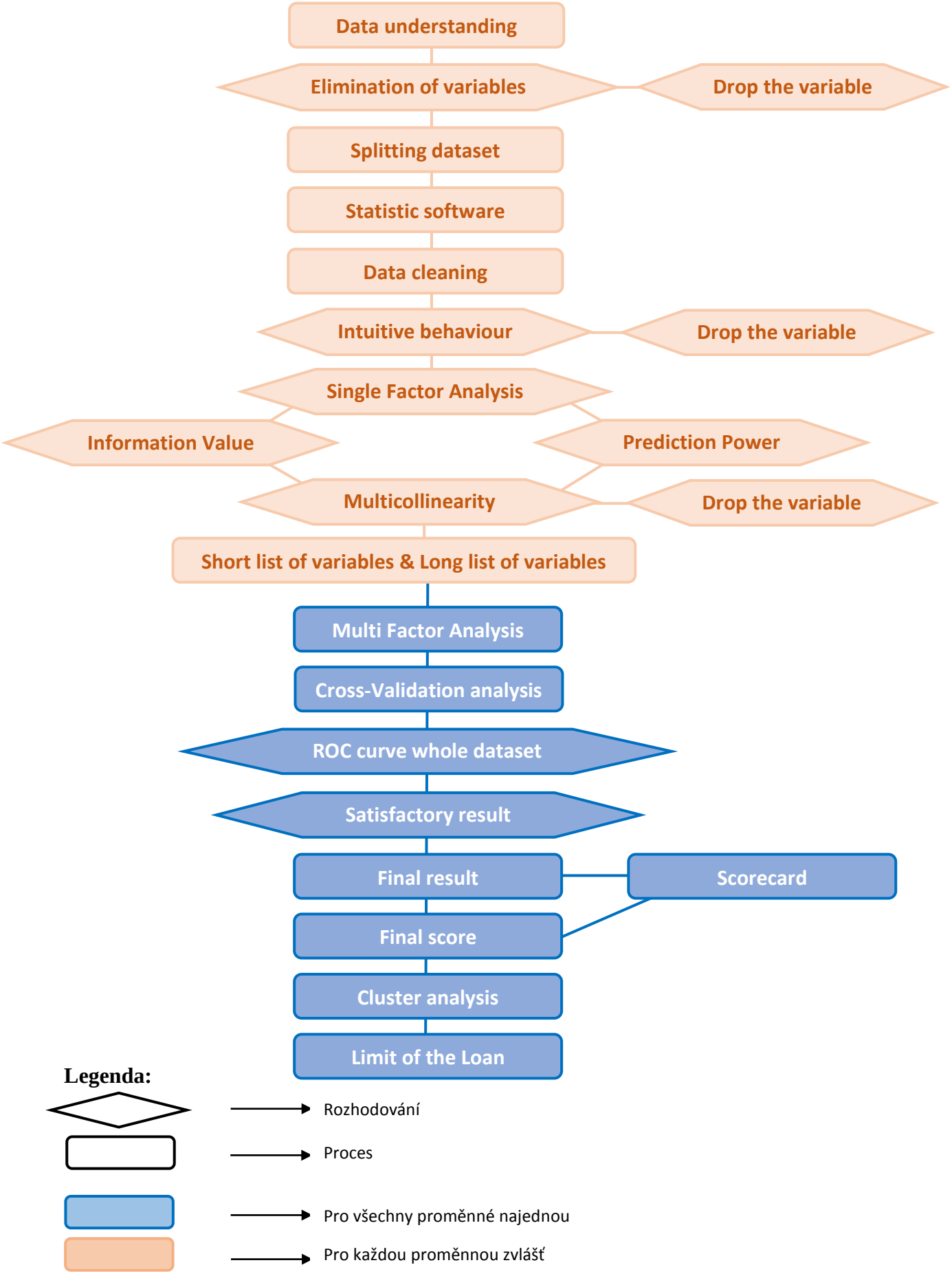
3.2 Preselekce klientů pomocí metody Desicion Tree

Před samotným sběrem dat, která dále vstupovala do modelu, musel klient projít rozhodovacím stromem. Cílem bylo absolvovat všechny uzly stromu a pokračovat dále do modelu.

Decision Tree je jedna z nejlépe interpretovatelných a jednoduše pochopitelných rozhodovacích dataminingových technik. Jde o princip kladení otázek, na které je následně odpovídáno.

Strom bylo velice jednoduché sestavit. V této metodě šlo pracovat s kvantitativními (numerickými) i kvalitativními (kategoriálními) proměnnými. Každý uzel se dal jednoduše interpretovat, jako *„if-then-else“,* což v doslovném překladu znamená *„pokud-tak-jinak“*. Všechna rozvětvení stromu představovala rozhodování podle podstaty objektu, z tohoto rozvětvení vedl konečný počet cest. Nejčastěji se jednalo o tzv. *binary outcome* – binární výstup (odpovědi: ano, ne). Nejlepší interpretace této metody je grafické znázornění. (Abbott, 2014, s. 214-223)

3.3 Diagram postupu práce s daty. Zdroj informací: Siddiqi, 2017



3.4 Práce s daty a podrobný postup prediktivní analýzy

Celý postup prediktivní analýzy jde v následujících krocích:

3.4.1 Data understanding

Podle Abbotta (2014) data understanding zahrnuje bližší ohledání, prohlídku a přípravu datového souboru. Tento krok byl rozhodující pro vyloučení neočekávaných budoucích problémů. Jedná se obvykle o nejdelší část projektu. Data se musela popsat, prozkoumat, připravit a ověřit jejich kvalita. Pochopení dat zahrnuje přístup k datům a jejich prozkoumání například pomocí tabulek a grafů. To umožňuje určit kvalitu dat a popsat výsledky hodnocení těchto kroků v projektové dokumentaci.

3.4.2 Eliminace irelevantních proměnných na základě expertního odhadu

V datovém vzorku byly proměnné, které nemělo smysl dále analyzovat, protože nepřinášely žádné informace, které by napomohly učinit rozhodnutí. Z toho důvodu byly datové soubory adjustovány od těchto proměnných.

3.4.3 Splitting data

V tomto kroku šlo o rozdělení datového souboru na dva vzorky (testovací a trénovací). Na trénovacím vzorku byl postaven model, na testovacím později vyzkoušena prediktabilita modelu pomocí metody Cross-Validace.

3.4.4 Stastistic software

Převod dat ze soboru MS Excel do vhodného statistického software. K vypracování statistických metod byl využit nejvíce software STATISTICA13 a software R.

3.4.5 Data cleaning a missing data

Klientská data jsou zadávána do počítače bankovními pracovníky nebo samotným klientem prostřednictvím internetového bankovnictví. Ztohoto důvodu bylo čištění dat velice důležité. Ve velkém datovém souboru je pravděpodobný výskyt chyb, popř. *outlierů* – odlehlých hodnot, které by mohly zkreslit celý model.

Jako vhodná metoda pro čištění dat bylo zvoleno grafické zobrazení pomocí Box plotů, které na první pohled velice přehledně znázorňovaly anomálie nebo chyby dat. Jedná se o znázornění umožňující posouzení dat pomocí kvartilů a je spolehlivou metodou k rychlému nahlédnutí do vlastností číselných dat, jako je průměr, medián, extrémy nebo odlehlé hodnoty (Abbott, 2014, s. 61).

Chybějící data (Missing data) byla například v aplikačním datasetu poměrně vzácným jevem vzhledem k charakteristice obsahu proměnných (věk, pohlaví, způsob bydlení, atp.).

Dle Finlayho (2010) je přístupů, jak se dá s tímto jevem vypořádat, hned několik. Jednou z možností bylo nahrazení dané proměnné hodnotou, která dává v kontextu celého klienta smysl. Druhým častým přístupem byla eliminace celé proměnné, pokud obsahovala velké množství chybějících dat, nebo odebrání celého klienta z datového souboru.

3.4.6 Intuitive Behaviour pomocí metody Weight of Evidence

Dle Siddiqiho (2017) je pro indikaci správného chování daných proměnných a dosažení maximální přesnosti predikce vhodná statistická metoda Weight of Evidence (dále jen WOE).

Celý postup byl postaven na rozkategorizování jednotlivých proměnných dle intuitivního chování. Proměnné následně vstupovaly do modelu již rozdělené. V případě, že chování dané proměnné nebylo logické, byla odstraněna ze souboru. U podobných skupin v rámci jednoho faktoru došlo k agregaci. Šlo o pohled na celou proměnnou, kde bylo důležité si na základě intuice a osobního zhodnocení uvědomit, že například ne všichni žadatelé o úvěr jakéhokoliv věku splácejí stejně. Proto nelze poměřovat stejně klienty 18leté a 54leté. U každé skupiny se vyskytovala jiná míra rizikovosti a jiná šance na splacení. Na základě rozdělení byl přiřazen každé skupině z rozkategorizované proměnné koeficient dle statistického vlivu na fakt, zda daný klient bude splácet dle pravidel či nikoliv. Takový model byl potom vyjádřen ze vzorce:

$$WoE = \left[ln\left(\frac{Distr\ Goods}{Distr\ Bads}\right)\right] * 100$$

Kde:

$$\frac{Distr\ Goods}{Distr\ Bads} = \frac{\frac{g_i}{G}}{\frac{b_i}{B}}$$

Distr Goods je zkratka pro *Distribution of Good Credit Outcomes* a *Distr Bads* je *Distribution of Bad Credit Outcomes.*

Distr Goods je označení pro poměr g_i/G, kde g_i je počet dobrých účtů v intervalu a G je suma všech dobrých účtů v celém souboru.

Distr Bads je označení pro poměr b_i/B, kde b_i je počet špatných účtů v intervalu a B suma všech špatných účtů v celém souboru (Finlay, 2014).

Multiplikace pomocí hodnoty 100 pomohla výsledky lépe interpretovat a dále s nimi jednodušeji pracovat.

Vzhledem k použití přirozeného logaritmu nabývala WOE hodnotu 0 v případě, že poměr dobrých a špatných účtů je roven 1. Kladné hodnoty značí větší počet správně splácejících klientů a záporné hodnoty větší poměr špatně splácejících klientů.

Při rozdělování dané proměnné na jednotlivé intervaly nebo kategorie musel být zohledněn fakt, že každý *bin* má dle Siddiqiho (2017) v ideálním případě obsahovat nejméně 5 % případů celého faktoru. Počet kategorií určuje chování proměnné. Menší počet binů zapříčiní hladší průběh proměnné a zároveň vylučuje šum. Kategorie s menším poměrem klientů, než je 5 %, nemusí být pravdivým obrazem distribuce dat a mohou způsobit nestabilitu modelu.

Výchozí hodnoty z modelu WOE dále vstupovaly do výpočtu Information Value, na základě kterého byly již rozkategorizované proměnné buď ponechány v modelu, nebo odstraněny.

3.4.7 Single Factor Analysis

Single Factor Analysis (jednofaktorová analýza), nebo také univariantní analýza dat, je technika analýzy faktorů, která se používá ke snížení velkého počtu proměnných na méně faktorů. Na základě single faktor analýzy docházelo k identifikaci proměnných, které měly největší vliv na rozhodování. Ačkoliv jsou finální modely postaveny pomocí více proměnných, bylo důležité, aby každá proměnná měla samostatně vysvětlující sílu a intuitivní vztah k predikci špatných účtů. Siddiqi (2017) udává, jako příklady metod používaných k posouzení prediktivní výkonnosti binovaných proměnných, využití dvou indikátorů: Information Value a Prediction Power.

Během designu celého modelu nebo rozšířeného přehledu bylo nutné analyzovat vztah s cílovou proměnnou, stabilitu v čase a vypovídající hodnotu proměnné.

3.4.7.1 Information Value

Dle Siddiquiho (2017) je Information Value jednou z nejužitečnějších metod pro výběr signifikantních proměnných v credit scoringovém modelu. Pomáhá třídit proměnné na základě jejich významu již po rozkategorizování pomocí metody WOE. Hodnota Information Value se vyjadřuje podle následujícího vzorce:

$$IV = \sum_{i-1}^{n} \left[(Distr\ Goods_i - Distr\ Bads_i) * ln\left(\frac{Distr\ Goods}{Distr\ Bads}\right) \right]$$

V tomto případě hodnota WOE vstupovala do vzorce jako desetinné číslo.

Information Value nabývá následujících hodnot:

Information Value	Prediction Variables
Less then 0,02	Generally unpredictive
od 0,02 do 0,1	Weak prediction
Od 0,1 do 0,3	Medium prediction
0,3+	Strong prediction

2) Tabulka – hodnoty Information Value, Zdroj: Siddiqi, 2017

Podle výsledných hodnot bylo možné interpretovat hodnoty Information Value v hodnocení credit scoringového modelu následovně:

Byla-li hodnota Information Value:

1. menší než 0,02, potom byl prediktor nepoužitelný pro modelování. Každá taková proměnná, u které klesla hodnota Information Value pod 0,02, byla z listu proměnných odebrána.

2. 0,02 až 0,1, prediktor měl pouze slabý vztah k poměru pravděpodobnosti Goods / Bads. V listu proměnných byl ale ponechán.

3. 0,1 až 0,3, potom měl prediktor střední vztah k poměru pravděpodobnosti.

4. 0,3 a více, prediktor měl silný vztah k poměru pravděpodobnosti.

5. nad 0,5, šlo detekovat podezřelý vztah a musela být proměnná prověřena na tzv. overpredicting (www.listendata.com, 2019).

Důležitá zjištěná fakta:

Information Value se zvyšovalo, jestliže se zvyšoval počet binů pro nezávislou proměnnou. Pokud se proměnná rozdělila na více než 20 binů (tzn. pod 5 % účast všech klientů z celého faktoru v každém binu), v některých kategoriích byl velice malý počet měření a vypovídající hodnota byla zavádějící.

3.4.7.2 Prediction Power

Jedná se o měření síly jednotlivých proměnných a následně i celého modelu. Využívají se metody jako: ROC křivka, Gini koeficient, Somer's D index nebo Kolmogorov-Smirnov index.

Receiver operating characteristic curve (dále ROC křivka)

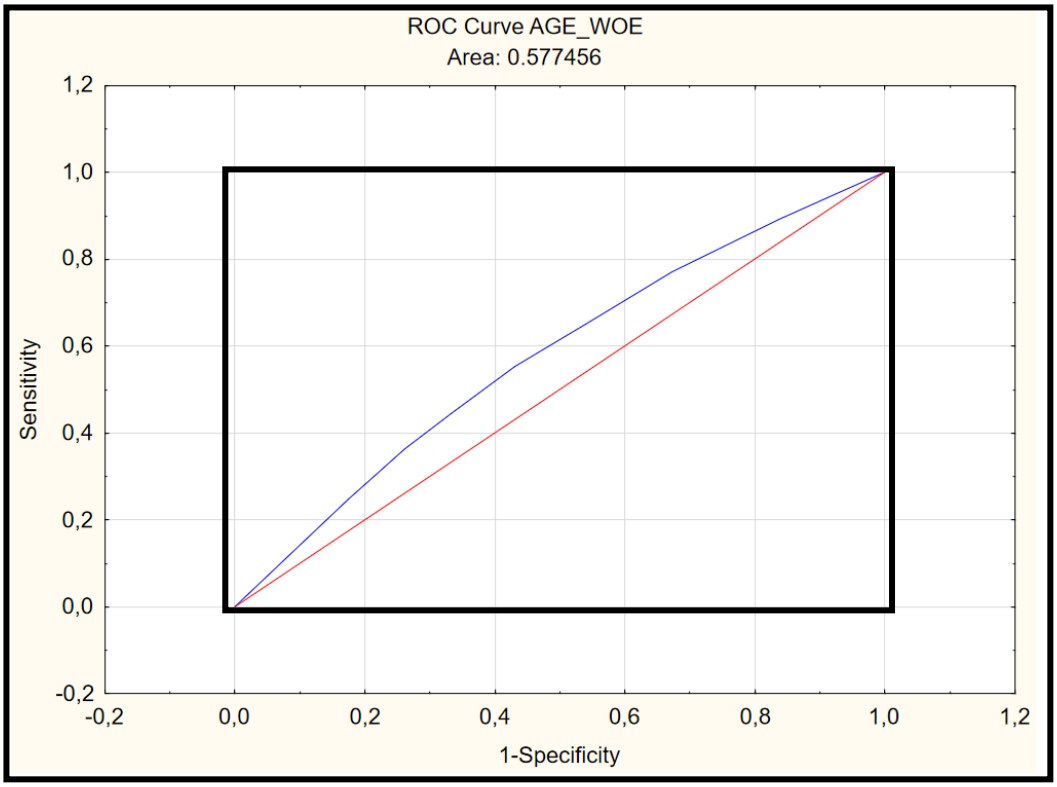
Tufféry (2011) pojednává o tom, že ROC křivka je grafické znázornění, které ukazuje diagnostické schopnosti binárního klasifikačního systému oddělit špatné a dobré účty. Měření odpovídá ploše pod křivkou – tzv. AUC, neboli area under curve. ROC křivka měří klasifikační výkonnost na jedné proměnné i celé scorekartě. Náhodná čára (45 stupňů) označuje sílu predikce 0,5. Proto by mělo skóre dosahovat vyšších výsledků než 0,5.

U ROC křivek se zkoumá citlivost a specifičnost modelu, což jsou statistické ukazatele výkonnosti binárního klasifikačního testu.

Citlivost (Sensitivity - nazývána také jako skutečná pozitivní míra) měří podíl správně identifikovaných pozitivních výsledků (například procento nemocných, kteří jsou správně označeni jako nemocní). V případě credit scoringového modelu by se jednalo o špatné účty, které byly správně identifikovány jako špatné.

Specifičnost (Specificity - nazývána též jako skutečná negativní míra) měří podíl správně identifikovaných negativních nálezů (například procento zdravých lidí správně označených jako zdraví). V případě credit scoringu by se jednalo o správné označení dobrých účtů.

ROC analýza je efektivním způsobem při rozhodování o diagnostice modelu. Poskytuje nástroje pro výběr optimálního řešení (Siddiqi, 2017, str. 542-548).



3) Graf: Ukázka ROC křivky na proměnné Age_WOE. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

Gini koeficient

Gini koeficient je vyjádřením míry diverzifikace scoringového modelu nebo dané proměnné. Nabývá hodnoty od 0 do 1. Hodnota 1 udává perfektní diverzifikační schopnost, 0 značí špatnou diverzifikační schopnost (Tufféry, 2011). Lze vyjádřit ve tvaru:

$$Gini = 2 * AUC - 1$$

Tufféry (2011) dále pojednává o tom, že Somers'D index je téměř identický s Gini.

Kolmogorov-Smirnov index (KS index)

Tento index měřil maximální svislou odchylku mezi kumulativním rozdělením Goods a Bads účtů. Je velmi rozšířeným využitím vyjádření rozdělení. Problém KS indexu spočívá v tom, že měří rozdělení pouze v bodě (který nemusí být kolem očekávaného mezního bodu) a nikoliv na celém rozsahu skóre. (Siddiqi, 2017, s.253)

3.4.8 Multikolinearita – vzájemná závislost dvou vysvětlujících proměnných

Ještě před vstupem proměnných do multifaktor analýzy bylo nutné otestovat případný výskyt multikolinearity mezi vysvětlujícími proměnnými na pravé straně rovnice, aby se předešlo k ovlivnění výpočtu jednotlivých prediktorů z důvodu nadbytečnosti některých proměnných.

Ve statistice je multikolinearita jev, kde dochází k nechtěné závislosti mezi dvěma vysvětlujícími proměnnými. To znamená, že lze jednu proměnnou prediktivně předpovídat od jiné. V případě silné závislosti mezi vysvětlovanou proměnnou a vysvětlující proměnnou se jedná o jev žádoucí. Silná závislost mezi dvěma vysvětlujícími proměnnými je však nežádoucí fenomén. Ponechání takové proměnné může vést k nestabilitě regresních koeficientů, proto musely být z listu proměnných, které dále postupují modelem, odebrány.

Pro výpočet závislostí vysvětlujících proměnných před finálním vstupem do modelu byly spojité proměnné testovány pro případný výskyt multikolinearity pomocí Pearsonových korelačních koeficientů uspořádaných do korelační matice (Correlation Matrix). V případě kategoriálních proměnných byly pro identifikaci multikolinearity využity Pearsonovy kontingenční koeficienty. Za vysokou multikolinearitu byly považovány hodnoty koeficientů vyšší nebo rovny 0,8 (v absolutní hodnotě). (Čechura et al., 2013)

V případě, že takový jev nastal, musela být jedna ze závislých proměnných z listu odebrána. Většinou se odebírala proměnná, která nabyla menší hodnoty Information Value nebo byla dražší, co se týká sběru dat, nebo nedávalo logiku ji nadále ponechávat v kontextu s ostatními proměnnými.

Dále pokračovaly do modelu vysvětlující proměnné, které po vzájemném testování neprokazovaly vysokou korelaci.

Pearsonův korelační koeficient

Může nabývat hodnot <-1; +1>. Vyjadřuje se pomocí vzorce (Hindls et al., 2007):

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2] * [n\sum y_i^2 - (\sum y_i)^2]}}$$

Korelační závislost:

Korelační koeficient r	Síla závislosti
r = 0	nulová závislost
0 < | r |< 0,3	slabá závislost
0,3 ≤ | r | < 0,5	střední závislost
0,5 ≤ | r | < 0,8	vysoká závislost
0,8 ≤ | r | ≤ 1	velmi silná závislost

4) Tabulka – korelační závislost, zdroj: https://docplayer.cz/

Personův kontingenční koeficient

Vyjadřuje se podle vzorce (Kába et al., 2013):

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

3.4.9 Long list of variables vs. Short list of variables

Po takto otestovaných proměnných vznikl tzv. Long list of variables a Short list of variables.

Long list of variables byl seznam proměnných s hodnotou Information Value větší nebo rovno 0,02.

Short list of variables byl seznam proměnných, které získaly v testování hodnotu IV větší nebo rovno 0,02 a hodnoty Prediction Power vyšší nebo rovno 0,1.

3.4.10 Multi Factor Analysis

Vzhledem k seznamu proměnných vybraných v analýze jednotlivých faktorů bylo cílem multifaktorové analýzy zvážit účinky různých kombinací těchto vstupních proměnných pro výběr konečné sady proměnných a jejich odpovídajících koeficientů. Modelové skóre bylo pak získáno násobením každé proměnné s její váhou a výsledným koeficientem.

Během analýzy více faktorů bylo třeba zkontrolovat, která kombinace proměnných je nejpříznivější, a přitom je stále přijatelná pro uživatele. K určení optimální kombinace proměnných existuje řada statistických metod. Pro regresní analýzu binární proměnné bylo vhodné využití regrese logistické.

Cílem regresní analýzy bylo navržení nejvhodnějšího modelu s největší možnou prediktivní sílou.

3.4.10.1 Logistická regrese

Siddiqi (2017) pojednává o tom, že logistická regrese je statistická metoda prediktivního modelování pro analýzu datové sady. Zabývá se problematikou odhadu pravděpodobnosti závislé proměnné na základě proměnných nezávislých, které určují výsledek. Od standardní regresní analýzy se odlišuje výskytem binární závislé proměnné *y* na levé straně rovnice. Tato binární proměnná *y (*konkrétně šlo v testovacích sadách vždy o proměnnou *Dependent_12M)* byla následně převedena logitovou transformací na pravděpodobnost *p,* což značí nastání určitého jevu. Logistická regrese je běžnou technikou používanou k vytváření scorekaret.

Logistická regrese generovala koeficienty, které předpovídaly logitovou transformaci pravděpodobnosti přítomnosti charakteristické vlastnosti. Rovnice pro transformaci pravděpodobnosti vycházela z následujících vztahů:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

A zároveň platí pro transformaci logitu vzorec:

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

Z uvedených vzorců plynulo odvození (dle vlastního odvození):

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k$$

$$e^{ln\left(\frac{p}{1-p}\right)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}$$

$$\left(\frac{p}{1-p}\right) = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}$$

$$p = (1-p) * e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}$$

$$p = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k} - p * e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}$$

$$p + p * e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}$$

$$p * (1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}) = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}$$

$$\boldsymbol{p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k}}}$$

Tento vzorec byl finální pro výpočet pravděpodobnosti splacení úvěru klientem.

Jak píše Siddiqi (2017), v praxi se využívají tři typy postupných logistických regresních technik, které byly jednotlivě testovány. Cílem bylo nalezení takové techniky, která vytvořila nejlepší model, který predikoval spolehlivé výsledky:

Forward selection (dopředný výběr)

Nejprve se vybírala proměnná, která měla nejlepší individuální prediktivní výkon. Dále se k modelu přidávaly další charakteristiky dle nejlepšího výkonu, aby se postupně vytvořil nejlepší charakteristický model, dokud žádná zbývající proměnná neměla hodnotu *p* pod hladinou významnosti (0,05). Tato metoda je účinná, ale může být slabá, pokud existuje příliš mnoho charakteristik nebo vysoká korelace mezi proměnnými.

Backward elimination (zpětné vyloučení)

Opačný postup byl backward elimination, kde do modelu nejdříve vstupovaly všechny proměnné a postupně se eliminovaly nejméně významné vzhledem k ostatním. Celý postup se opakoval, dokud nebyly v modelu pouze ty proměnné, které měly hodnotu *p* pod hladinou významnosti (0,05).

Stepwise (postupně)

Jedná se o kombinaci předešlých dvou technik. Zahrnovala přidávání a odstraňování charakteristik z výsledné karty v každém kroku, dokud nebyla dosažena nejlepší kombinace. Použita byla vždy v kombinaci forward stepwise nebo backward stepwise.

3.4.11 Cross-Validation (validace modelu na testovacích datech)

Po sestavení modelu byla otestována přesnost predikce na nových datech – testovacím vzorku, který byl připraven v kroku dělení dat. Model byl testován na datech jemu neznámých, ale u kterých byl znám výsledek, zda klient splatil, či nesplatil. Tato metoda se nazývá Cross-Validace (Tufféry, 2011).

Dále bylo důležité správné nastavení hodnotu cut-off. Je to hodnota oddělující pravděpodobnosti predikující splacení úvěru od nesplacení a banky si ji určují samy. Je odrazem strategie a míry rizika, kterou je instituce ochotná podstoupit. V ideálním případě by se mělo jednat o takovou hodnotu, při které dochází k maximální správné predikci výsledků a zároveň minimalizaci rizika a ušlých obchodních příležitostí. Vztah by se dal popsat takto:

Confusion Matrix		Predicted	
		Good (0)	Bad (1)
Real	Good (0)	True Negative	False Positive
	Bad (1)	False Negative	True Positive

5) Tabulka – Confusion Matrix. Zdroj: Siddiqi, 2017, s. 248.

True negative – klienti co byli správně označeni jako správně splácející.

False positive – klienti, kteří byli označení jako špatně splácející, ale v reálném případě spláceli dobře. Tato skupina prezentuje ušlou obchodní příležitost. Prakticky se často tito klienti dostanou na individuální schválení, kde může dále banka klientům úvěr poskytnout. Popřípadě poskytnout klientovi úvěr za vyšší úrok, aby byl úrok ekvivalentní s mírou rizika.

False negative – nesprávně odhalené špatné účty a jejich následná akceptace. U těchto klientů je důležitá průměrná doba, kdy ještě splácí dle smluvních podmínek. Jedná se o návratnost poskytnutých prostředků. Pokud by se i po odepsání části prostředků, které se bance nevrátí, úvěr i tak vyplatil poskytnout, může přistupovat banka k riziku shovívavěji. Dále záleží na postupu a úspěšnosti vymáhání prostředků zpět.

True Positive – klienti, kteří byli správně označeni jako špatně splácející (Siddiqi, 2017).

3.4.12 Satisfactory result, nastavení cut-off hladiny (kalibrace)

Na základě výsledků Cross-Validace bylo nutné správně kalibrovat hodnotu cut-off hladiny. Pokud byla cut-off hladina nastavena například na 0,85, znamenalo to, že klientům, kterým vyšla hodnota stejná nebo vyšší, byl úvěr poskytnut. Pokud byla predikována nižší pravděpodobnost, úvěr jim poskytnut nebyl.

3.4.13 Scorekarta

Scorekarta je výsledný seznam signifikantních proměnných a jejich jednotlivých koeficientů. U rozbinovaných proměnných získal každý bin samostatnou hodnotu (Tufféry, 2017, s. 564 – 567)

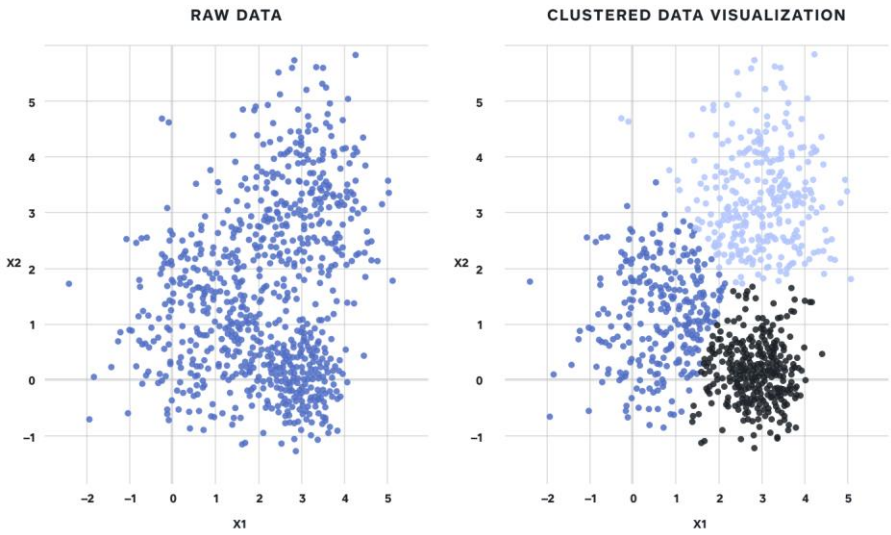
3.4.14 Final Score

Součet výsledných hodnot skóre, který klient získal po kompletní analýze jednotlivých modelů.

3.4.15 Cluster analysis (shluková analýza)

Cluster analysis, též shluková analýza, je vícerozměrná statistická metoda, která se používá ke klasifikaci objektů. Slouží například k třídění klientů do jednotlivých skupin tak, aby si klienti náležící do stejné skupiny byli podobnější s objekty z ostatních skupin. Každou skupinu lze charakterizovat prostřednictvím určitého souboru znaků.

V případě credit scoringu jde o zařazení klientů do určité skupiny dle míry rizika. Banky většinou klienty řadí do čtyř až pěti skupin od *Very Low* skupiny, kde jsou zařazení klienti s velice malou predikcí splacení úvěru, až po skupinu *Excellent*, kde jsou zařazení klienti s výborným chováním a malou pravděpodobností nesplacení.



6) Obrázek – příklad Clusterové analýzy, zdroj: https://medium.com

3.4.16 Limit of the Loan

Po kompletní analýze dochází k výpočtu limitu úvěru, který lze klientovi poskytnout. V praxi se každé skupině (clusteru klientů) přidělí maximální limit, který lze této skupině poskytnout. Postupně se testují částky od nejvyšší až po nejnižší. V každém kroku se částka snižuje o 10 000 Kč do té doby, dokud nevyjde částka finální.

4 Teoretická východiska

4.1 Dlužník a věřitel

Dlužník je fyzická nebo právnická osoba, která má dluh vůči věřiteli.

Věřitel je fyzická nebo právnická osoba, která má pohledávku za dlužníkem. Většinou si klade podmínky, za jakých je ochotna „věřit“, že daný dluh dlužník splatí za sjednaných podmínek. (Zákon č. 257/2016 Sb. o spotřebitelském úvěru)

4.2 Debt Service To Income (DSTI), Debt To Income (DTI) a Minimální disponibilní příjem (DISPO)

Debt Service To Income (dále jen DSTI) v doslovném překladu znamená „dluhové služby k příjmu“. Využívá se pro vyjádření procentuálního poměru všech splátek úvěrů (započítávají se i kreditní karty, kontokorenty, atp.) k čistému měsíčnímu příjmu. Pokud DSTI přesáhne 60 %, banky většinou další žádost o úvěr rovnou zamítají nebo jde žádost na individuální schválení (www.banky.cz).

Debt To Income (dále jen DTI) v překladu doslova znamená „dluh k příjmu“. Jedná se o poměr celkové dlužné částky všech úvěrů k ročnímu čistému příjmu klienta. Nová regulace vydaná Českou národní bankou udává, že celkový poskytnutý limit úvěrů by neměl přesáhnout devítinásobek čistého ročního příjmu (www.mesec.cz).

Minimální disponibilní zdroje (DISPO) vyjadřují volné finanční zdroje klienta, které mu zbydou (z čistého měsíčního příjmu) po zaplacení všech nutných nákladů, které daný měsíc má (používá se také výraz existenční minimum). Po odečtení této částky zůstanou klientovi volné finanční prostředky, se kterými banka může počítat jako s možnou maximální výší měsíční splátky úvěru (www.hypotecnibanka.cz).

Klient musí podmínky DTI, DSTI a DISPO splňovat vždy najednou.

4.3 Úvěr

Úvěr je dočasné zapůjčení finančních prostředků věřitele dlužníkovi za účelem zisku věřitele. Ziskem je v tomto případě myšlený úrok a poplatky, které je dlužník ochoten a schopen zaplatit. Dlužník se zavazuje splácet věřiteli po částech poskytnuté prostředky spolu s úrokem, dokud nedojde k úplnému umoření dluhu.

Úvěry v současnosti poskytují banky a různé nebankovní společnosti, což je jejich hlavní činností. Dne 1.12.2016 vešel v platnost nový zákon č. 257/2016 Sb., o spotřebitelském úvěru. Podstatnou částí zákona bylo omezení počtu nebankovních subjektů poskytujících úvěry na trhu. Počet se omezil pouze na takové poskytovatele, kteří získají oprávnění od České národní banky a splní zákonné podmínky.

Hlavní položkou aktiv bank je poskytování úvěrů. Díky nim přerozdělují volné prostředky a regulují množství peněz, které jsou v oběhu (Bessis, 2015)

4.4 Druhy úvěrů

Pro tvorbu credit scoringového modelu postačí dělení úvěrů podle: **zdroje, dlužníka, doby splatnosti, způsobu zajištění a účelu.**

4.4.1 Rozdělení podle zdroje

Podle zdroje lze úvěry definovat na bankovní a nebankovní. Bankovní úvěry poskytují bankovní instituce akreditované Českou národní bankou, která má také nad poskytnutými úvěry dohled. Klient podléhá velice složitému schvalovacímu procesu, ve kterém se zkoumá jeho bonita a solventnost.

Oproti tomu nebankovní úvěry bývají snáze dostupné, někdy i bez nutnosti dokládání nebo prokazování příjmu, většinou se však poskytují v nižších částkách. Žadatel prochází výrazně jednodušším schvalovacím procesem. Úvěry však bývají dražší. Dohled nad nebankovními úvěry provádí Česká obchodní inspekce (www.wikipedia.cz).

4.4.2 Rozdělení podle dlužníka

Podle profilu dlužníka můžeme rozlišovat nejčastější dělení na úvěry pro fyzické osoby, pro fyzické osoby podnikatele (FOP), firemní úvěry a korporátní úvěry (www.csas.cz).

4.4.3 Podle doby splatnosti

Dle doby splatnosti rozlišujeme úvěry krátkodobé (většinou do jednoho roku), střednědobé (1 – 4 roky) a dlouhodobé (4 a více let) (www.wikipedia.cz).

4.4.4 Podle způsobu zajištění

Úvěry zajištěné bývají nejčastěji zajištěny nebo ručeny nemovitostí, a to buď nemovitostí kupovanou, nebo i jinou, má-li dostatečně vysokou odhadní částku kupní ceny. Nezajištěné úvěry, tedy takové, u nichž dlužník nemusí ručit žádnou zástavou, se dále kategorizují na úvěry revolvingové a peněžní.

Zajištěné úvěry jsou například hypotéky, nebo úvěry ze stavebního spoření.

Revolvingové úvěry jsou takové úvěry, které při vyčerpání a splacení nezaniknou a dají se proto opakovaně čerpat. Jsou to například kreditní karty (od cca 2tis až do 500tis) nebo kontokorenty (poskytují se například do výše jednoho měsíčního čistého příjmu).

Revolvingové úvěry patří mezi nejdražší typy úvěrů. Velkou výhodou pro klienta je, že pokud se dostane do finanční tísně, má peníze ihned k dispozici a může je čerpat. Většinou ale klienti žádají o úvěr až ve chvíli, kdy se dostanou do finančních problémů a za takových podmínek banka s největší pravděpodobností úvěr neschválí. Další předností je, že i když klient úvěr nečerpá nebo čerpá a řádně splácí, zaznamenává banka do bankovního registru dobré chování klienta, které mu pak může pomoci při žádosti o další úvěr. (Cipra, 2015)

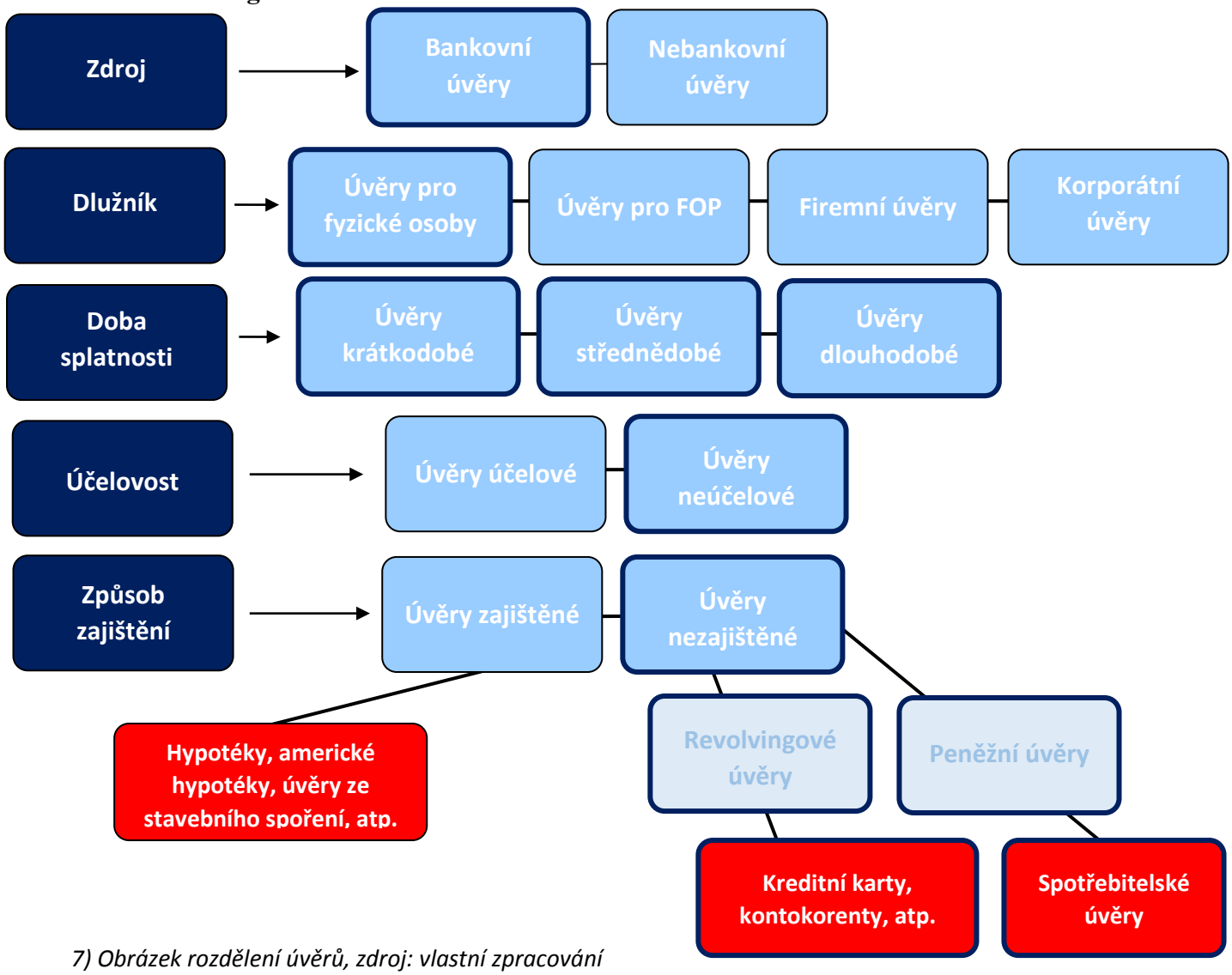
Nezajištěné úvěry peněžního typu jsou spotřebitelské úvěry, které jsou určené pro koncového klienta – spotřebitele a fyzickou osobu nepodnikatele – za účelem nákupu spotřebního zboží, auta, apod. Od 1. prosince 2016 upravuje problematiku spotřebitelských úvěrů zákon č. 257/2016 Sb., o spotřebitelském úvěru.

4.4.5 Dělení dle účelovosti

Poslední dělení je na úvěry účelové a neúčelové. Účelové úvěry bývají použity například na koupi auta, vybavení bytu, atp. V takovém případě chce banka doložení účelovosti. Úvěry neúčelové lze použít téměř na cokoliv (www.wikipedia.cz).

V této práci bude popsán model, který slouží pro vyhodnocování kredibility klienta u úvěrů pro fyzické osoby, v domácí měně (v českých korunách), bankovních, neúčelových, nezajištěných, krátkodobých, střednědobých, dlouhodobých a úvěrů bez ručitele. Jedná se o vyhodnocení u těchto typů úvěrů: kreditní karty, kontokorenty, úvěry typu „na klik“, povolená přečerpání účtu, spotřebitelské úvěry a jedna výjimka - úvěry na bydlení nezajištěné s maximálním limitem do 1 000 000 Kč.

Diagram rozdělení úvěrů:



7) Obrázek rozdělení úvěrů, zdroj: vlastní zpracování

4.5 Credit Scoring

Siddiqi ve své knize *Intelligent Credit Scoring* pojednává o tom, že credit scoring je složitý a sofistikovaný proces hodnocení bonity a rizikovosti klienta pomocí skórování úvěrové žádosti. Jedná se o soubor statistických a matematických metod, technik a rozhodovacích modelů a zároveň také osobního vyhodnocení, které slouží ke správnému a bezpečnému úvěrování klientů a rozpoznání kreditního a úvěrového rizika. Při tvorbě modelu jde o výběr signifikantních proměnných, které mají na rozhodnutí zásadní vliv, a odstranění jejich duplicitních informací. Výsledek se většinou odvíjí od pravděpodobnosti, zda bude dlužník schopen dostát svých závazků (*probability of default*) a odhaduje se na základě historických dat. Naopak také slouží k ohodnocení bonity klienta (čím vyšší je bonita, tím méně rizikový se může klient jevit pro případného věřitele).

Cílem credit scoringového modelu je předpovědět s co možná největší přesností statistickou šanci (*odds ratio)* nebo pravděpodobnost (*probability*), s jakou bude daný klient schopný dostát svým závazkům v budoucnosti. Jinak řečeno, zda daný klient úvěr, o který žádá, splatí či nesplatí. (Mejstřík, 2014)

Výsledkem credit scoringového modelu bývá nejčastěji konkrétní procentuální pravděpodobnost splacení úvěru nebo přepočet na číselnou hodnotu, tzv. skóre, podle kterého se daná instituce rozhodne klientovi půjčit a za jakých podmínek. Dle výsledného skóre se bude banka rozhodovat, do jaké skupiny klienta zařadí. Na základě zařazení klienta prostředky poskytne v plné výši, či jen část, nebo žádost zcela zamítne. Následně pak záleží na tom, jak velké riziko je banka schopna podstoupit, tzn. vhodné zvolení hranice *cut-off hladiny*, kdy už bude žádost zamítnuta s konečným rozhodnutím. Přiměřenou hranici si banky určují samy na základě optimální procentuální průchodnosti žádostí. Zjednodušeně jde tedy o stanovení co možná nejlepšího modelu, který bude schopen predikovat spolehlivé výsledky na základě analýzy historických dat a pokusit se tak o co nejlepší odhad chování žadatelů o úvěr a zároveň o maximalizaci zisku při současné minimalizaci a řízení rizika a ztrát. Model často banky využívají i při výpočtu pojištění k úvěru (Siddiqi, 2017).

Model je postaven na využití statistických metod, jako jsou například Weight of Evidence, Information Value, Gini index, Kolmogorov - Smirnov index, Somers'D index, ROC křivky, logistická regrese nebo Cross-Validace. Celý proces nezahrnuje pouze vytvoření a postupné aplikování modelu, ale také ohodnocení a následné monitorování celého schématu. Jakákoliv legislativní, populační nebo regulační změna musí být zohledněna, tudíž následná kontrola zahrnuje testování optimálního modelu při využití nových dat za účelem zlepšení celého procesu.

Důležité je také zvolení takového postupu, který bude následně jednoduše interpretovatelný. Pro snadnější interpretaci se používá skórovací karta – *scorekarta*, kde jsou uvedeny statisticky významné proměnné, které model využívá. Každá proměnná má určitou váhu, která byla zvolena na základě vypovídajícího charakteru a míry závislosti na vysvětlované proměnné. (Siddiqi, 2017)

Díky přispění Basilejského výboru pro bankovní dohled, který napomohl utvořit standardizovaný hodnotící postup, používají tento typ složitého úvěrového procesu téměř všechny instituce, které úvěry poskytují. Standardy vydávané tímto výborem se nazývají v čase Basel I, Basel II a Basel III (též zvané jako Báze). V současnosti platí Basel III a jeho rozšířená forma. (Cipra, 2015)

4.6 Bankovní data

Banky využívají obrovské množství vstupních informací. Dá se říct, že do modelu použijí všechna dostupná data, která o klientovi mají nebo zjistí (z různých databází, dotazováním klienta, atp.). Rozhodují se na základě dat nasbíraných v minulosti a hledají podobnost s určitým typem klientů. Výsledné skóre vzniká na základě *scorekarty (skórovací karty),* která nejlépe předpovídá a odděluje dobré (goods) a špatné (bads) úvěrové účty. Dobrými úvěrovými účty jsou myšleny úvěry, které byly splaceny za smluvních podmínek – bez defaultu. Špatné úvěrové účty jsou naopak účty vykazující nesplacené závazky. (Siddiqi, 2017)

Ukázka Scorekarty:

Characteristics	Attribute	Scorecard Point
Age	A:0-30	49
Age	A:30-40	51
Age	A:40-50	53
Age	A:50-60	57
Age	A:60-70	65
Age	A:70-80	70
Age	A:80-90	72
Age	A:90-130	71
DebtRatio	DR:0-0.2	57
DebtRatio	DR:0.2-0.4	59
DebtRatio	DR:0.4-0.6	55
DebtRatio	DR:0.6-0.8	52
DebtRatio	DR:0.8-1.0	50
DebtRatio	DR:1.0-1.2	49
DebtRatio	DR:1.2-1.4	48
DebtRatio	DR:1.4-1.6	53
MonthlyIncome	MI:0-2000	53
MonthlyIncome	MI:2000-4000	52
MonthlyIncome	MI:4000-6000	57
MonthlyIncome	MI:6000-8000	58
MonthlyIncome	MI:8000-10000	60
MonthlyIncome	MI:10000-12000	63
MonthlyIncome	MI:12000-14000	63
MonthlyIncome	MI:14000-16000	61
MonthlyIncome	MI:16000+	61

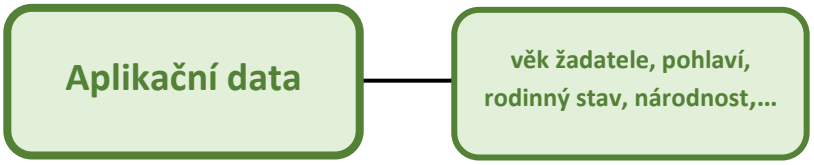
8) Obrázek ukázka scorekarty. Zdroj obrázku: https://weclouddata.com

4.7 Rozdělení bankovních dat

Jsou tři druhy dat, které banky sbírají. Ke každé skupině dat přísluší jeden scoringový model. Výsledkem testování jsou číselné hodnoty (skóre). Ty se na konci sčítají. Dle výsledku se banky nakonec rozhodnou, zda klientům úvěr půjčí, či nikoliv. Jedná se o tyto datasety:

9) Obrázek rozdělení bankovních dat. Zdroj informací: Siddiqi

4.7.1 Aplikační data



10) Obrázek příklad aplikačních dat. Zdroj informací: Siddiqi

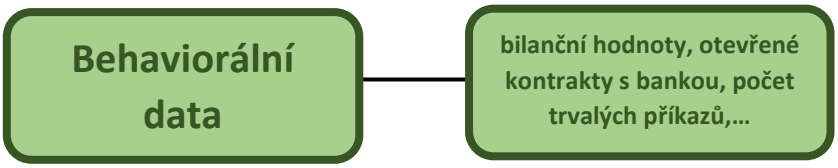
Data, se kterými se klient setká osobně a sám sděluje při žádosti o úvěr, vstupují do aplikačního skóre.

Obsahuje zejména demografické proměnné: věk žadatele, pohlaví, rodinný stav, národnost, způsob bydlení, počet odpracovaných měsíců v současném zaměstnání, doba pobytu na kontaktní adrese, příjem, obor zaměstnavatele, počet členů domácnosti bez příjmu, atp. (Siddiqi, 2017).

Data zadávají do systému klientští pracovníci nebo si je klient vyplňuje sám v rámci internetového bankovnictví nebo online žádosti o úvěr. Zde je velké riziko, že se můžou objevit fraudy popř. zkreslené informace, které klient nesdělí dle skutečnosti, popř. klientský pracovník zadá špatně. Například v dnešní době už nemusí mít rodiče uvedené děti v občanském průkazu, tudíž klient může uvést počet dětí úmyslně špatně (počet dětí zvyšuje měsíční minimální existenční náklady, se kterými musí banka počítat).

Dále je zde velké riziko výskytu fraudu při úmyslném zkreslování informací ze strany klientského pracovníka za účelem poskytnutí úvěru klientovi za každou cenu. Důvodem je plnění KPI kritérií (z anglického Key Performance Indicator, což v překladu znamená klíčový ukazatel výkonnosti), které často obsahují plnění objemu poskytnutých úvěrů. Pokud by se jednalo o odlehlé hodnoty (např. namísto příjmu třiceti tisíc korun by bylo zadáno tři sta tisíc korun), dají se detekovat. Z tohoto důvodu je u aplikačního skóre velice důležité čištění dat.

4.7.2 Behaviorální data



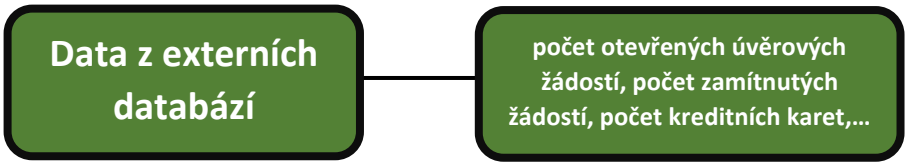
11) Obrázek příklad behaviorálních dat. Zdroj informací: Siddiqi

Nejdůležitější a nejcennější data pro banku jsou data sbíraná za účelem výpočtu behaviorálního skóre. Klient se s nimi nesetká přímo, protože se sbírají na pozadí. Jedná se o sběr cca 300 – 400 typů proměnných, ze kterých se následně skládá model. Vzhledem k tomu, že se jedná o bankovní data, která si banky sbírají samy z vlastních zdrojů, nemusí za jejich získání platit jiným dodavatelům.

Jedná se především o sběr proměnných definujících existující vztah s bankou: jak dlouho má klient vedený účet v bance, platební styk, příjem na účtu (poslední kvartál až čtyři kvartály), bilanční hodnoty, počet otevřených kontraktů s bankou, počet trvalých příkazů, chování na účtu u kreditní karty, atp. (Siddiqi, 2017).

Na základě behaviorálních dat lze podstoupit s klientem zjednodušený úvěrový proces schvalování, který je velice rychlý. Klient většinou nemusí dokladovat příjmy a jiné dokumenty. Banky na základě těchto dat přepočítávají také předschválené limity na účtech. Jsou to takové limity, které můžou být poskytnuty žadateli bez nutnosti prokazování bonity.

4.7.3 Data z externích databází



12) Obrázek příklad dat z externích databází. Zdroj informací: Siddiqi

Jedná se o tzv. kreditní data: počet otevřených úvěrových žádostí od všech společností, počet zamítnutých žádostí, počet kreditních karet, otevřené kontrakty, delikvence, jiné veřejné záznamy, atp. (Siddiqi, 2017)

V bankovním sektoru se využívá hned několik externích databází. Tři nejčastější databáze jsou Bankovní Registr Klientských Informací, Nebankovní Registr Klientských Informací a SOLUS.

4.7.3.1 Bankovní registr klientských informací (dále jen BRKI)

Nejvýznamnější a nejčastěji využívaná externí databáze u retailového klienta je BRKI. Též nazývána jako CBCB neboli Czech Banking Credit Bureau. Jedná se o databázi údajů o smluvních úvěrových vztazích mezi bankami a jejich klienty. Jsou zde uvedeny údaje o klientech, jejich celkové úvěrové angažovanosti a také o čerpání jednotlivých úvěrových produktů a o platební morálce.

Jak již bylo zmíněno, v této databázi se nachází nejen negativní, ale i pozitivní informace o klientech. Databáze umožňuje bankám důkladně prověřit úvěrovou historii klienta nejen na produktech, které byly uzavřeny uvnitř banky, ale také v jiné finanční instituci.

Klientům databáze umožňuje budování dobré úvěrové historie a důvěryhodnosti. V BRKI jsou také vedeni klienti, kteří o úvěrový produkt teprve žádají nebo klienti, s nimiž banka požadovaný kontrakt neuzavřela. Je možné dohledat i počet těchto zamítnutých žádostí. Eviduje se pouze po dobu jednoho roku od podání žádosti. Pak je tato informace o zamítnutí smazána. Banky své zamítnuté žádosti evidují déle. Data o uzavřených smlouvách jsou evidována po celou dobu existence úvěrového vztahu a dále po dobu čtyř let po jeho ukončení. Finanční instituce většinou berou v potaz data stará dva roky.

Banky do této databáze můžou nahlédnout na základě podepsané žádosti klientem. Provozovatelem tohoto registru je společnost CBCB - Czech Banking Credit Bureau, a.s., která je vlastněna pěti zakládajícími bankami. Jsou to Česká spořitelna, a.s., Československá obchodní banka, a.s., Komerční banka, a.s., Moneta Money Bank, a.s. a Unicredit Bank Czech Republic and Slovakia, a.s. (www.cbcb.cz).

4.7.3.2 Nebankovní Registr Klientských Informací (dále jen NRKI)

Další databáze je NRKI. Též nazývána jako CNCB neboli Czech Non-banking Credit Bureau. Tato databáze funguje velice podobně jako BRKI. Shromažďuje pouze jinak zaměřená klientská data. Jedná se o sběr informací od věřitelských subjektů a zprostředkování informací leasingových a úvěrových společností. Konkrétně jde o údaje vypovídajících o bonitě, důvěryhodnosti a platební morálce klientů. Majiteli sdružení jsou společnosti nebankovní, působící na trhu splátkového prodeje a leasingu. Jako příklad lze uvést: ČSOB Leasing, a.s., Providen Financial s.r.o. nebo Mercedes-Benz Financial Services Česká republika, s.r.o. (www.cncb.cz).

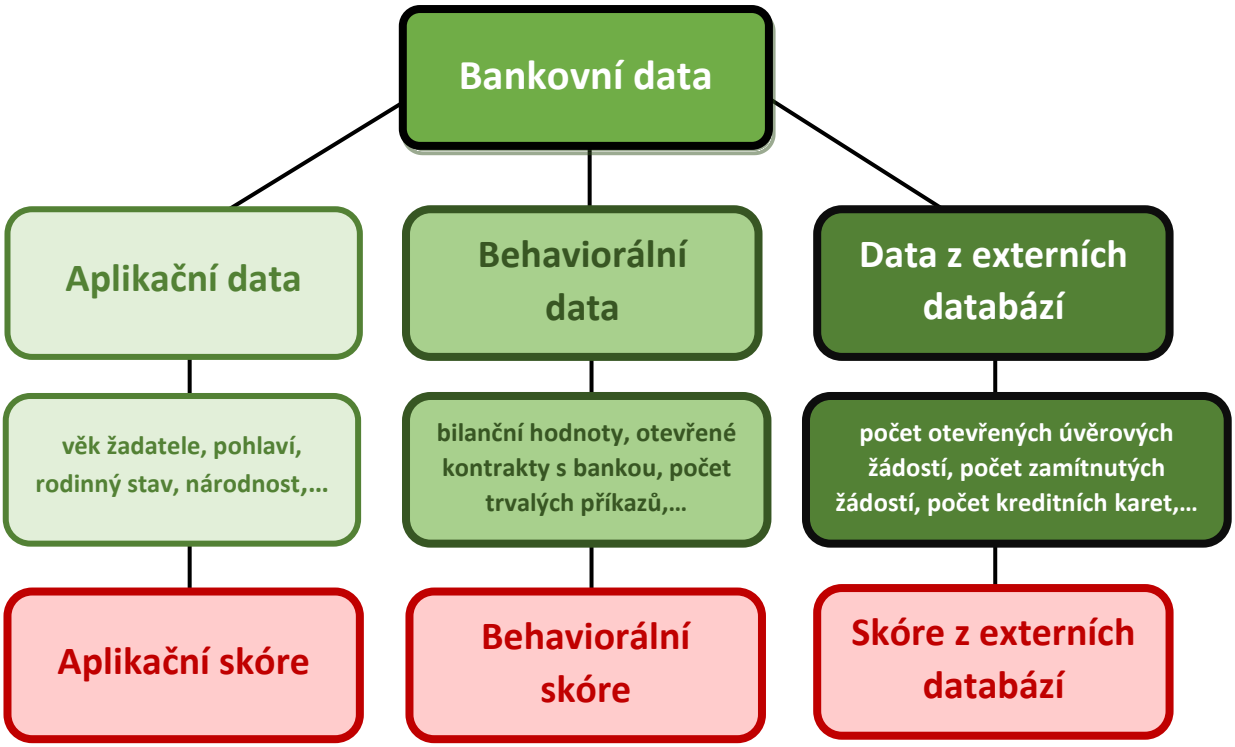
Velkou slabinou těchto databází je jejich dostupnost a cena získaných dat. Za každý dotaz, který je zaslán do této databáze, banky platí a získané informace můžou být již zastaralé. Finanční instituce je aktualizují pouze jednou za měsíc.

4.7.3.3 Sdružení na Ochranu Leasingu a Úvěrů Spotřebitelům (SOLUS)

„SOLUS (Sdružení na Ochranu Leasingu a Úvěrů Spotřebitelům) sdružuje řadu společností z různých ekonomických sektorů. Jsou mezi nimi banky a stavební spořitelny, nebankovní finanční instituce, poskytovatelé telekomunikačních služeb, distributoři energií, poskytovatelé P2P půjček a další společnosti z oblasti obchodu a služeb.“ (www.solus.cz).

V této databázi se nachází pouze negativní záznamy. Banky je většinou berou v potaz od dlužné částky cca 1000Kč, kde žádost rovnou zamítají.

Diagram rozdělení bankovních dat:



13) Obrázek rozdělení bankovních dat. Zdroj informací: Siddiqi

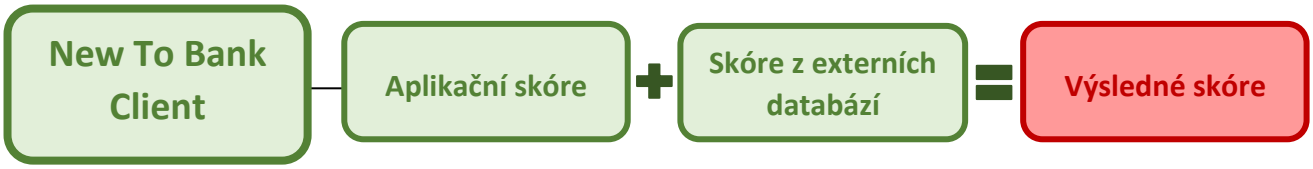
4.8 Rozdělení klientů

14) Obrázek: příklad dělení bankovních klientů, zdroj: vlastní zpracování

Banky rozdělují klienty na tři skupiny. Tzv.: Nový bankovní klient (New To Bank Client), nový klient pro banku i úvěrový trh (New To Market Client) a klient banky (Bank Client). Podle tohoto rozdělení různě žadatel vstupuje do kombinace modelů, aby banky byly schopné učinit nejlepší rozhodnutí a aplikovat co nejvíce dostupných dat.

4.8.1 New To Bank Client (Nový bankovní klient)

Výsledné skóre obsahuje výstup z modelu aplikačního skóre a z dat externích databází.



15) Obrázek součtu skóre NTB klienta, zdroj: vlastní zpracování

New To Bank Client (dále jen NTB) je klient, který u banky neměl nikdy žádný produkt, nebo ho měl, ale po dobu delší než dvanáct měsíců už u banky žádný kontrakt obnovený nemá (www.quora.com).

Od NTB klienta banka nemůže sbírat data o chování na účtech, tudíž se nedá zpracovat jeho behaviorální skóre.

U tohoto typu klienta pracovník klientských služeb shromáždí informace, které se použijí pro výpočet aplikačního skóre. Většinou banka od klienta vyžaduje doložení některých dodatečných dokumentů. Jedná se o doložení příjmu (potvrzení o příjmech nebo daňové přiznání), pracovní smlouvy, dohody o pracovní činnosti, výplatních pásek, doložení určitého počtu výpisů z účtu, atp.

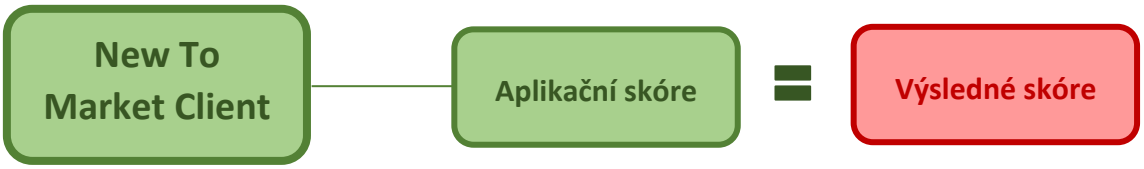
Zároveň také záleží na osobním vyhodnocení pracovníka, zde se klient nechová podezřele, či jsou relevantní a důvěryhodné jeho sdělené informace a doložené doklady. Pokud by bylo něco podezřelého, co by mohlo nasvědčovat úvěrovému podvodu, pracovník může vyhodnotit klienta jako neakceptovatelného a úvěr zamítnout ihned. Dále pak probíhá šetření útvaru Compliance, zda se opravdu jednalo o podvod. Poskytnutí úvěru není nijak právně vymahatelné. Banka se může kdykoliv rozhodnout úvěr neposkytnout.

Vzhledem k narůstajícím požadavkům na poskytnutí úvěru a regulacím ČNB se v poslední době velice často objevují falšovaná potvrzení o příjmech nebo potvrzení o příjmech z fiktivních společností, popř. se klienti nechávají fiktivně zaměstnat na pár měsíců kvůli potřebné výši příjmů. Klient většinou zaplatí všechny náklady s tím spojené a banka plní např. limit hypotéky, který by neměl být klientovi poskytnut.

U NTB klientů se sledují dvě rozhraní. Klient je sice nový pro banku, ale ne nový pro úvěrový trh. Znamená to, že v minulosti již nějaký úvěr měl nebo v současnosti disponuje např. kreditní kartou nebo kontokorentem. První, jak již bylo zmíněno, je aplikační skóre. Další informace jsou z dostupných externích databází. Celkově tedy banka dostává dostatečné informace, které jí poslouží k relevantnímu rozhodnutí, zda klientovi finanční prostředky poskytne.

4.8.2 New to Market Client (Nový klient pro banku i pro úvěrový trh)

Výsledné skóre pouze z modelu aplikačního skóre.



16) Obrázek součtu skóre NTM klienta, zdroj: vlastní zpracování

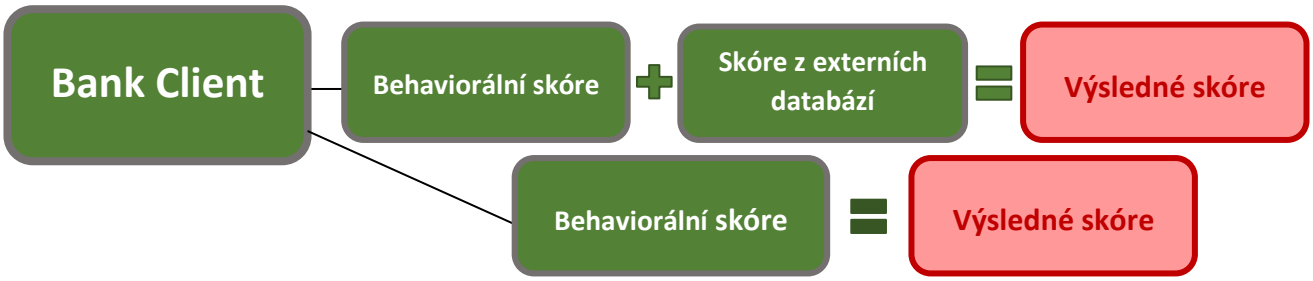
Další kategorií je klient New To Market (dále jen NTM). Jedná se o klienta, u kterého nelze zjistit informace z externích databází, protože v minulosti neměl úvěr. Nelze ani vyhodnotit behaviorální skóre, protože u posuzující banky nemá žádný účet ani produkt (www.segmentationstudyguide.com).

U klienta tohoto typu jsou zjišťována pouze vstupní data do aplikačního skóre. Jedná se většinou o studenty nebo klienty, kteří mají účet u jiné společnosti. V takovém případě se většinou banka rozhodne klientovi poskytnout pouze úvěr s nižším limitem (v řádech jednotek nebo desítek tisíc) nebo neposkytnout úvěr vůbec, popř. jde úvěr na individuální schválení.

Často se takovým klientům doporučuje otevření úvěrového účtu s nižším limitem (např. kreditní karty nebo kontokorentu) pro získání kladné úvěrové historie v databázi BRKI.

4.8.3 Bank Client (Klient banky)

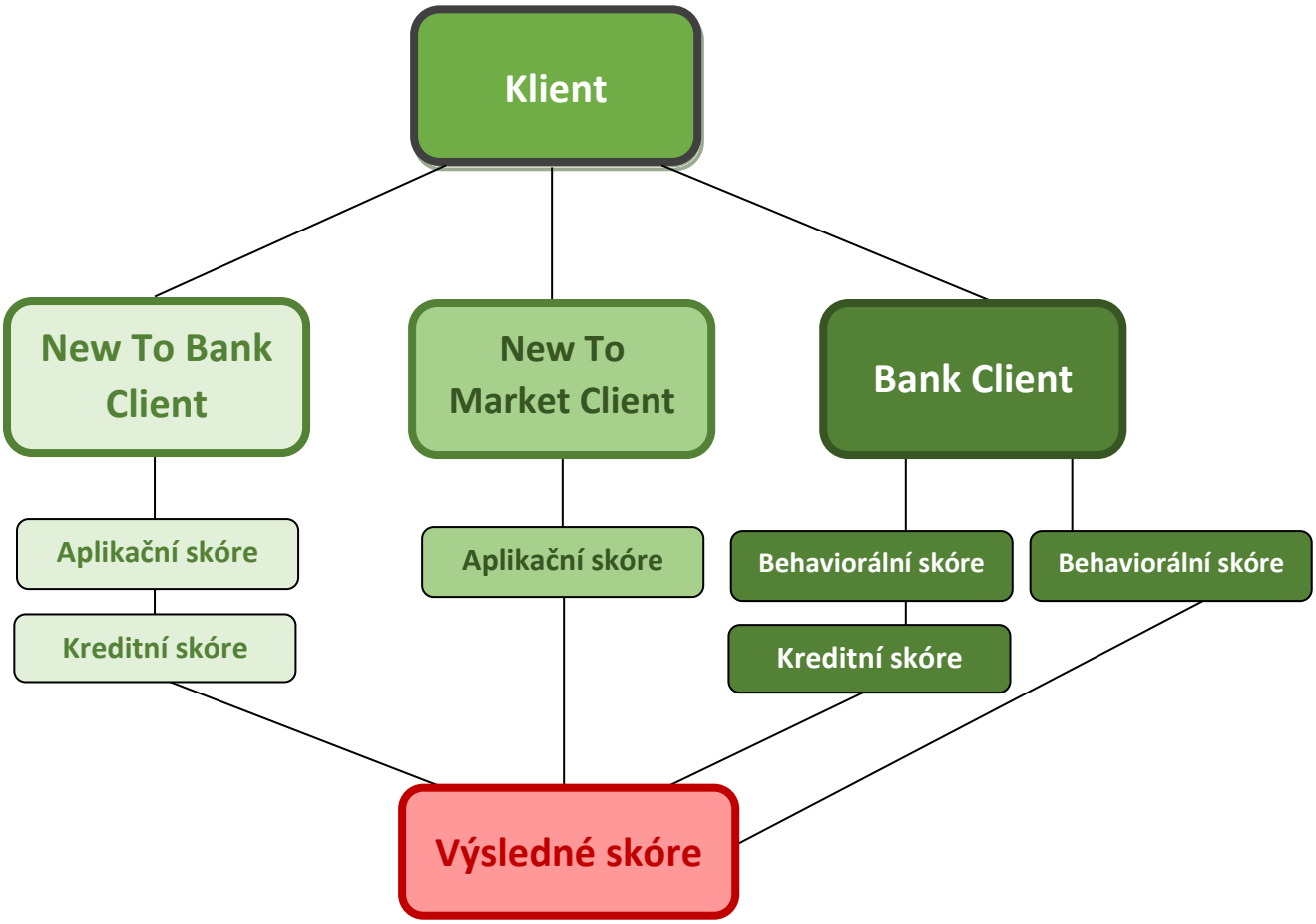
Výsledné skóre z modelu behaviorálního skóre a z modelu externích databází.



17) Obrázek součtu skóre BC klienta, zdroj: vlastní zpracování

Pravděpodobně nejčastější žadatel o úvěr bývá ten, který má u dané instituce otevřený účet (www.segmentationstudyguide.com).

Sběr dat probíhá do modelu behaviorálního skóre. Dále se pak takový klient ještě dělí na klienta s úvěrovou historií a bez úvěrové historie. Podle toho banka sbírá data z externích databází. Behaviorální data jsou dostatečně silná, aby banky pro finální rozhodnutí nemusely data z externích databází sbírat. Aby však nedocházelo např. k přeúvěrování klienta z důvodu přesáhnutí maximální možné celkové úvěrové angažovanosti, banky informace z externích databází zjišťují.



18) Diagram součtů skóre klientů, zdroj: vlastní zpracování

4.9 Úvěrový proces

Celý úvěrový proces má přesně definované kroky, jak postupovat. Pro dobrou orientaci v celé problematice úvěrového schvalování je lepší celou cestu znázornit i graficky.

Proces probíhá v těchto krocích:

1) **Identifikace** klienta pracovníkem banky nebo systémem v internetovém bankovnictví
2) **Klient žádá o úvěr**
3) **Národnost** (Klient s jinou národností, než českou, je posuzován zvlášť)
4) **Akceptace klienta** - osobní ohodnocení klienta pracovníkem banky, zda je akceptovatelný (zda se nejedná o „bílého koně“, popř. se klient nechová podezřele)
5) **Rozdělení klientů** (NTB, NTM, CB)
6) **Sběr dat** dle nutnosti do modelu AS nebo vyhodnocení BS
7) **Dostatečnost skóre z AS nebo BS** (pokud ne, nemá smysl se dále dotazovat do externích databází)
8) **Prověření úvěru** (vhodný typ úvěru)
9) **Dokumentace** (doložení nutné dokumentace k úvěru a její posouzení důvěryhodnosti a pravosti)
10) **DTI, DSTI a DISPO** výpočet (prodloužení doby splatnosti za účelem snížení měsíčních splátek, snížení limitu úvěru, atp)
11) **Externí databáze a podpis žádosti** (klient dává souhlas k nahlédnutí do externích databází a podepisuje žádost)
12) **Sběr dat z externích databází** (dostatečnost skóre z externích databází)

13) **Součet skóre** (z použitých modelů)

14) **Schválení klienta** (nebo zamítnutí)

15) **Výpočet limitu úvěru** možného k poskytnutí

16) **Schválení úvěru**

17) **Souhlas klienta s podmínkami** úvěru, podpis úvěrové dokumentace

18) **Poskytnutí** úvěru

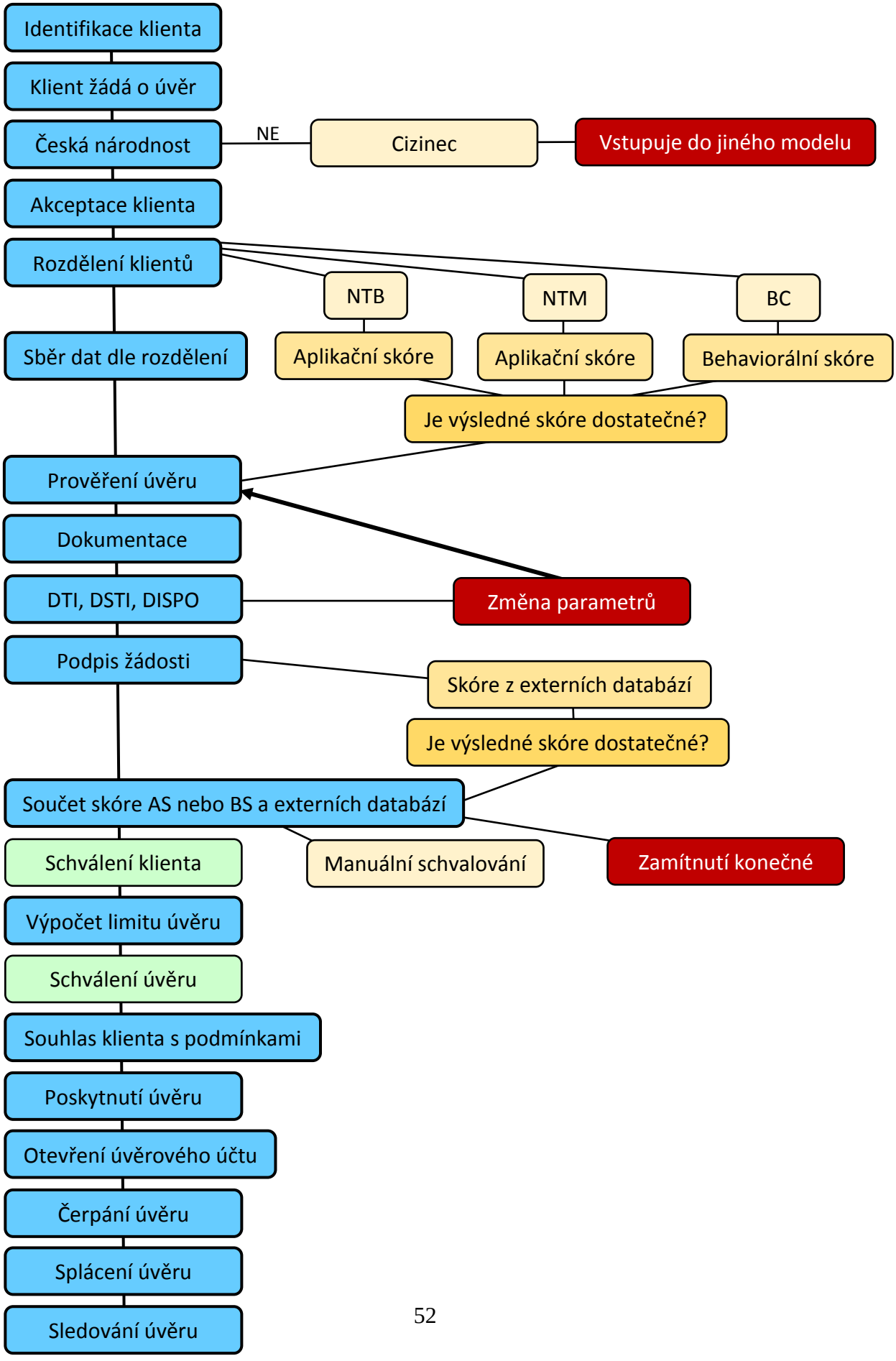
19) **Otevření úvěrového účtu**

20) **Čerpání** úvěru

21) **Splácení**

22) **Sledování** úvěru

4.9.1 Diagram úvěrového procesu (zdroj: vlastní zpracování)



5 Vlastní práce

V následující kapitole bude podrobně rozebrán a popsán celý proces tvorby credit scoringového modelu a prakticky demonstrovány statistické metody a dataminingové techniky na třech sadách reálných dat. Vyhotoveny budou tři modely - aplikační, kreditní a behaviorální. Kompletní postup bude z důvodu nadměrné rozsáhlosti problematiky podrobně demonstrován pouze na aplikačních datech. V případě dat kreditních a behaviorálních bude kladen důraz již jen na nejzásadnější kroky daného postupu. Cílem je nalezení nejvhodnějšího modelu, který bude predikovat spolehlivé výsledky, pro jednotlivé sady reálných dat.

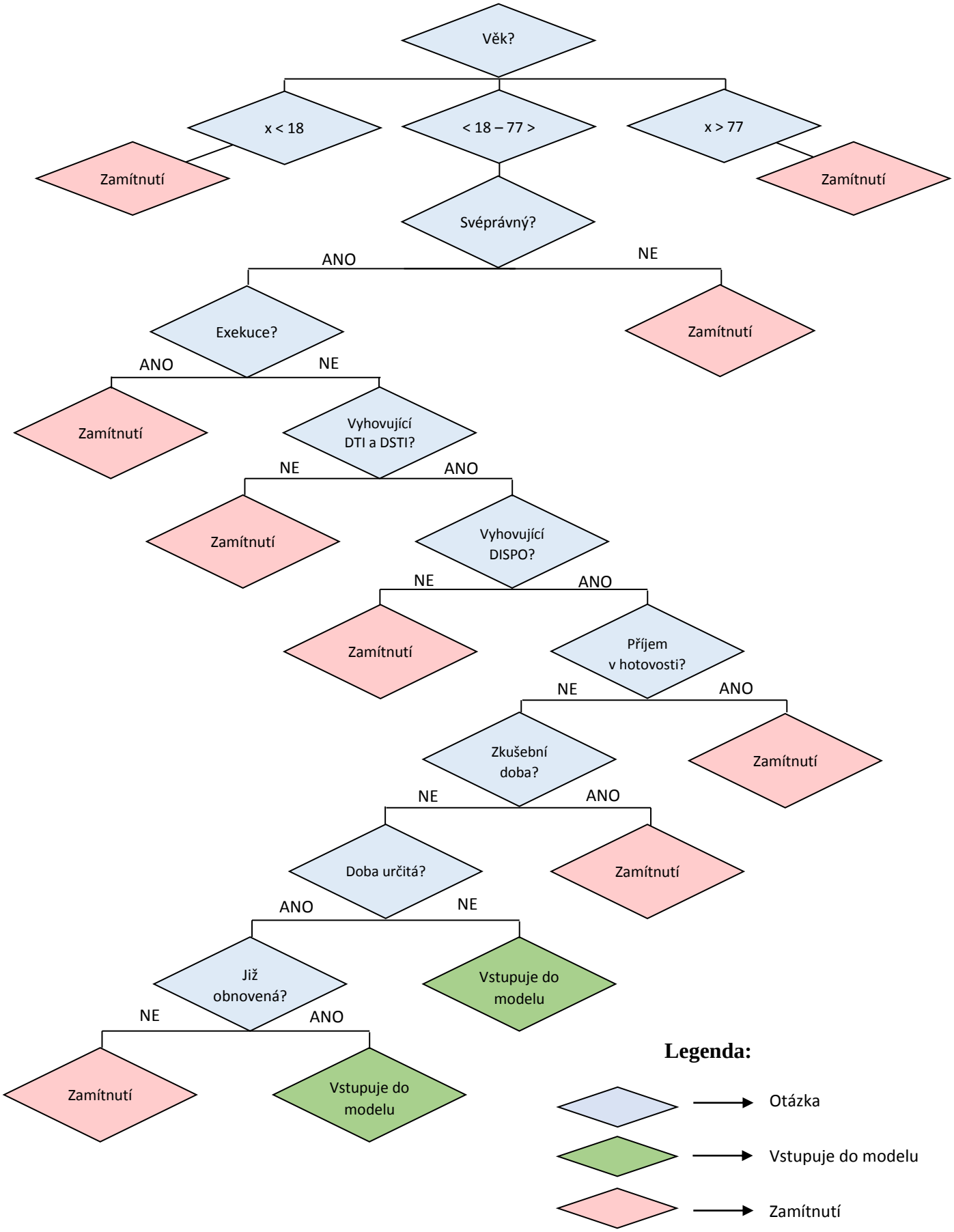
5.1 Preselekce klientů pomocí metody Decision Tree

Před vstupem klienta do jednotlivých modelů prochází klient rozhodovacím stromem. Zde dochází k rozhodnutí bankovního pracovníka, zda je klient vhodný k úvěrování.

Klient, který projde celým tímto modelem, bude ve věku od 18 do 77, svéprávný, bez exekuce, bude mít vyhovující DTI, DSTI i DISPO, příjem mu chodí na účet, není ve zkušební době, bude pracovat na dobu neurčitou nebo mít v případě doby určité minimálně jednou obnovenou pracovní smlouvu. Pokud by klient neprošel přes uzel DTI a DSTI, banka má možnost u takového klienta využít individuální manuální schválení žádosti, pokud dává žádost v celém kontextu klienta smysl.

Následující diagram znázorňuje přesný chod celým rozhodovacím stromem:

5.1.1 Diagram Decision Tree (zdroj: vlastní zpracování)



5.2 Aplikační model

Celý dataminingový proces postupuje v následujících krocích:

Práce s daty

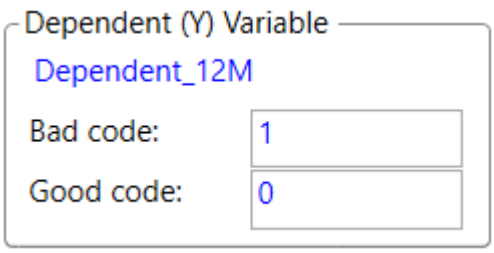
Cílem práce s daty bude nalezení signifikantních proměnných, které budou následně vstupovat do modelu. Celý postup vychází z diagramu na straně 17.

5.2.1 Data understanding

Datový soubor aplikačního modelu čítá data cca 15 000 klientů. Sbírá se přibližně 50 proměnných, jako například *Income_M (měsíční příjem), InsurancePay (platba pojištění), HouseMembers (počet členů v domácnosti), OtherIncome (ostatní příjmy)* nebo *RegionCont (kraj kontaktní adresy).*

V celém souboru se vyskytují proměnné *spojité (continuous)* a *kategoriální (categorical)*. Dále je důležité určit vysvětlovanou proměnnou *y*, což v tomto případě bude proměnná *Dependent_12M*. Jedná se o vektor binárních proměnných (tj. nabývající hodnot 0 nebo 1). Dobré účty (Goods) získávají hodnoty 0 a špatné (Bads) hodnoty 1. Ostatní proměnné x_i jsou vysvětlující. Poměr dobrých a špatných účtů je cca 14 200 Goods : 800 Bads.

Vzhledem k ručnímu zadávání dat do počítače musí proběhnout čištění dat.



19) Obrázek – rozdělení proměnné y, zdroj: vlastní zpracování

5.2.2 Eliminace irelevantních proměnných na základě expertního odhadu

Proměnné typu *NoRequest (číslo žádosti), RequestDate (datum otevření žádosti)* nebo *IDDoc (ID dokumentu)* nemají dále žádný vypovídající charakter a jejich obsah je proto irelevantní. Proměnné tohoto typu budou ze soboru vyjmuty.

5.2.3 Splitting data

Z důvodu nutné finální validace modelu pomocí techniky Cross-Validace bylo nutné soubor rozdělit na dva vzorky – testovací a trénovací.

Na trénovacím vzorku je postaven model a na testovacím vyzkoušena přesnost a prediktabilita modelu. Celý datový soubor bude nejdřív připraven v programu MS Excel.

Existují dvě cesty, jak soubor dat rozdělit. První cesta je randomizace pomocí náhodně generovaných čísel z intervalu. Vzhledem k datování žádostí je lepší využít cestu druhou, což je rozdělení dle data žádosti. Do trénovacího vzorku jsou zařazena starší data, která budou predikovat chování klientů v budoucnu. Po vytvoření se model otestuje na novějších datech – testovacím vzorku.

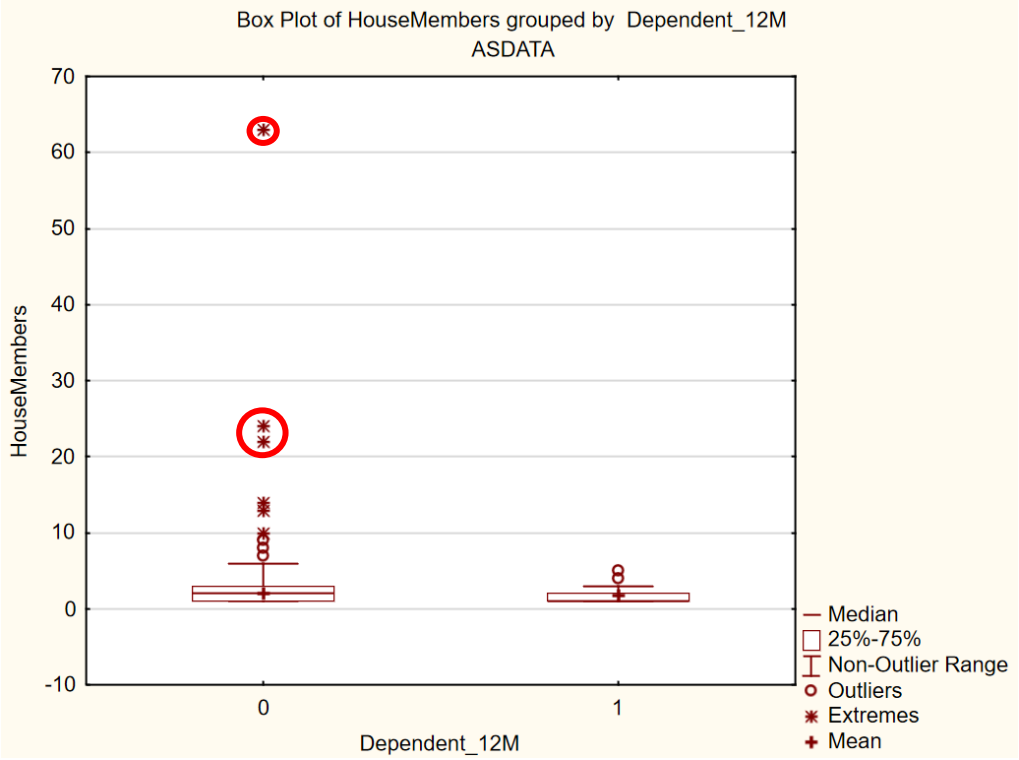
Soubor byl tedy rozdělen na 11 000 klientů (trénovací skupina) a cca 4 000 klientů (testovací skupina).

5.2.4 Statistic software

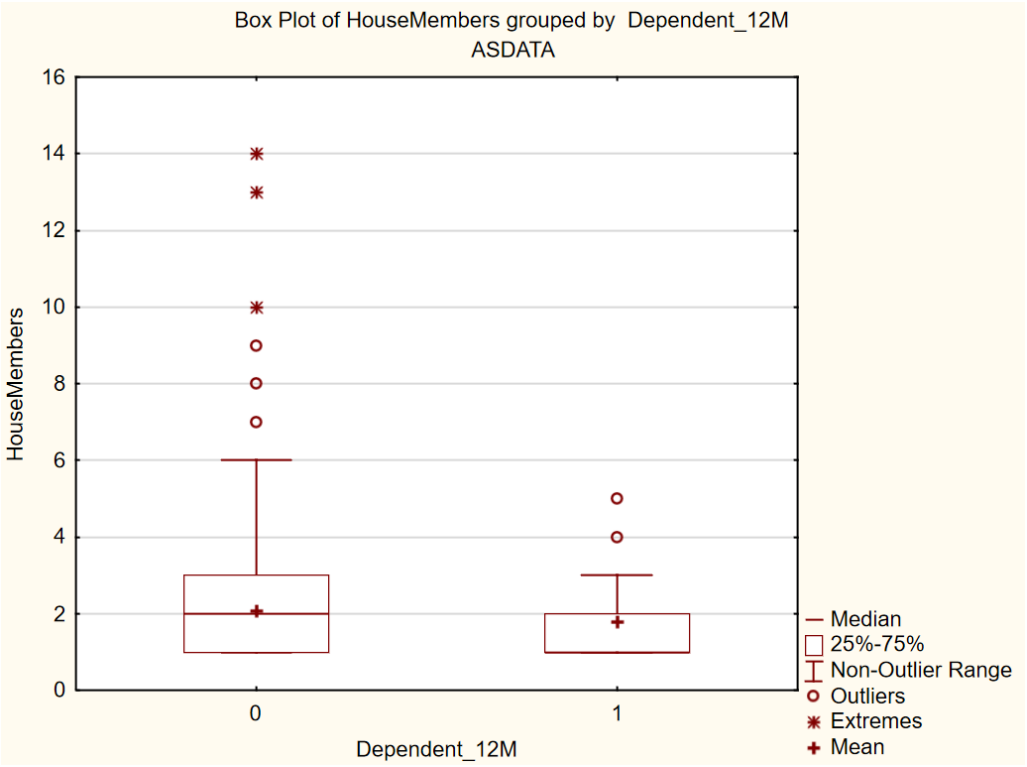
MS Excel nemá vhodné komponenty pro výpočet složitějších statistických metod, jako je například Weight of Evidence nebo logistická regrese. Naopak je vhodný pro jednodušší obsluhu, zpracování a přípravu dat. Po přípravě dat je důležité převést soubor do vhodného software, který má složitější dataminingové metody zakomponované. Pro tuto práci byl využit software STATISTICA13 a software R.

5.2.5 Data cleaning

Jako příklad proměnné, na které se dá čištění dat demonstrovat, byla zvolena proměnná *HouseMembers (počet členů v domácnosti)*. Zde se objevily tři odlehlé hodnoty (63, 24 a 22 členů v domácnosti), které by při ponechání v datovém souboru zkreslily celý model. Při čištění dat je důležitý pohled na celého klienta v kontextu všech ostatních proměnných. Pokud je celkový počet členů v domácnosti 5 a počet závislých osob bez příjmu v domácnosti 63, jedná se o chybu zápisu. Jsou dvě možnosti, jak se vypořádat s chybou tohoto druhu. Datový řádek kompletně vyjmout ze souboru nebo ponechat a predikovat chybu na logickou hodnotu, která vyplývá z kontextu. V případě velkého datového souboru je možné celý datový řádek odstranit. Pokud by se jednalo o malý vzorek nebo například data, jejichž sběr provází finanční náročnost, datový řádek se ponechá a změní. V souvislosti s obsáhlým datovým vzorkem byl každý takový řádek odstraněn. Pokud by se ovšem jednalo o řádek, který nabývá v proměnné *Dependent_12M* hodnoty 1, data klienta budou pozměněna vzhledem k poměrně nízkému počtu *Bads* účtů v celém souboru.



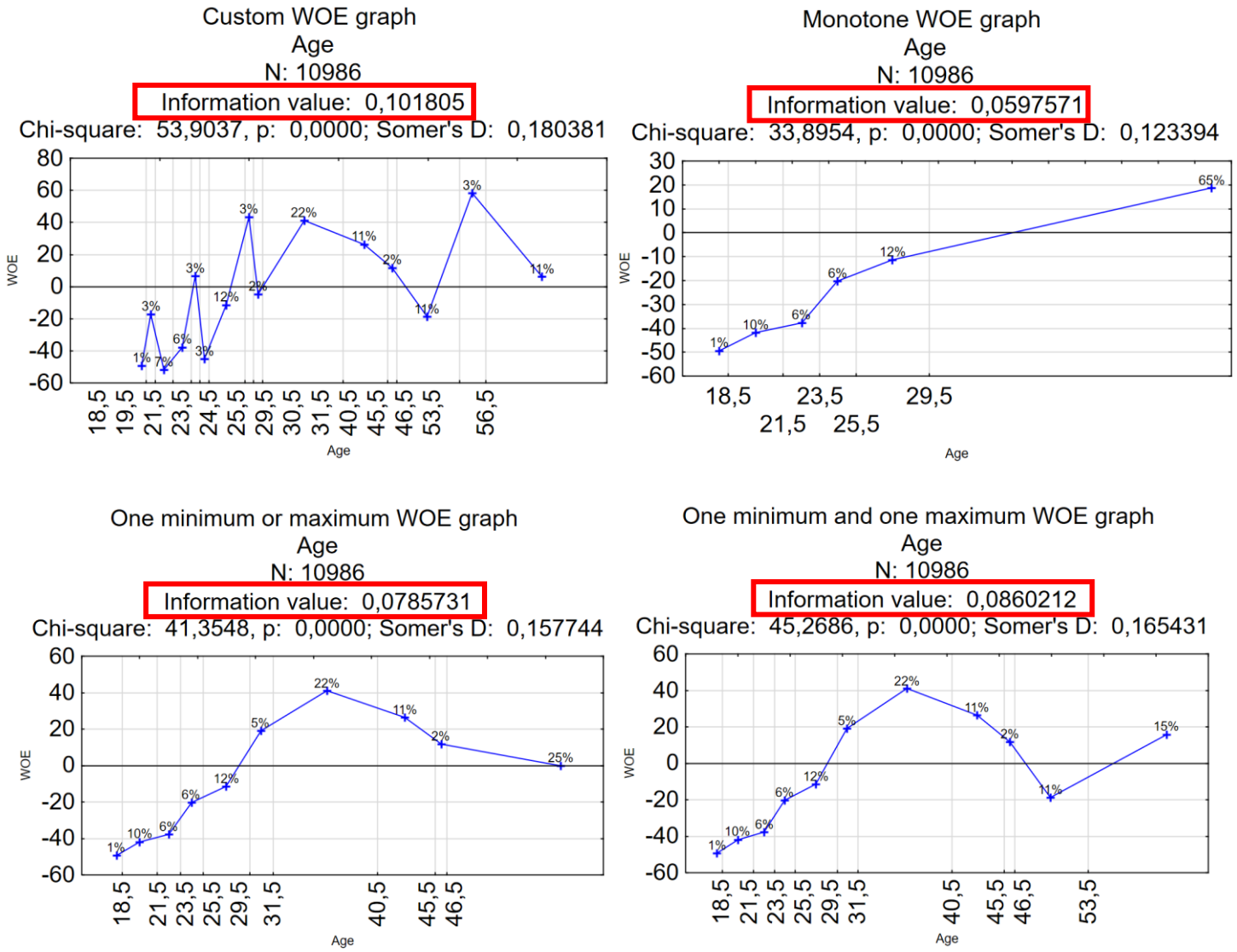
20) Obrázek - Box plot – pro proměnnou HouseMembers ***- před očištěním odlehlých hodnot****. Výstup ze software STATISTICA13. Zdroj: vlastní zpracování*



21) Obrázek - Box plot – pro proměnnou HouseMembers ***- po očištěním odlehlých hodnot****. Výstup ze software STATISTICA13. Zdroj: vlastní zpracování*

5.2.6 Intuitive Behaviour – Weight of Evidence (WOE)

Pro demonstraci celého postupu statistické metody WOE bude vhodné zvolit nejdříve spojitou proměnnou *Age*, která představuje *věk* všech klientů v souboru. Software STATISTICA13 nabízí defaultně čtyři možné varianty přístupu k interpretaci proměnných, se kterými lze dále pracovat:



22) Graf – WOE grafy pro proměnnou Age – různé přístupy. Výstup ze software STATISTICA13. Zdroj: vlastní zpracování

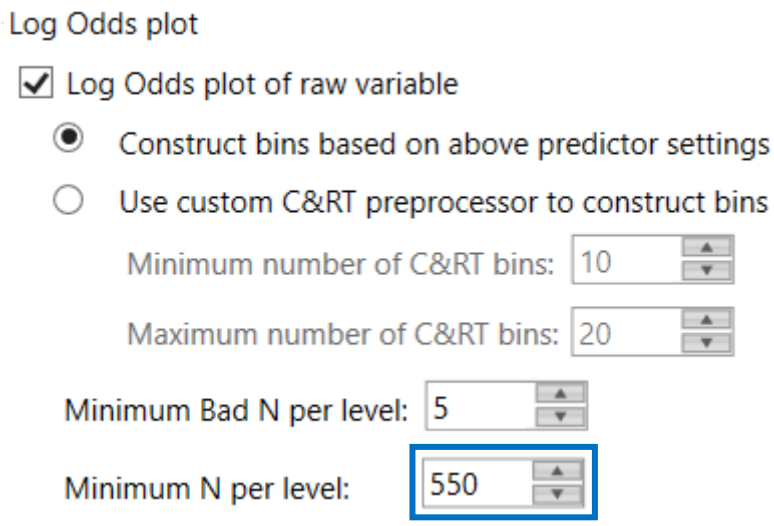
První varianta Custom WOE nabízí rozdělení proměnné na 16 kategorií. Tato metoda nabývá nejvyšší hodnotu Information Value ze všech čtyř přístupů.

Druhá varianta zobrazuje náhled na proměnnou z pohledu monotonie. Tato metoda je vzhledem ke srovnání hodnot Information Value ostatních metod poměrně vágní a ztrácí důležité informace.

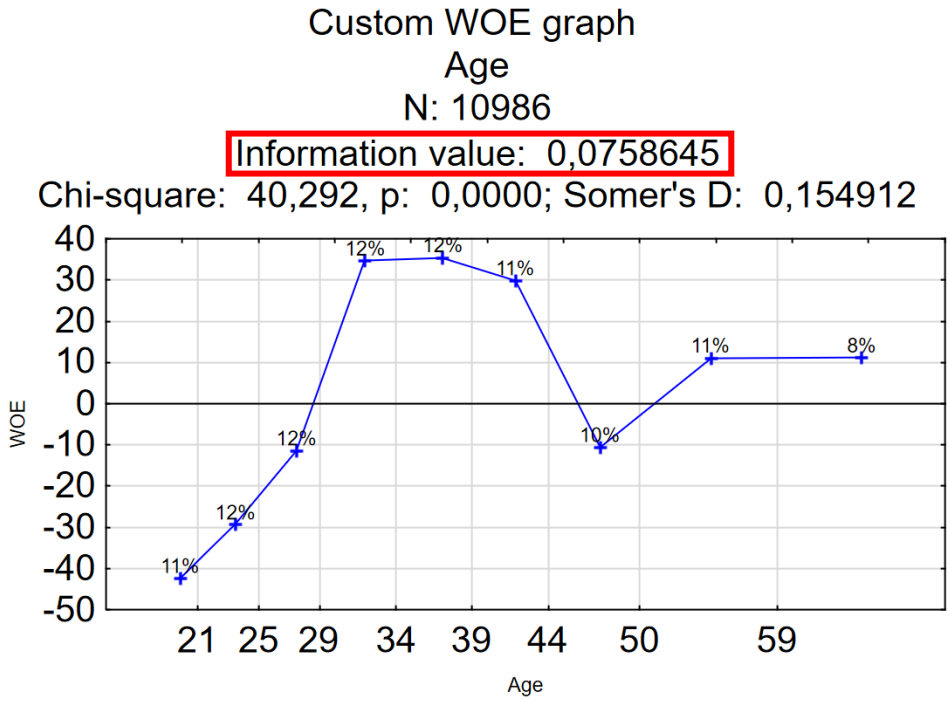
Na základě třetí a čtvrté metody se proměnná rozděluje dle výskytu extrémů. Třetí metoda hledá buď jedno globální minimum, nebo jedno globální maximum hodnoty WOE. Čtvrtá metoda hledá jedno minimum a zároveň jedno maximum.

Siddiqi ve své knize *Intelligent Credit Scoring* popisuje rozdělení proměnných na základě minimálně 5 % zastoupení všech klientů ze soboru v každém binu, aby měly všechny kategorie dostatečně silný vypovídající charakter a nedošlo k překategorizování dané proměnné ve finálním modelu.

Proto bude v případě celého souboru 10 986 klientů nastavena 5 % hranice na 550 klientů.



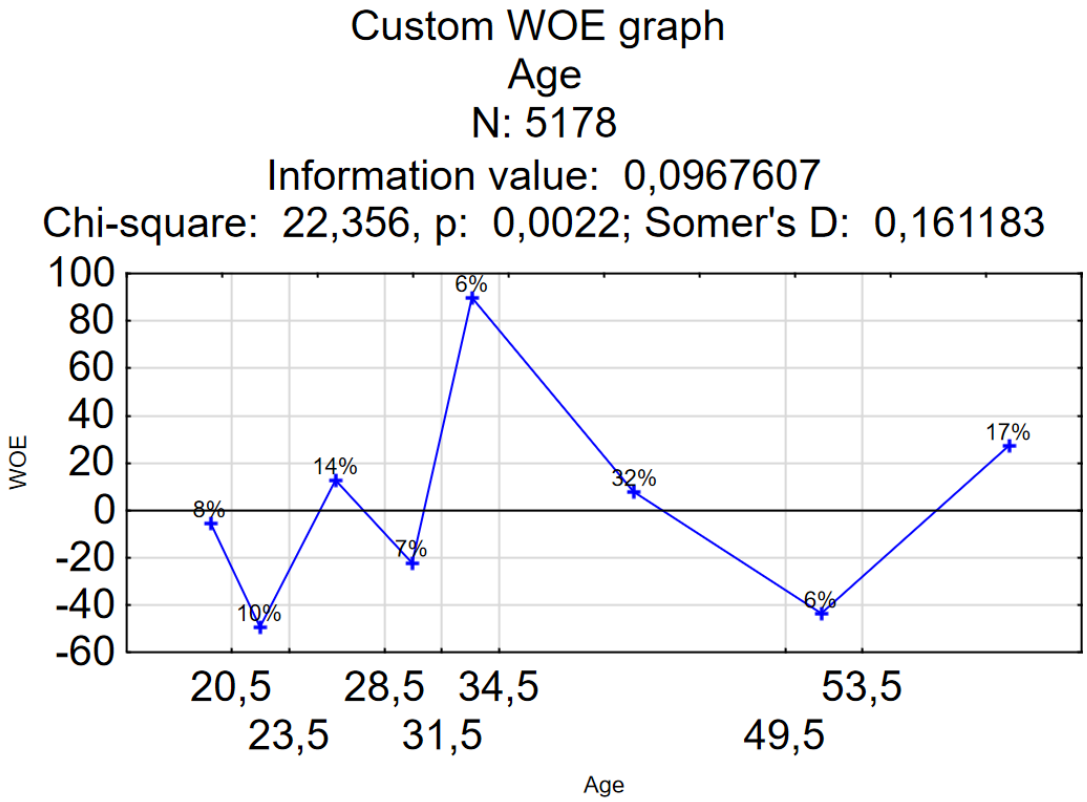
23) Obrázek – Nastavení hodnot v programu STATISTICA13. Zdroj: vlastní zpracování



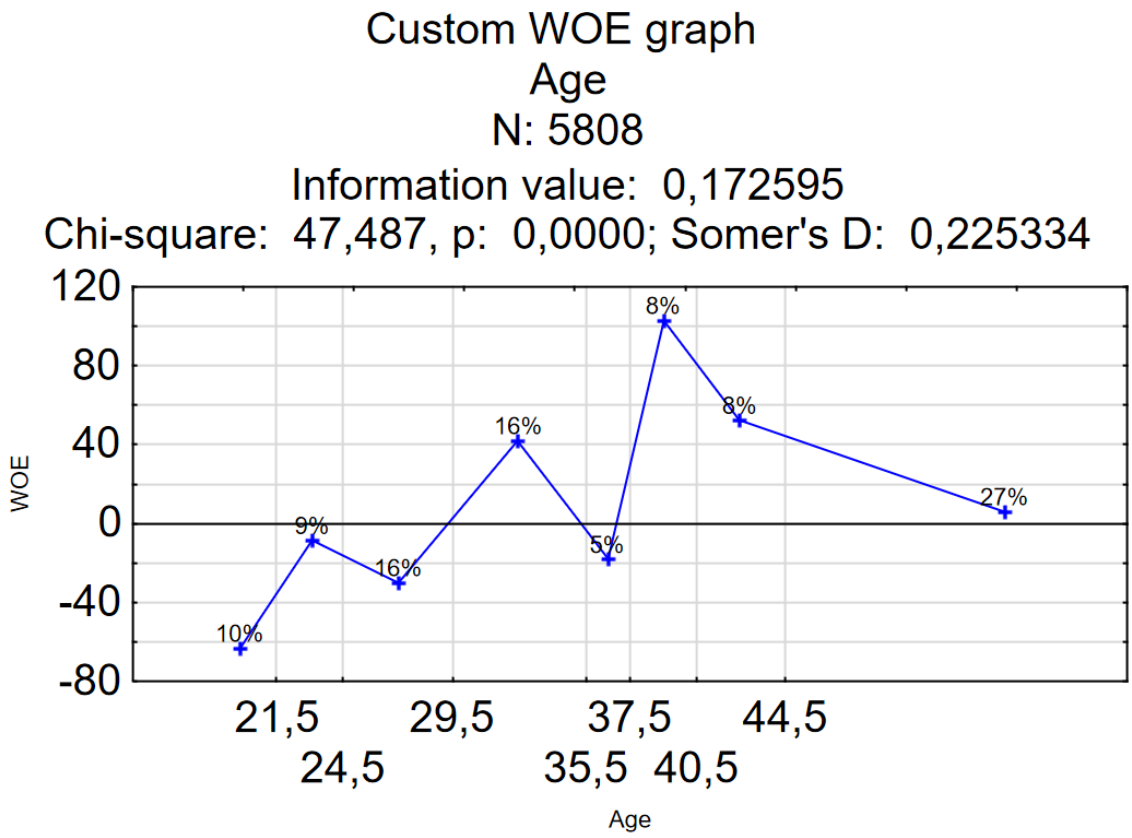
24) Graf – WOE graf pro proměnnou Age s 5 % účastí v jednom binu. Výstup ze software STATISTICA13. Zdroj: vlastní zpracování

Proměnná *Age* se chová dle předpokladu velice intuitivně a logicky. Mladí klienti od 18 do 29 let splácí podstatně hůř, ale zároveň je jejich tendence splácení rostoucí se stářím klientů. Zajímavostí je věková kategorie od 44 do 50 let, kde by se dalo předpokládat, že klienti v této kategorii budou splácet bez problémů. Při bližším ohledání a rozpadu proměnné dle charakteristiky *Gender* lze pozorovat, že ženy od mužů splácí rozdílně (podstatně hůře), viz obr. č. 24. Takový jev se ve statistice nazývá interakce**.** V případě věku se dá tento fenomén interpretovat různými způsoby. Například pokud se vezme v úvahu statistika rozvodů v Praze, tak věková kategorie nejčastějších rozvodů je ekvivalentní s touto hůře splácející kategorií. Po rozvodu se zvýší existenční výdaje oběma zúčastněným stranám a je velká pravděpodobnost, že se dostanou do dluhové pasti. Ženy většinou mívají platový medián nižší, než je tomu u mužů, proto po rozdělení proměnné na muže a ženy lze tento jev takto logicky odůvodnit (www.czso.cz). V této práci dále s interakcemi nebude pracováno z důvodu poměrně obsáhlé problematiky.

WOE Age - Ženy



WOE Age - Muži



25) Grafy – WOE grafy pro proměnnou Age Muži a Age Ženy. Výstup ze software STATISTICA13. Zdroj: vlastní zpracování

V případě dodržení pravidla 5 % celého souboru klientů v jedné kategorii, nabývá proměnná *Age* Information Value hodnotu 0,0759, což udává dostatečné rozdělení faktoru a následný postup proměnné do multivariantní analýzy. Rozdělena byla na 9 binů. Jednotlivé kategorie, jejich koeficienty a celkový přehled je uveden v následující tabulce:

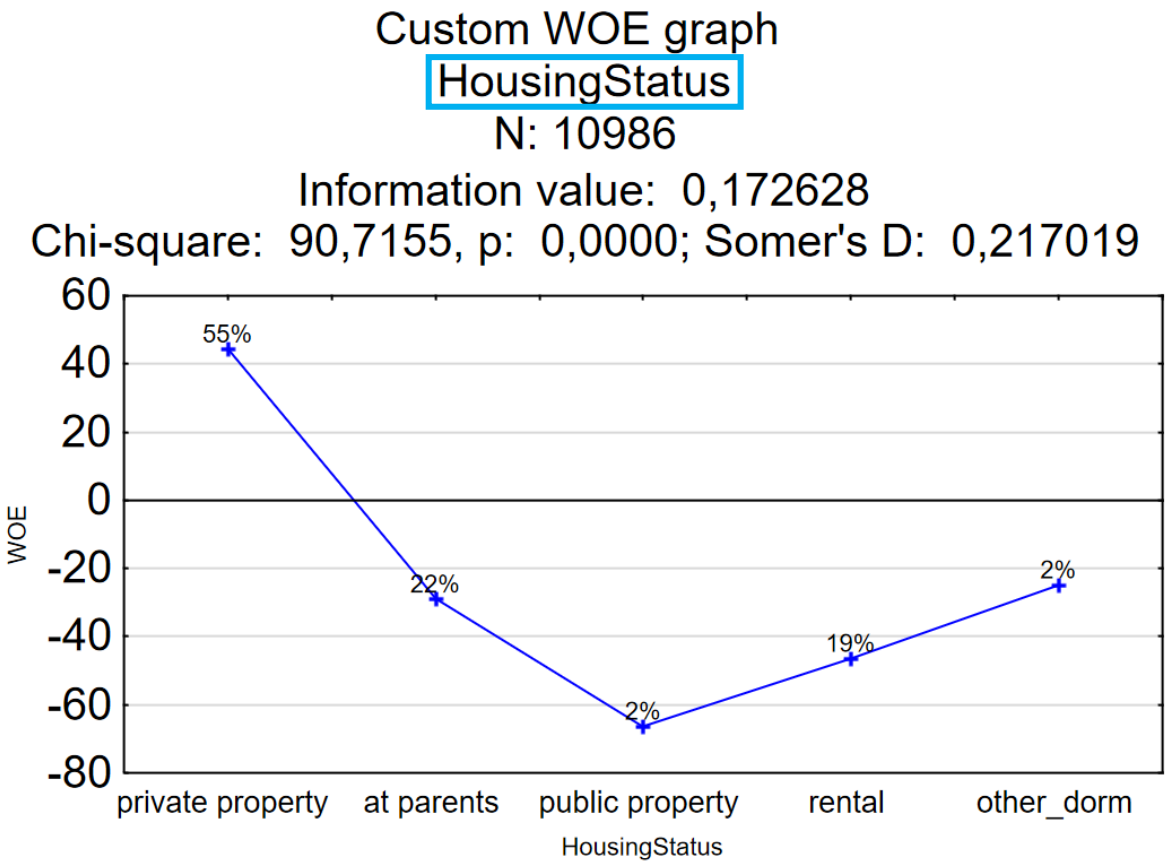
Custom Crosstabulation for Age (ASDATATraning)
Cramer's V=0,0605605
Information value=0,0758645
Somer's D=0,154912
Chi-square= 40,292, p= 0,0000

	1 Goods	2 Bads	3 Gini	4 Information value	5 WOE	6 Boundary
1	1118	89	0,13659892	0,02396996	-42,398188	(-Inf.; 21]
2	1275	89	0,1219836	0,01214192	-29,257708	(21; 25]
3	1199	70	0,1042375	0,00157763	-11,389425	(25; 29]
4	1277	47	0,06847669	0,01247364	34,7479087	(29; 34]
5	1313	48	0,06804868	0,01328595	35,4226696	(34; 39]
6	1216	47	0,07165636	0,00897261	29,8532294	(39; 44]
7	1070	62	0,10354106	0,00122315	-10,636262	(44; 50]
8	1115	52	0,08514644	0,00123923	11,0723798	(50; 59]
9	859	40	0,08502835	0,00098043	11,2247301	(59; Inf.)
10	0	0				MD

26) Tabulka – WOE hodnoty proměnné Age, výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

Jako příklad kategoriální proměnné byla zvolená proměnná *HousingStatus*, která vyjadřuje způsob bydlení klientů. Faktor nabývá hodnot *privat property (soukromá vlastnictví nemovitosti), at parents (u rodičů), public property (státní bydlení), rental (pronájem), other (ostatní).*

Z grafu se dá vyčíst, že v případě proměnné *HousingStatus* lépe splácí klienti, kteří bydlí ve vlastní nemovitosti. Naopak nejhůř splácí klienti, kteří bydlí ve státním podnájmu. Graf a tabulka hodnot jsou uvedeny dále:



27) Graf – WOE pro HousingStatus, výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

Custom Crosstabulation for HousingStatus (ASDATATraning)
Cramer's V=0,0908701
Information value=0,172628
Somer's D=0,217019
Chi-square= 90,7155, p= 0,0000

	1 Goods	2 Bads	3 Gini	4 Information value	5 WOE	6 Levels
1	5891	197	0,0626233	0,08956771	44,3335146	{private property}
2	2243	156	0,12159714	0,02079105	-28,892864	{at parents}
3	178	18	0,1668055	0,01063934	-66,323037	{public property}
4	1920	159	0,14126005	0,05024276	-46,34659	{rental}
5	210	14	0,1171875	0,00138688	-24,659197	{other_dorm}
6	0	0				{MD}

28) Tabulka – výsledné hodnoty proměnné HousingStatus, výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

Model Weight of Evidence bude aplikován na všechny ostatní proměnné, které budou následně postupovat do multivariantní analýzy již rozbinované.

5.2.7 Single Factor Analysis

Rozkategorizované proměnné pomocí WOE metody se dále analyzují pomocí indikátorů Information Value a Prediction power.

U hodnoty Information Value jsou odebrány faktory, které nepřekročily hodnotu 0,02.

Dále se určí hodnoty Gini koeficientu, Kolmogorov-Smirnov indexu a Somers'D indexu. Tyto indexy můžou být pouze informativního charakteru a je na tvůrci modelu, zda proměnné ponechá nebo je odebere z listu proměnných na základě výsledných hodnot. Vybrané proměnné následně vstupují do multifaktor analýzy. U každého faktoru by se navíc mělo brát v potaz, zda se jedná o proměnnou, do které se jednoduše sbírají data, popřípadě zdali se jedná o proměnnou, u které je sběr dat pro banku drahý a postrádá efektivitu.

Výsledné hodnoty pro jednotlivé proměnné jsou uvedeny v následující tabulce:

Variable name	Meaning of the variable	Variable type	Information Value	Gini index	KS	Somer's D
Income_M	Příjem	Continuous	0,24378	0,25758	0,200	0,260
InsurancePay	Platba pojištění	Continuous	0,24247	0,19866	0,200	0,200
CurrentEmpSince	Počet měsíců v současném zaměstnání	Continuous	0,20162	0,23256	0,160	0,230
Education	Vzdělání	Categorical	0,18711	0,10586	0,110	0,207
ProductType	Typ produktu	Categorical	0,17493	0,13508	0,120	0,140
HousingStatus	Způsob bydlení	Categorical	0,17263	0,20203	0,200	0,200
WorkingPosition	Pracovní pozice	Categorical	0,14520	0,18716	0,160	0,203
RegionCont	Kraj trvalého bydliště	Categorical	0,13704	0,09002	0,090	0,190
MaritalStatus	Rodinný stav	Categorical	0,12692	0,15918	0,160	0,170
PersonWithNoInc	Počet členů bez příjmu	Continuous	0,11663	0,11964	0,110	0,120
CommonEquity	Spojené jmění manželů	Categorical	0,11169	0,15037	0,150	0,150
RegionPerm	Kraj přechodného bydliště	Categorical	0,10687	0,12724	0,090	0,175
HouseMembers	Počet členů v domácnosti	Continuous	0,07916	0,13932	0,110	0,140
Age	Věk	Continuous	0,07586	0,15491	0,120	0,155
DifferentAdress	Rozdíl adres	Categorical	0,06230	0,08436	0,080	0,084
Savings	Výdaje spoření	Continuous	0,04781	0,05416	0,050	0,054
DepartmentEmpl	Obor zaměstnavatele	Categorical	0,04384	0,03981	0,040	0,097
CountryCont	Země kontaktní adresy	Categorical	0,03905	0,03883	0,040	0,040
EmployerType	Druh zaměstnavatele	Categorical	0,02654	0,04382	0,020	0,077
Title	Titul	Categorical	0,02604	0,02278	0,020	0,020
ContAddrSince_M	Počet měsíců na kontaktní adrese	Continuous	0,02048	0,08095	0,060	0,080

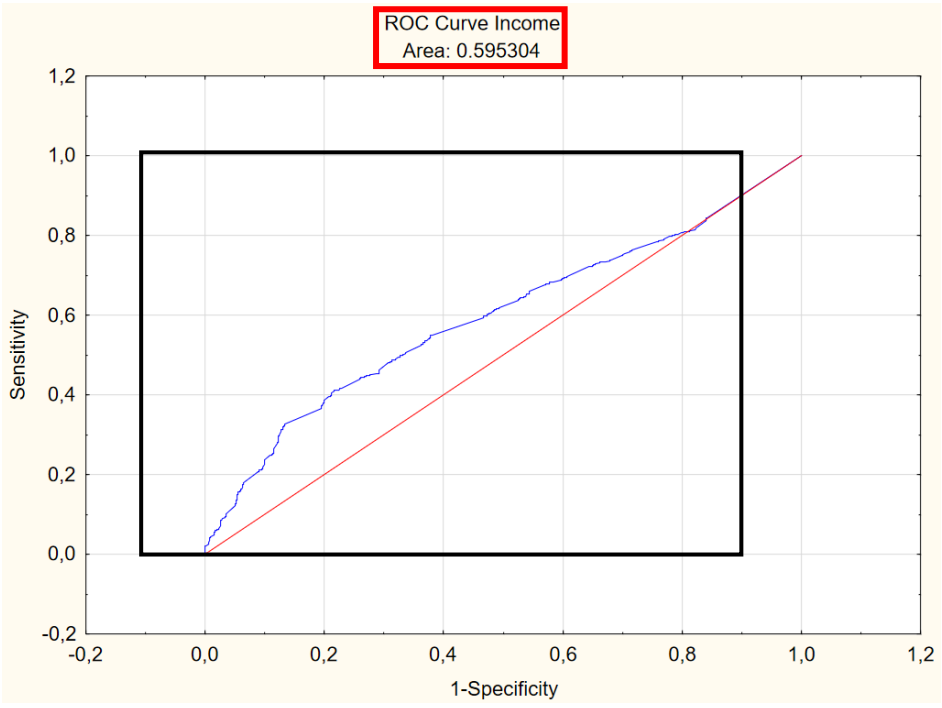
29) Tabulka – výsledné hodnoty IV, Gini indexu, Kolmogorov-Smirnov indexu, Somer's D indexu pro jednotlivé proměnné. Zdroj: vlastní zpracování

Na základě výsledků Information Value prošlo do modelu 21 proměnných. Zbylé proměnné byly ze souboru vyjmuty.

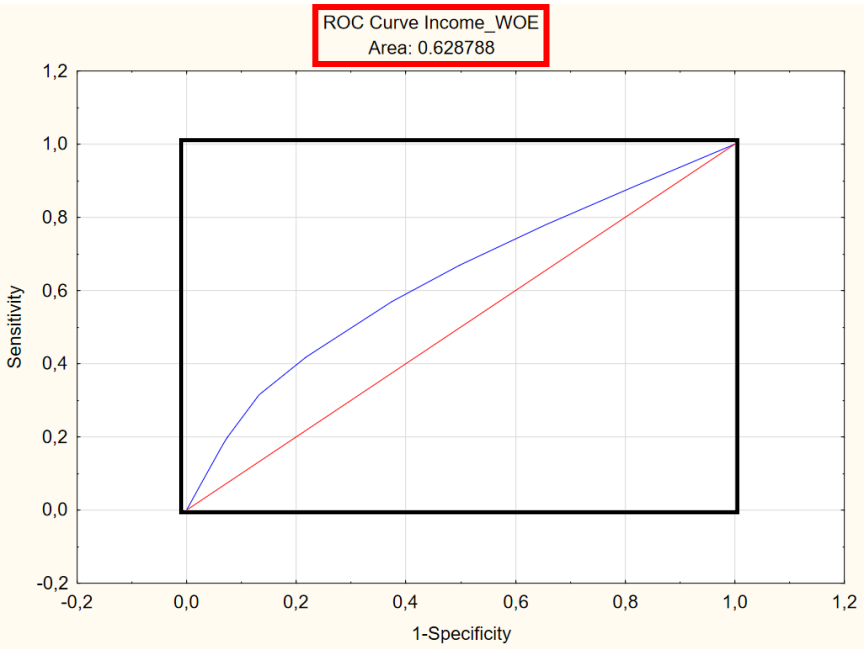
Dále bude testováno, který přístup je lepší. Zda nechat postoupit všechny proměnné, které projdou na základě Information Value předvýběru do multivariantní analýzy, nebo zda pracovat pouze s kratším výběrem proměnných na základě všech zmiňovaných indexů. Siddiqi pojednává o tom, že power indexy mají pouze informativní charakter, na který buď bude brán zřetel, nebo ne. Z toho důvodu je nutné udělat test obou přístupů. Finální stanovisko bude nakonec potvrzeno na základě Cross-Validace a hodnoty AUC výsledného modelu.

Pro demonstraci síly je vhodná metoda například pomocí ROC křivek.

Na následujících dvou grafech je pomocí hodnot AUC křivek demonstrovaná síla proměnné *Income* před rozkategorizováním a po rozkategorizování pomocí metody WOE. Na základě výsledku ROC křivky (vyšší hodnota AUC) je patrné, že proměnná po rozkategorizování má větší vypovídající charakter, než před rozkategorizováním. Vzhledem k síle a rozdílu jednotlivých hodnot je důležitost rozdělování proměnných enormní.



*30) Graf – ROC křivka – Pro proměnnou příjem v **původních hodnotách**. Výstup ze software STATISTICA13. Zdroj: vlastní zpracování*



*31) Graf – ROC křivka– pro proměnnou **příjem v hodnotách WOE**. Výstup ze software STATISTICA13. Zdroj: vlastní zpracování*

5.2.8 Multikolinearita

Dále je nutné zjistit, zda se mezi vysvětlujícími proměnnými nevyskytuje silná nežádoucí závislost. Před finálním vstupem do modelu byly spojité proměnné testovány pro případný výskyt multikolinearity pomocí Pearsonových korelačních koeficientů uspořádaných do korelační matice *(Correlation Matrix).* V případě kategoriálních proměnných byly pro identifikaci multikolinearity využity Pearsonovy kontingenční koeficienty. Za vysokou multikolinearitu byly považovány hodnoty koeficientů vyšší nebo rovny 0,8 (v absolutní hodnotě).

Korelační závislost:

Jak je patrné z korelační matice ze strany 26, v případě spojitých proměnných nedošlo k překročení žádné závislosti nad 0,8. V takovém případě dále postupují do modelu všechny spojité proměnné.

Correlation Matrix	WOE_Income_M	WOE_InsurancePay	WOE_CurrentEmpSince	WOE_PersonWithNoInc	WOE_Age	WOE_HouseMembers	WOE_ContAddrSince_M	WOE_Savings
WOE_Income_M	1,0000	-0,0173	-0,2070	-0,0284	-0,1822	0,0161	0,0301	0,0092
WOE_InsurancePay	-0,0173	1,0000	-0,0457	0,0089	-0,1029	-0,0290	0,0240	-0,2171
WOE_CurrentEmpSince	-0,2070	-0,0457	1,0000	0,0182	-0,0791	-0,0376	-0,0532	0,0270
WOE_PersonWithNoInc	-0,0284	0,0089	0,0182	1,0000	-0,1217	-0,6931	-0,0642	-0,0028
WOE_Age	-0,1822	-0,1029	-0,0791	-0,1217	1,0000	-0,0491	0,2410	-0,0012
WOE_HouseMembers	0,0161	-0,0290	-0,0376	-0,6931	-0,0491	1,0000	-0,0498	-0,0120
WOE_ContAddrSince_M	0,0301	0,0240	-0,0532	-0,0642	0,2410	-0,0498	1,0000	-0,0078
WOE_Savings	0,0092	-0,2171	0,0270	-0,0028	-0,0012	-0,0120	-0,0078	1,0000

32) Tabulka – Correlation Matrix. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

Personův kontingenční koeficient

Vzhledem k výstupu kontingenční tabulky, kde jsou získané hodnoty $\boldsymbol{\chi^2}$ testu, bude eliminace proměnných vycházet z následujícího vztahu (dle vlastního odvození):

$$C = \sqrt{\frac{\boldsymbol{\chi^2}}{\boldsymbol{\chi^2} + n}}$$

$$C^2 = \frac{\boldsymbol{\chi^2}}{\boldsymbol{\chi^2} + n}$$

$$C^2 * (\boldsymbol{\chi^2} + n) = \boldsymbol{\chi^2}$$

$$C^2 * \boldsymbol{\chi^2} + C^2 * n = \boldsymbol{\chi^2}$$

$$C^2 * \boldsymbol{\chi^2} - \boldsymbol{\chi^2} = -C^2 * n$$

$$\boldsymbol{\chi^2} * (C^2 - 1) = -C^2 * n$$

$$\boldsymbol{\chi^2 = \frac{-C^2 * n}{C^2 - 1}}$$

Síla závislosti dle kontingenčního koeficientu se posuzuje podle stejné škály, jako u korelačního koeficientu. Z toho vyplývá, že pokud nemá být závislost větší nebo rovna hodnotě 0,8, musí platit následující vztah:

$$\boldsymbol{\chi^2 < \frac{-C^2 * n}{C^2 - 1}}$$

n = počet klientů v celém souboru

C = hodnota Pearsonova kontingenčního koeficientu

χ^2= hodnota chi kvadrát testového kritéria

Po dosazení číselných hodnot vychází hraniční hodnota chi kvadrát testu pro identifikaci výskytu multikolinearity v modelu tato:

$$\chi^2 < \frac{-0{,}8^2 * 10\,968}{0{,}8^2 - 1}$$

$$\boldsymbol{\chi^2 < 19498,\overline{6}}$$

V tomto případě splňují podmínku všechny kategoriální proměnné, tudíž žádná nepřekročí závislost vyšší než 0,8. Všechny kategoriální proměnné postupují dále do modelu. **(**viz Příloha B). U kategoriálních proměnných se zřídka vyskytuje velmi silná závislost.

Pro příklad jsou uvedené dvě proměnné WOE_MaritalStatus a WOE_CountryCont, a jejich výsledné hodnoty chi kvadrát testů (ostatní testy jsou přiloženy v příloze A).

WOE_MaritalStatus	Chi-square	p-value
WOE_Title	2013,219	0,000000
WOE_WorkingPosition	1226,584	0,000000
WOE_DepartmentEmpl	313,471	0,000000
WOE_EmployerType	256,043	0,000000
WOE_ProdectType	209,427	0,000000
WOE_DifferentAdress	49,380	0,000000
WOE_RegionCont	35,268	0,000000
WOE_CountryCont	26,831	0,000000
WOE_HousingStatus	22,366	0,000002
WOE_RegionPerm	16,401	0,000275
WOE_MaritalStatus	9,261	0,002341
WOE_CommonEquity	5,018	0,025088

WOE_CountryCont	Chi-square	p-value
WOE_RegionCont	3864,593	0,000000
WOE_HousingStatus	711,864	0,000000
WOE_WorkingPosition	532,333	0,000000
WOE_MaritalStatus	349,882	0,000000
WOE_CommonEquity	297,575	0,000000
WOE_RegionPerm	223,910	0,000000
WOE_ProdectType	165,151	0,000000
WOE_DifferentAdress	72,734	0,000000
WOE_EmployerType	65,974	0,000000
WOE_DepartmentEmpl	29,192	0,000000
WOE_Education	26,831	0,000000
WOE_Title	16,028	0,000062

33) Tabulky – výsledné hodnoty chi kvadrát testů pro proměnné WOE_MaritalStatus a WOE_CountryCont. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.2.9 Short list proměnných a long list proměnných

Long list of variables je seznam proměnných, které dosáhly hodnoty Information Value větší nebo rovno 0,02. Na základě výsledků hodnot Information Value vznikl tzv. Long list of Variables, kde prošlo testem 21 proměnných.

Long list of variables	
Variable name	**Meaning of the variable**
Income_M	Příjem
CurrentEmpSince	Počet měsíců v současném zaměstnání
Education	Vzdělání
ProductType	Typ produktu
InsurancePay	Platba pojištění
WorkingPosition	Pracovní pozice
HousingStatus	Způsob bydlení
MaritalStatus	Rodinný stav
Age	Věk
CommonEquity	Spojené jmění manželů
PersonWithNoInc	Počet členů bez příjmu
HouseMembers	Počet členů v domácnosti
RegionCont	Kraj trvalého bydliště
Savings	Výdaje spoření
RegionPerm	Kraj přechodného bydliště
DifferentAdress	Rozdíl adres
DepartmentEmpl	Obor zaměstnavatele
EmployerType	Druh zaměstnavatele
ContAddrSince_M	Počet měsíců na kontaktní adrese
Title	Titul
CountryCont	Země kontaktní adresy

34) Tabulka – Long list of variables – dlouhý seznam proměnných. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

Dále na základě takto otestovaných proměnných vzniká i Short list of variables (krátký seznam proměnných).

Short list of variables je seznam proměnných, které dosáhly hodnoty Information Value větší nebo rovno 0,02. Pokud alespoň u jednoho indexu (Gini indexu, Kolmogorov-Smirnov indexu a Somer's D indexu) nabyla proměnná hodnotu vyšší nebo rovno 0,1, postoupila dále do seznamu proměnných. Takto prošlo 14 proměnných.

Short list of variables	
Variable name	**Meaning of the variable**
Income_M	Příjem
InsurancePay	Platba pojištění
CurrentEmpSince	Počet měsíců v současném zaměstnání
Education	Vzdělání
ProductType	Typ produktu
HousingStatus	Způsob bydlení
WorkingPosition	Pracovní pozice
RegionCont	Kraj trvalého bydliště
MaritalStatus	Rodinný stav
PersonWithNoInc	Počet členů bez příjmu
CommonEquity	Spojené jmění manželů
RegionPerm	Kraj přechodného bydliště
HouseMembers	Počet členů v domácnosti
Age	Věk

35) Tabulka – Short list of variables – krátký seznam proměnných. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.2.10 Multi Factor analysis pomocí logistické regrese

Pro další práci s daty bude sestaveno a porovnáno 12 modelů v následující kombinaci:

Model	Short/long list	Kombinace proměnných dle WOE hodnot	Metoda logistické regrese
MODEL 01	Long list	Spojité WOE	Forward Stepwise
MODEL 02	Long list	Spojité WOE	Backward Stepwise
MODEL 03	Short list	Spojité WOE	Forward Stepwise
MODEL 04	Short list	Spojité WOE	Backward Stepwise
MODEL 05	Long list	Kategoriální a spojité WOE	Forward Stepwise
MODEL 06	Long list	Kategoriální a spojité WOE	Backward Stepwise
MODEL 07	Short list	Kategoriální a spojité WOE	Forward Stepwise
MODEL 08	Short list	Kategoriální a spojité WOE	Backward Stepwise
MODEL 09	Long list	Kategoriální WOE	Forward Stepwise
MODEL 10	Long list	Kategoriální WOE	Backward Stepwise
MODEL 11	Short list	Kategoriální WOE	Forward Stepwise
MODEL 12	Short list	Kategoriální WOE	Backward Stepwise

36) Tabulka – Kombinace modelů Multi Factor Analysis. Výstup z programu MS Excel. Zdroj: vlastní zpracování

Pomocí metody Cross-Validace budou v následujícím kroku validovány výsledky těchto 12 modelů.

5.2.11 Cross-Validation (Porovnání na testovacích datech), hladina cut-off

Pro další práci s daty je nutné vybrat správnou kombinaci modelu, který bude nejlépe a spolehlivě predikovat výsledky.

Dále je důležité správné nastavení hodnoty cut-off. Tuto hodnotu si banky určují samy. Je odrazem strategie a míry rizika, kterou je instituce ochotná podstoupit. V ideálním případě by se mělo jednat o takovou hodnotu, při které dochází k maximální správné predikci výsledků a zároveň minimalizaci rizika a ušlých obchodních příležitostí.

Pro první testování Cross-Validace byla zvolena hodnota cut-off hladiny na 0,95. Následující tabulky ukazují výsledky pro MODELY 03, 04, 09 a 10. Ostatní testy viz příloha B.

	Summary Frequency Table MODEL 03			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**3015**	383	3398
Total Percent		**82,94%**	10,54%	93,48%
Count	1	217	**20**	237
Total Percent		5,97%	**0,55%**	6,52%
Count	All Grps	3232	403	3635
Total Percent		88,91%	11,09%	

37) Tabulka – Cross-Validace – MODEL 03, CUT-OFF 0,95. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

	Summary Frequency Table MODEL 04			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**3015**	383	3398
Total Percent		**82,94%**	10,54%	93,48%
Count	1	217	**20**	237
Total Percent		5,97%	**0,55%**	6,52%
Count	All Grps	3232	403	3635
Total Percent		88,91%	11,09%	

38) Tabulka – Cross-Validace – MODEL 04, CUT-OFF 0,95. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

	Summary Frequency Table MODEL 09			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**2338**	1060	3398
Total Percent		**64,32%**	29,16%	93,48%
Count	1	48	**189**	237
Total Percent		1,32%	**5,20%**	6,52%
Count	All Grps	2386	1249	3635
Total Percent		65,64%	34,36%	

39) Tabulka – Cross-Validace – MODEL 09, CUT-OFF 0,95. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

	Summary Frequency Table MODEL 10			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**2331**	1067	3398
Total Percent		**64,13%**	29,35%	93,48%
Count	1	49	**188**	237
Total Percent		1,35%	**5,17%**	6,52%
Count	All Grps	2380	1255	3635
Total Percent		65,47%	34,53%	

40) Tabulka – Cross-Validace – MODEL 10, CUT-OFF 0,95. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

Na základě výsledků jednotlivých modelů nelze brát v potaz modely 05 – 08. Výsledné hodnoty nejsou dostačující. I přes celkem vysoce nastavenou hladinu cut-off na 0,95 model není schopný rozeznat rizikové klienty a poskytnul by úvěr 100 % klientům – nerizikovým i rizikovým. Zvýšení hladiny cut-off v tomto případě nedává smysl z důvodu vysokých hodnot výsledných pravděpodobností u klientů, které se lišily pouze v setinách či tisícinách.

Modely s kategoriálními nebo spojitými proměnnými naopak předpovídaly výsledky spolehlivě. Z kategoriálních modelů nejlépe predikují modely MODEL 09 a MODEL 10, ze spojitých MODEL 03 a MODEL 04.

MODEL 03 a MODEL 04 správně odhadl dohromady 83,49 % a MODEL 09 a MODEL 10 odhadl správně chování u 69,3 % klientů.

U MODELU 09 a 10 se vyskytuje vysoký podíl klientů, kterým bylo predikováno špatné chování, ale chování měli dobré. V tomto případě jde o 29,35 % klientů a jedná se o ušlou příležitost. Vzhledem k této vysoké hodnotě lze dále posunout hodnotu cut-off tak, aby model nebyl tolik přísný. Na druhou stranu podíl klientů, kterým byl úvěr poskytnut, ale chování měli v reálném případě špatné, je poměrně malý, a to 1,32 %.

MODEL 03 a 04 také vykazuje poměrně vysoký podíl klientů, kterým byl úvěr zamítnut, přitom měl být poskytnut. Jedná se o podíl 10,54 %. V tomto případě posun hladiny cut-off nemá smysl. S každým dalším posunem by model nevykazoval relevantní výsledky a zvyšoval by se podíl nesplacených úvěrů, a to víc než 5,97 %. Modely se sníženou hladinou cut-off také nedetekovaly téměř žádné rizikové klienty. Dále s těmito modely nebude pracováno.

U MODELU 09 a 10 bude testována hranice cut-off na hladině 0,85.

Výsledky jsou následující:

	Summary Frequency Table MODEL 09, CUT-OFF 0,85			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**3208**	190	3398
Total Percent		**88,25%**	5,23%	93,48%
Count	1	186	**51**	237
Total Percent		5,12%	**1,40%**	6,52%
Count	All Grps	3394	241	3635
Total Percent		93,37%	6,63%	

41) Tabulka – Cross-Validace – MODEL 09, CUT-OFF 0,85. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

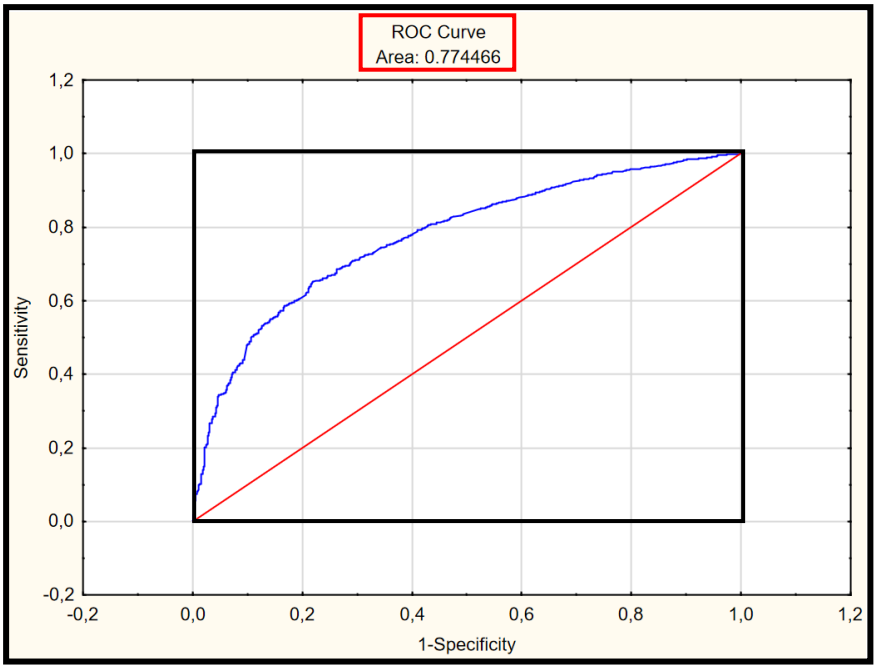
	Summary Frequency Table MODEL 10, CUT-OFF 0,85			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**3208**	190	3398
Total Percent		**88,25%**	5,23%	93,48%
Count	1	186	**51**	237
Total Percent		5,12%	**1,40%**	6,52%
Count	All Grps	3394	241	3635
Total Percent		93,37%	6,63%	

42) Tabulka – Cross-Validace – MODEL 10, CUT-OFF 0,85. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

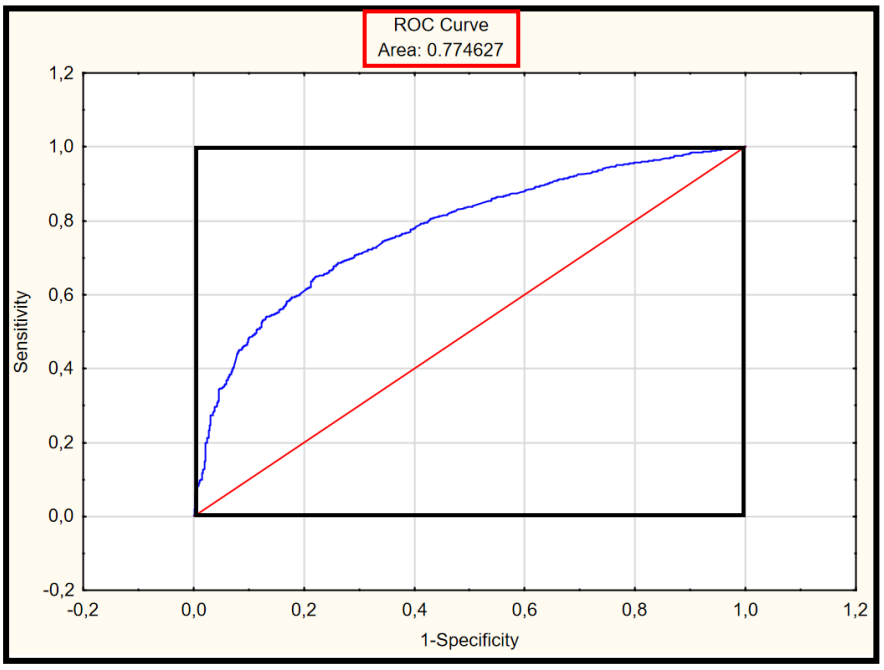
Posun cut-off hranice se pozitivně projevil u MODELU 09 a 10, kde se snížil podíl špatně zařazených klientů. U obou modelů vyšly stejné výsledky, proto pro finální výběr bude nutné přihlédnout k hodnotě ROC křivky.

5.2.12 ROC křivka

Z MODELU 09 a 10 bude vybrán MODEL 10 a to z důvodů vyšší hodnoty ROC křivky:



43) Graf – ROC křivka celý dataset – MODEL 09. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

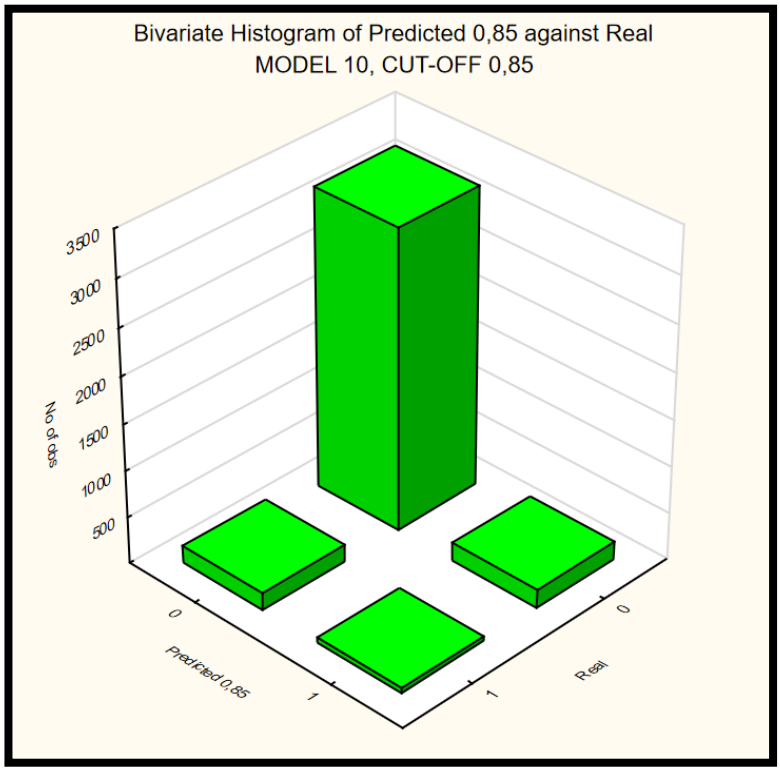


44) Graf – ROC křivka celý dataset – MODEL 10. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

Výsledný model je tedy MODEL 10, který je kombinací WOE kategoriálních proměnných, složený z Long listu proměnných a logistická regrese je provedena metodou backward stepwise. Model vykazuje při hodnotě cut-off hladiny 0,85 tyto hodnoty:

	Summary Frequency Table MODEL 10; CUT-OFF 0,85			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**3208**	**190**	3398
Total Percent		**88,25%**	**5,23%**	93,48%
Count	1	**186**	**51**	237
Total Percent		**5,12%**	**1,40%**	6,52%
Count	All Grps	3394	241	3635
Total Percent		93,37%	6,63%	

45) Tabulka – Cross-Validace – MODEL 10, CUT-OFF 0,85. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování



46) Graf – Histogram – MODEL 10, CUT-OFF 0,85. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

5.2.13 Satisfactory result (Kalibrace vybraného modelu)

U MODELU 10 bude dále testováno správné nastavení hladiny cut-off.

Na základě výstupů nastavení jednotlivých hladin cut-offů lze interpretovat následující výsledky (zbylé testy jsou přiloženy v příloze C):

Čím vyšší cut-off je nastaven, tím lépe model detekuje špatné chování klientů. Zároveň se ale zvyšuje chybovost u klientů, kteří by spláceli v pořádku, protože jim model přiřadí pravděpodobnost nesplacení závazků. Z toho důvodu banka při nastavení vyšší hladiny cut-offu přichází o dobré klienty, tím pádem i o ušlý zisk. Zároveň pokud nesprávně detekuje klientské špatné chování, zvýší se počet nesplacených úvěrů.

Vzhledem k těmto okolnostem by bylo nejvhodnější nastavit hladinu cut-offu na 0,87 z důvodu velkého nárůstu klientů, kteří by spadli do kategorie neposkytnutí úvěru.

	Summary Frequency Table MODEL 10; CUT-OFF 0,87			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**3121**	277	3398
Total Percent		**85,86%**	7,62%	93,48%
Count	1	161	**76**	237
Total Percent		4,43%	**2,09%**	6,52%
Count	All Grps	3282	353	3635
Total Percent		90,29%	9,71%	

47) Tabulka – Cross-Validace – MODEL 10 – finální verze, CUT-OFF 0,87. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

Výsledné nastavení hladiny cut-off záleží na strategii dané finanční instituce, která buď přijme vyšší riziko za účelem získání podílu market share (což by znamenalo hladinu cut-off nastavit na nižší hodnotu), nebo se staví k riziku negativně a raději zvolí bezpečnější cestu i za riziko ztráty obchodní příležitosti a částečně zisku (nastavení hladiny na vyšší hodnotu). Dále je důležitá abilita vymáhání nesplacených pohledávek bankou.

5.2.14 Výsledná scorekarta aplikačních dat:

Finální scorekarta vytvořená na sadě aplikačních dat. Jako referenční hodnoty byly zvoleny hodnoty s nejnižšími koeficienty:

APPLICATION SCORECARD			
#	**Variable**	**Cathegory**	**Estimate**
0		Intercept	-1,0062
1	WOE_ProdectType	OVD , CC , CL	0
2		HL , CO	1,5878
3	WOE_HousingStatus	at parents , other_dorm , rental , public property	0
4		private property	0,5881
5	WOE_WorkingPosition	NONE_V domacnosti , manual_definite	0
6		manual_indefinite , NONE_Duchodce , NONE_Student	0,1530
7		intellectual_definite , manager_definite , NONE_Rentier/Jine , intellectual_indefinite , manager_indefinite , NONE_Podnikatel	0,4675
8	WOE_Education	Maturita , VOS , Vyucen , Zakladni	0
9		VS	0,9288
10	WOE_MaritalStatus	Svobodny , DruhDruzka , Vdova , Rozvedeny	0
11		Marriage , Reg_partner	0,4393
12	WOE_CommonEquity	No/Rozdel , Opusteni	0
		Yes	0,4287
13	WOE_InsurancePay	InsurancePay = 0	0
		0 < InsurancePay <= 500	0,9203
14		InsurancePay > 500	1,0991
15	WOE_Age	44 < Age <= 50	0
16		50 < Age <= 59	0,2360
17		Age <= 21	0,2426
18		39 < Age <= 44	0,3202
19		21 < Age <= 25	0,5656
20		34 < Age <= 39	0,6049
21		Age > 59	0,6165
22		25 < Age <= 29	0,6403
23		29 < Age <= 34	0,8487
24	WOE_RegionPerm	Foreign , CZ_Jihomoravsky kraj , CZ_Ustecky kraj	0

25		CZ_Jihocesky kraj , CZ_Zlinsky kraj , CZ_Plzensky kraj , CZ_Pardubicky kraj , CZ_Vysocina , SK , CZ_Stredocesky kraj , CZ_Olomoucky kraj , CZ_Moravskoslezsky kraj , CZ_Liberecky kraj , CZ_Hlavni mesto	0,3944
26		CZ_Kralovehradecky kraj , CZ_Karlovarsky kraj	1,1433
27	WOE_DifferentAdress	1	0
28		0	0,5340
29	WOE_EmployerType	Podnikatel/OSVC	0
30		NONE_Student , Other , NONE_V domacnos , Soukroma inst , NONE_Duchodce , Verejna inst , NONE_Podnikatel , NONE_Rentier	0,4594
31		Finance	1,6462
32	WOE_PersonWithNoInc	0 < PersonWithNoInc <= 1	0
33		PersonWithNoInc = 0	0,1833
34		PersonWithNoInc > 1	0,8048
35	WOE_CurrentEmpSince	CurrentEmpSince <= 4	0
36		12 < CurrentEmpSince <= 27	0,1999
37		27 < CurrentEmpSince <= 51	0,2007
38		4 < CurrentEmpSince <= 12	0,3485
39		88 < CurrentEmpSince <= 101	0,5611
40		51 < CurrentEmpSince <= 88	0,5828
41		101 < CurrentEmpSince <= 103	0,6888
42		103 < CurrentEmpSince <= 180	0,8994
43		CurrentEmpSince > 180	1,3758

48) Tabulka – Finální scorekarta aplikačních dat. Výstup ze softwaru STATISTICA13. Zdroj: vlastní zpracování

5.2.15 Výsledná rovnice logistické regrese a výpočtu splacení aplikačního modelu:

***P* = [**exp(-1,0062 + 1,5878 * WOE_Product_Type (HL , CO) + 0,5881 * WOE_HousingStatus (private property) + 0,1530 * WOE_WorkingPosition (manual_indefinite , NONE_Duchodce , NONE_Student) + 0,4675 * WOE_WorkingPosition (intellectual_definite , manager_definite , NONE_Rentier/Jine , intellectual_indefinite , manager_indefinite , NONE_Podnikatel) + 0,9288 * WOE_Education (VS) + 0,4393 * WOE_MaritalStatus (Marriage , Reg_partner) + 0,4287 * WOE_CommonEquity (Yes) + 0,9203 * WOE_InsurancePay (0 < InsurancePay <= 500) + 1,0991 * WOE_InsurancePay (InsurancePay > 500) + 0,2360 * WOE_Age (50 < Age <= 59) + 0,2426 * WOE_Age (Age <= 21) + 0,3202 * WOE_Age (39 < Age <= 44) + 0,5656 * WOE_Age (21 < Age <= 25) + 0,6049 * WOE_Age (34 < Age <= 39) + 0,6165 * WOE_Age (Age > 59) + 0,6403 * WOE_Age (25 < Age <= 29) + 0,8487 * WOE_Age (29 < Age <= 34) + 0,3944 * WOE_RegionPerm (CZ_Jihocesky kraj , CZ_Zlinsky kraj , CZ_Plzensky kraj , CZ_Pardubicky kraj , CZ_Vysocina , SK , CZ_Stredocesky kraj , CZ_Olomoucky kraj , CZ_Moravskoslezsky kraj , CZ_Liberecky kraj , CZ_Hlavni město) + 1,1433 * WOE_RegionPerm (CZ_Kralovehradecky kraj , CZ_Karlovarsky kraj) + 0,5340 * WOE_DifferentAdress (0) + 0,4594 * WOE_EmployerType (NONE_Student , Other , NONE_V domacnos , Soukroma inst , NONE_Duchodce , Verejna inst , NONE_Podnikatel , NONE_Rentier) + 1,6462 * WOE_EmployerType (Finance) + 0,1833 * WOE_PersonWithNoInc (PersonWithNoInc = 0) + 0,8048 * WOE_PersonWithNoInc (PersonWithNoInc > 1) + 0,1999 * WOE_CurrentEmpSince (12 < CurrentEmpSince <= 27) + 0,2007 * WOE_CurrentEmpSince (27 < CurrentEmpSince <= 51) + 0,3485 * WOE_CurrentEmpSince (4 < CurrentEmpSince <= 12) + 0,5611 * WOE_CurrentEmpSince (88 < CurrentEmpSince <= 101) + 0,5828 * WOE_CurrentEmpSince (51 < CurrentEmpSince <= 88) + 0,6888 * WOE_CurrentEmpSince (101 < CurrentEmpSince <= 103) + 0,8994 * WOE_CurrentEmpSince (103 < CurrentEmpSince <= 180) + 1,3758 * WOE_CurrentEmpSince (CurrentEmpSince > 180)**]** / **[**1 + exp(-1,0062 + 1,5878 * WOE_Product_Type (HL , CO) + 0,5881 * WOE_HousingStatus (private property) + 0,1530 * WOE_WorkingPosition (manual_indefinite , NONE_Duchodce , NONE_Student) + 0,4675 * WOE_WorkingPosition (intellectual_definite ,

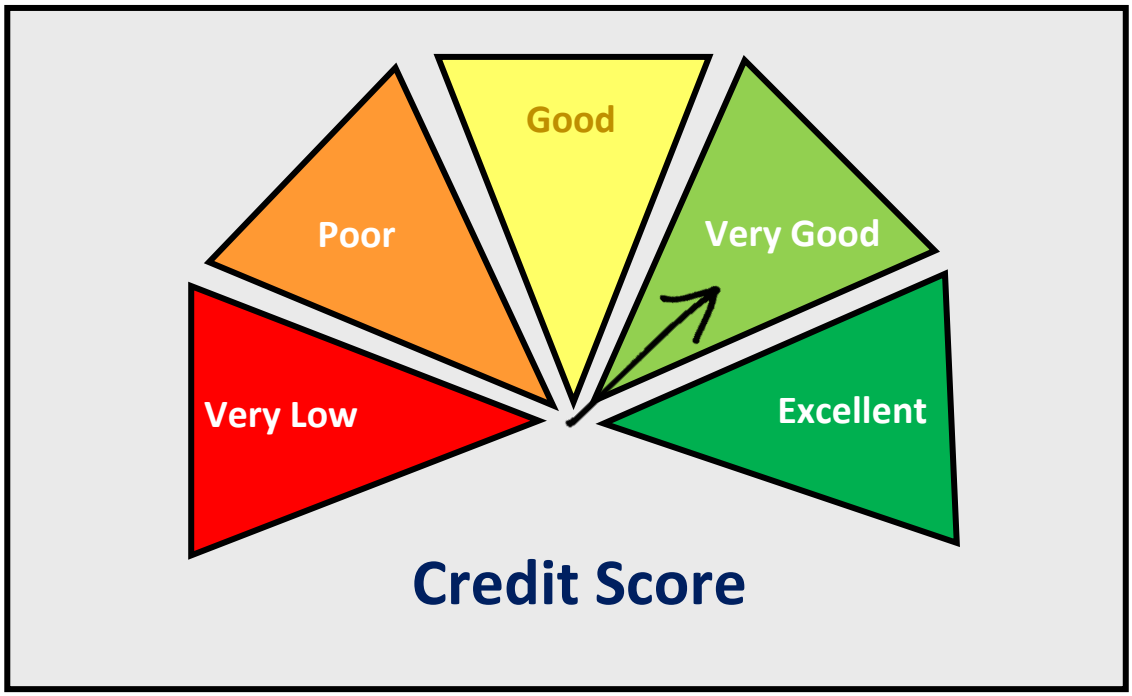
manager_definite , NONE_Rentier/Jine , intellectual_indefinite , manager_indefinite , NONE_Podnikatel) + 0,9288 * WOE_Education (VS) + 0,4393 * WOE_MaritalStatus (Marriage , Reg_partner) + 0,4287 * WOE_CommonEquity (Yes) + 0,9203 * WOE_InsurancePay (0 < InsurancePay <= 500) + 1,0991 * WOE_InsurancePay (InsurancePay > 500) + 0,2360 * WOE_Age (50 < Age <= 59) + 0,2426 * WOE_Age (Age <= 21) + 0,3202 * WOE_Age (39 < Age <= 44) + 0,5656 * WOE_Age (21 < Age <= 25) + 0,6049 * WOE_Age (34 < Age <= 39) + 0,6165 * WOE_Age (Age > 59) + 0,6403 * WOE_Age (25 < Age <= 29) + 0,8487 * WOE_Age (29 < Age <= 34) + 0,3944 * WOE_RegionPerm (CZ_Jihocesky kraj , CZ_Zlinsky kraj , CZ_Plzensky kraj , CZ_Pardubicky kraj , CZ_Vysocina , SK , CZ_Stredocesky kraj , CZ_Olomoucky kraj , CZ_Moravskoslezsky kraj , CZ_Liberecky kraj , CZ_Hlavni město) + 1,1433 * WOE_RegionPerm (CZ_Kralovehradecky kraj , CZ_Karlovarsky kraj) + 0,5340 * WOE_DifferentAdress (0) + 0,4594 * WOE_EmployerType (NONE_Student , Other , NONE_V domacnos , Soukroma inst , NONE_Duchodce , Verejna inst , NONE_Podnikatel , NONE_Rentier) + 1,6462 * WOE_EmployerType (Finance) + 0,1833 * WOE_PersonWithNoInc (PersonWithNoInc = 0) + 0,8048 * WOE_PersonWithNoInc (PersonWithNoInc > 1) + 0,1999 * WOE_CurrentEmpSince (12 < CurrentEmpSince <= 27) + 0,2007 * WOE_CurrentEmpSince (27 < CurrentEmpSince <= 51) + 0,3485 * WOE_CurrentEmpSince (4 < CurrentEmpSince <= 12) + 0,5611 * WOE_CurrentEmpSince (88 < CurrentEmpSince <= 101) + 0,5828 * WOE_CurrentEmpSince (51 < CurrentEmpSince <= 88) + 0,6888 * WOE_CurrentEmpSince (101 < CurrentEmpSince <= 103) + 0,8994 * WOE_CurrentEmpSince (103 < CurrentEmpSince <= 180) + 1,3758 * WOE_CurrentEmpSince (CurrentEmpSince > 180)**]**

5.2.16 Final Score

Součet výsledných hodnot skóre jednotlivých modelů, kterými klient prošel

5.2.17 Cluster analysis (Shluková analýza)

Po součtu výsledného skóre klient dále postupuje do zařazení do určité skupiny – clusteru klientů, kteří si jsou vzájemně podobní svým chováním. Banky mohou mít několik kategorií podle výsledků. Často se klienti dělí na čtyři nebo pět kategorií dle chování od *Very Low (velmi špatné chování)* až po *Excellent (excelentní chování).*



49) Obrázek– ukázka kategorií. Zdroj: vlastní zpracování

5.2.18 Limit of the Loan

V praxi se jednotlivým clusterům určí maximální částka, kolik se dá klientům v této kategorii půjčit. Dále se postupuje od nejvyšší hodnoty v clusteru a odečítá se při každém testu 10 000 Kč, dokud nevyjde výsledná maximální částka úvěru, která může být klientovi poskytnuta.

5.3 Kreditní model

Tvorba kreditního modelu bude postupovat ve stejných krocích, jako práce s datasetem aplikačních dat. Vypsány budou pouze nejdůležitější kroky, výsledky a fakta.

5.3.1 Data understanding

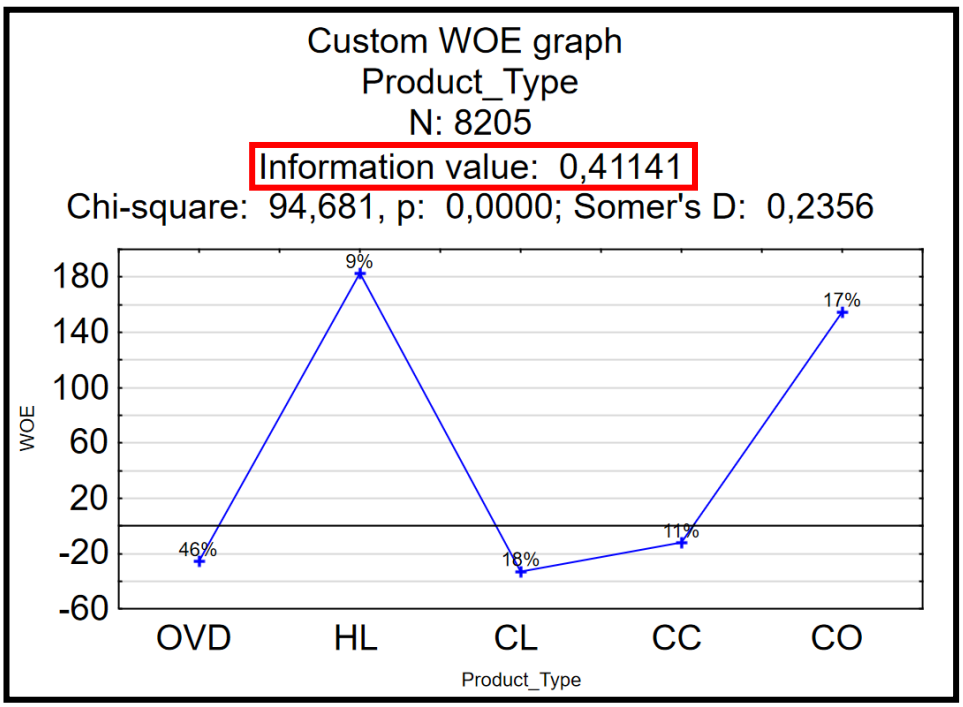
Datový soubor kreditního modelu obsahuje data od cca 11 000 klientů. Sbírá se 96 ks proměnných, jako jsou například *NumOfClosProd (počet ukončených produktů), Product_Type (typ produktu), NumOfExProd (počet existujících produktů), nebo Avg_KU (průměrná výše kontokorentů).* V celém souboru se vyskytují proměnné *spojité (continuous).* Kategoriální je pouze jedna proměnná, a to *Product_Type*. Vysvětlovaná proměnná *y* je *Dependent_12M.* Jedná se o vektor binárních proměnných (tj. nabývající hodnot 0 nebo 1). Dobré účty (Goods) získávají hodnoty 0 a špatné (Bads) hodnoty 1. Ostatní proměnné x_i jsou vysvětlující. Poměr dobrých a špatných účtů je 10 450 Goods : 550 Bads.

5.3.2 Splitting data

Soubor byl rozdělen na dva vzorky – testovací a trénovací. Trénovací obsahuje 8 200 klientů a testovací cca 2 800 klientů.

5.3.3 Intuitive Behaviour – Weight of Evidence

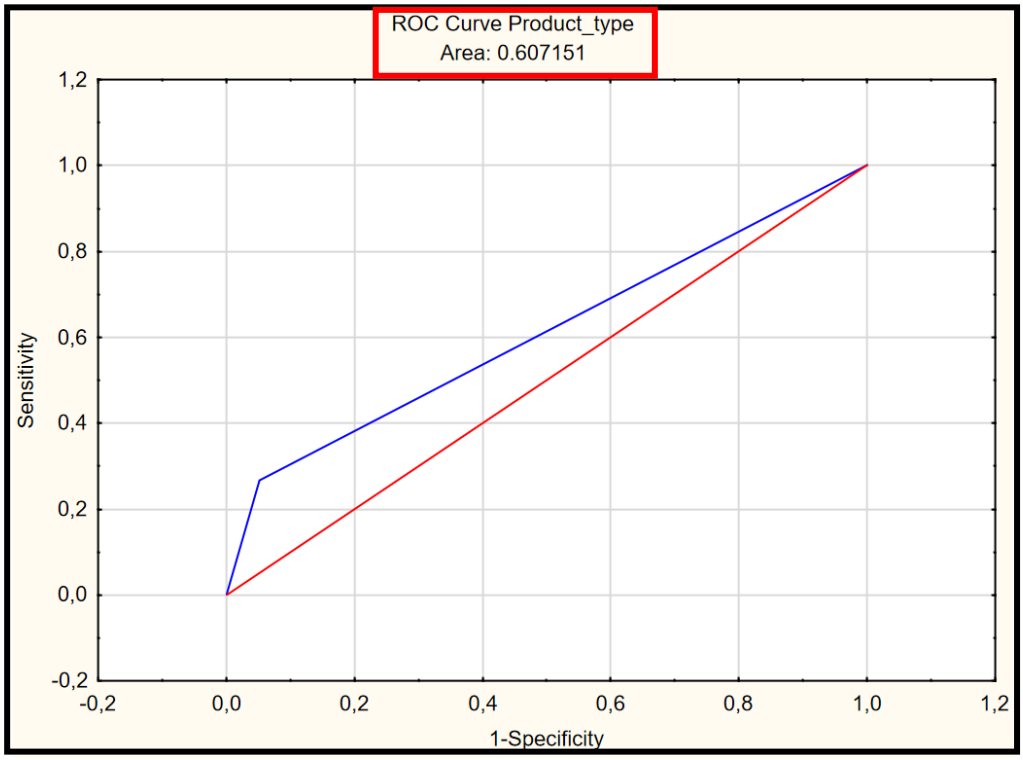
Proměnná s nejvyšší hodnotou Information Value z celého souboru je *Product_Type.* Její hodnoty a chování je vidět v následujícím grafu: Intuitive Behaviour – Weight of Evidence



50) Graf – WOE graf pro proměnnou Product_Type. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

Power statistic

Následující ROC křivka znázorňuje sílu proměnné *Product_Type*, která získala nejvyšší hodnotu Information Value v souboru kreditních dat:



51) Graf – ROC křivka pro proměnnou Product_Type. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.3.4 Long list of variables

V následující tabulce jsou uvedené proměnny z Long listu, které dále vstupují do testování:

Long list of variables	
Variable name	**Meaning of the variable**
Product_Type	Typ produktu
NumOfClosProd	Počet uzavřených produktů poslední dva roky
NumOfPreMatProd	Počet předčasně ukončených produktů
Avg_KU	Průměrná výše kontokorentů
NumOfExProd	Počet existujících produktů
NumOfExHU	Počet existujích hypoték
NumOfRejProd_off_us	Počet zamítnutých žádostí, kromě dané společnosti
Avg_KU_ex	Průměrná výše existujících kontokorentů
NumOfExProd_RoleS	Počet existujících produktů v roli spolužadatele
NumOfExProd_360_720	Počet existujících produktů otevřených před 1 nebo 2 lety
NumOfExProd_on_us	Počet existujících produktů v dané společnosti
CL_limit	Součet limitů s úrokem - spotřebitelské úvěry
NumOfProd_3M_off_us	Celkový počet produktů - poslední 3 měsíce, kromě dané společnosti
NumOfExKK	Počet otevřených kreditních karet
NumOfCL_3M	Součet limitů produktů poslední 3 měsíce
NumOfProd_2y	Celkový počet žádostí poslední 2 roky
NumOfExHU_RoleS	Počet existujících hypoték v roli spolužadatele
NumOfClosCL_2y	Celkový počet zavřených produktů poslední dva roky
CL_ex	Počet existujících produktů - spotebitelské úvěry
NumOfProd_1y	Celkový počet produktů/žádostí poslední 1 rok

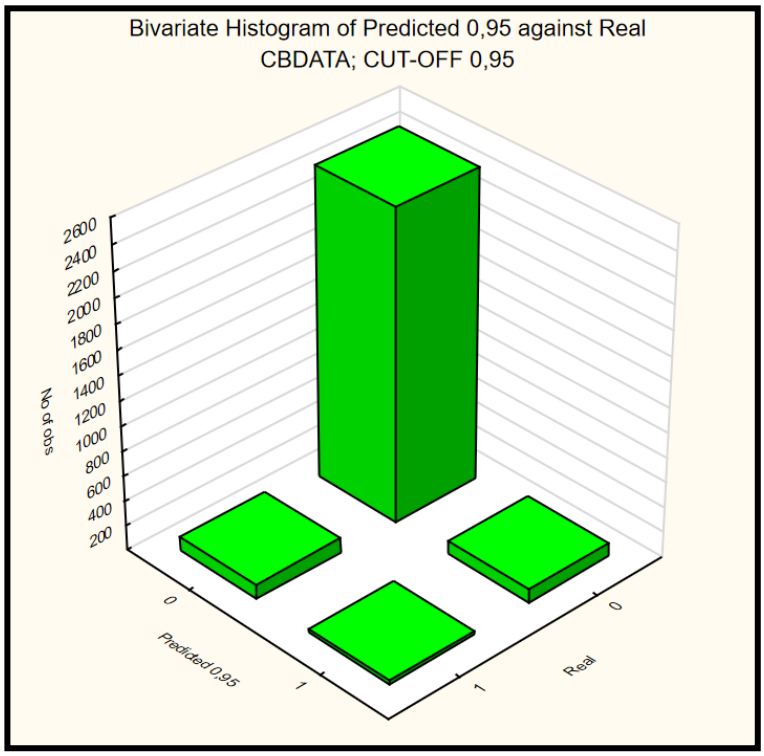
52) Tabulka – Long list of variables CBDATA. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.3.5 Výsledky modelu

Na základě testování je zvolen finální model, u kterého byla nastavena hladina cut-off na 0,95. Model byl otestován na základě Cross-Validace. Výsledky jsou uvedeny v následující tabulce a histogramu:

	Summary Frequency Table CBDATA, CUT-OFF 0,95			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**2491**	105	2596
Total Percent		**90,88%**	3,83%	94,71%
Count	1	115	**30**	145
Total Percent		4,20%	**1,09%**	5,29%
Count	All Grps	2606	135	2741
Total Percent		95,07%	4,93%	

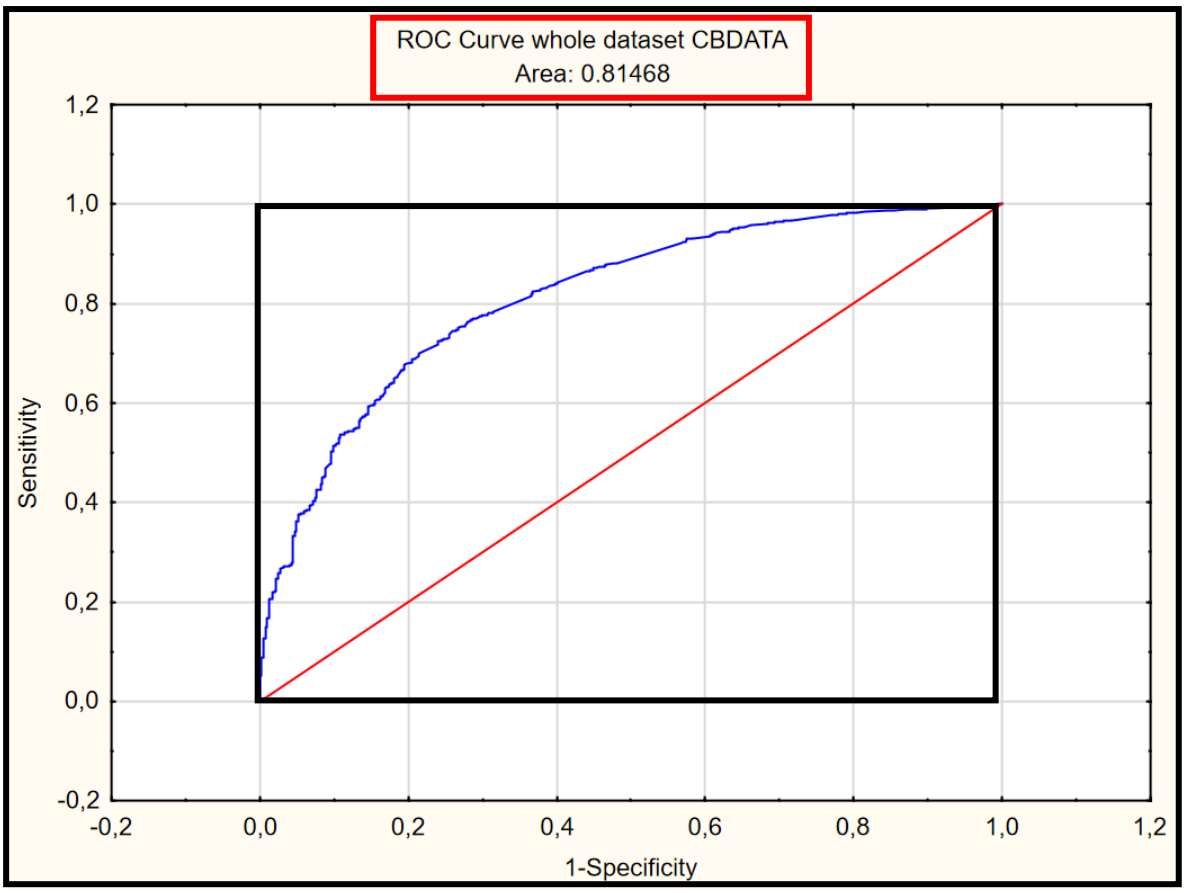
53) Tabulka – Výsledky testu Cross-Validace kreditního modelu. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování



54) Graf – Výsledky testu Cross-Validace na kreditních datech. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.3.6 ROC křivka celého modelu

Následující graf zobrazuje sílu výsledného modelu pomocí ROC křivky. AUC dosáhla hodnoty 0,8147.



55) Graf – ROC křivka pro proměnnou Product_Type. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.3.7 Výsledná scorekarta kreditního modelu:

CREDIT BUREAU SCORECARD			
#	**Variable**	**Cathegory**	**Estimate**
0		Intercept	-2,8197
1	WOE_Product_Type	OVD , CL , CC	0
2		HL , CO	1,7792
3	WOE_NumOfClosProd	NumOfClosProd = 0	0
4		0 < NumOfClosProd <= 1	0,4166
5		2 < NumOfClosProd <= 4	0,8949
6		1 < NumOfClosProd <= 2	0,9838
7		NumOfClosProd > 4	1,2202
8	WOE_NumOfPreMatProd	NumOfPreMatProd = 0	0
9		0 < NumOfPreMatProd <= 1	0,5784
10		1 < NumOfPreMatProd <= 3	1,1380
11		NumOfPreMatProd > 3	1,3049
12	WOE_Avg_KU	0 < Avg_KU <= 5000	0
13		10000 < Avg_KU <= 20000	0,4768
14		Avg_KU = 0	0,4837
15		5000 < Avg_KU <= 10000	0,5712
16		Avg_KU > 20000	1,1067
17	WOE_NumOfExProd	NumOfExProd = 0	0
18		0 < NumOfExProd <= 1	0,7957
19		2 < NumOfExProd <= 3	0,9610
20		1 < NumOfExProd <= 2	1,1430
21		NumOfExProd > 5	1,1933
22		3 < NumOfExProd <= 5	1,3587
23	WOE_NumOfExHU	NumOfExHU = 0	0
24		NumOfExHU > 0	0,8842
25	WOE_NumOfRejProd_off_us	NumOfRejProd_off_us > 1	0
26		0 < NumOfRejProd_off_us <= 1	0,7232
27		NumOfRejProd_off_us = 0	1,4568
28	WOE_NumOfClosCL_2y	NumOfClosCL_2y = 0	0
29		NumOfClosCL_2y > 0	0,9049
30	WOE_NumOfProd_2y	NumOfProd_2y > 4	0
31		1 < NumOfProd_2y <= 2	0,8666
32		2 < NumOfProd_2y <= 4	1,0253
33		0 < NumOfProd_2y <= 1	1,1436
34		NumOfProd_2y = 0	1,6863
35	WOE_NumOfCL_3M	NumOfCL_3M > 0	0
36		NumOfCL_3M = 0	0,5280
37	WOE_NumOfExProd_on_us	NumOfExProd_on_us = 0	0
38		NumOfExProd_on_us > 0	0,5819
39	WOE_NumOfProd_3M_off_us	NumOfProd_3M_off_us > 0	0
40		NumOfProd_3M_off_us = 0	0,7715

56) Tabulka: scorekarta kreditního modelu. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.3.8 Výsledná rovnice logistické regrese a výpočtu pravděpodobnosti splacení kreditního modelu:

***P* = [**exp(-2,8197 + 1,7792 * WOE_Product_Type (HL , CO) + 0,4166 * WOE_NumOfClosProd (0 < NumOfClosProd <= 1) + 0,8949 * WOE_NumOfClosProd (2 < NumOfClosProd <= 4) + 0,9838 * WOE_NumOfClosProd (1 < NumOfClosProd <= 2) + 1,2202 * WOE_NumOfClosProd (NumOfClosProd > 4) + 0,5784 * WOE_NumOfPreMatProd (0 < NumOfPreMatProd <= 1) + 1,1380 * WOE_NumOfPreMatProd (1 < NumOfPreMatProd <= 3) + 1,3049 * WOE_NumOfPreMatProd (NumOfPreMatProd > 3) + 0,4768 * WOE_Avg_KU (10000 < Avg_KU <= 20000) + 0,4837 * WOE_Avg_KU (Avg_KU <= 0) + 0,5712 * WOE_Avg_KU (5000 < Avg_KU <= 10000) + 1,1067 * WOE_Avg_KU (Avg_KU > 20000) + 0,7957 * WOE_NumOfExProd (0 < NumOfExProd <= 1) + 0,9610 * WOE_NumOfExProd (2 < NumOfExProd <= 3) + 1,1430 * WOE_NumOfExProd (1 < NumOfExProd <= 2) + 1,1933 * WOE_NumOfExProd (NumOfExProd > 5) * 1,3587 * WOE_NumOfExProd (3 < NumOfExProd <= 5) + 0,8842 * WOE_NumOfExHU (NumOfExHU > 0) + 0,7232 * WOE_NumOfRejProd_off_us (0 < NumOfRejProd_off_us <= 1) + 1,4568 * WOE_NumOfRejProd_off_us (NumOfRejProd_off_us = 0) + 0,9049 * WOE_NumOfClosCL_2y (NumOfClosCL_2y > 0) + 0,8666 * WOE_NumOfProd_2y (1 < NumOfProd_2y <= 2) + 1,0253 * WOE_NumOfProd_2y(2 < NumOfProd_2y <= 4) + 1,1436 * WOE_NumOfProd_2y (0 < NumOfProd_2y <= 1) + 1,6863 * WOE_NumOfProd_2y (NumOfProd_2y = 0) + 0,5280 * WOE_NumOfCL_3M (NumOfCL_3M = 0) + 0,5819 * WOE_NumOfCL_3M (NumOfExProd_on_us > 0) + 0,7715 * WOE_NumOfProd_3M_off_us (NumOfProd_3M_off_us = 0)**] / [**(1+ exp(-2,8197 + 1,7792 * WOE_Product_Type (HL , CO) + 0,4166 * WOE_NumOfClosProd (0 < NumOfClosProd <= 1) + 0,8949 * WOE_NumOfClosProd (2 < NumOfClosProd <= 4) + 0,9838 * WOE_NumOfClosProd (1 < NumOfClosProd <= 2) + 1,2202 * WOE_NumOfClosProd (NumOfClosProd > 4) + 0,5784 * WOE_NumOfPreMatProd (0 < NumOfPreMatProd <= 1) + 1,1380 * WOE_NumOfPreMatProd (1 < NumOfPreMatProd <= 3) + 1,3049 * WOE_NumOfPreMatProd (NumOfPreMatProd > 3) + 0,4768 * WOE_Avg_KU (10000 < Avg_KU <= 20000) + 0,4837 * WOE_Avg_KU (Avg_KU <= 0) + 0,5712 * WOE_Avg_KU (5000 < Avg_KU <= 10000) + 1,1067 * WOE_Avg_KU (Avg_KU > 20000) + 0,7957 * WOE_NumOfExProd (0 < NumOfExProd <= 1) + 0,9610 * WOE_NumOfExProd (2 < NumOfExProd <= 3) + 1,1430 * WOE_NumOfExProd (1 < NumOfExProd <= 2) + 1,1933 * WOE_NumOfExProd (NumOfExProd > 5) * 1,3587 * WOE_NumOfExProd (3 < NumOfExProd <= 5) + 0,8842 * WOE_NumOfExHU (NumOfExHU > 0) +

+ 0,7232 * WOE_NumOfRejProd_off_us (0 < NumOfRejProd_off_us <= 1) + 1,4568 * WOE_NumOfRejProd_off_us (NumOfRejProd_off_us = 0) + 0,9049 * WOE_NumOfClosCL_2y (NumOfClosCL_2y > 0) + 0,8666 * WOE_NumOfProd_2y (1 < NumOfProd_2y <= 2) + 1,0253 * WOE_NumOfProd_2y(2 < NumOfProd_2y <= 4) + 1,1436 * WOE_NumOfProd_2y (0 < NumOfProd_2y <= 1) + 1,6863 * WOE_NumOfProd_2y (NumOfProd_2y = 0) + 0,5280 * WOE_NumOfCL_3M (NumOfCL_3M = 0) + 0,5819 * WOE_NumOfCL_3M (NumOfExProd_on_us > 0) + 0,7715 * WOE_NumOfProd_3M_off_us (NumOfProd_3M_off_us = 0)**]**

5.4. Behaviorální model

Tvorba behaviorálního modelu bude postupovat ve stejných krocích, jako práce s datasetem aplikačních dat. Vypsány budou pouze nejdůležitější kroky, výsledky a fakta.

5.4.1 Data understanding

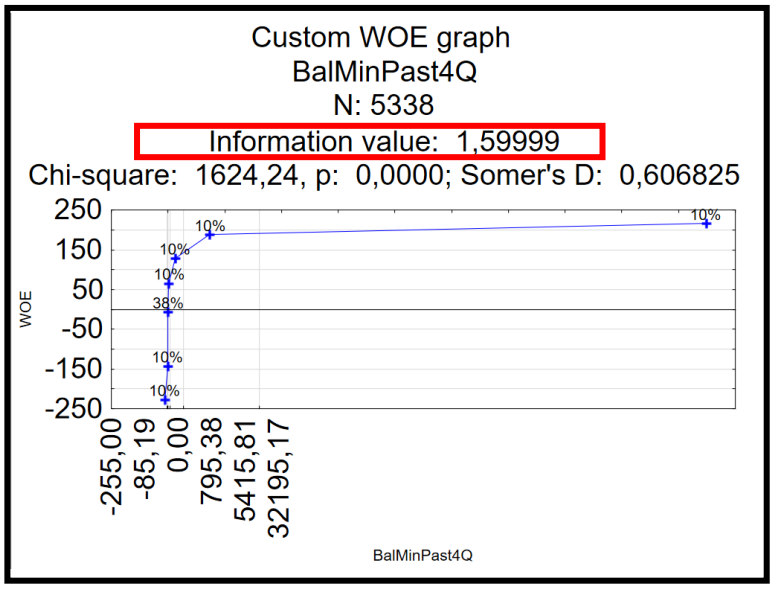
Datový soubor behaviorálního modelu čítá data od cca 8 000 klientů. Sbírá se cca 300 - 400 proměnných, jako jsou například *NumOfExProd (počet existujících produktů), nebo Avg_KU (průměrná výše komtokorentů).* V celém souboru se vyskytují pouze proměnné *spojité*. Vysvětlovaná proměnna *y* je *Dependent_12M.* Jedná se o vektor binárních proměnných (tj. nabývající hodnot 0 nebo 1). Dobré účty (Goods) získávájí hodnoty 0 a špatné (Bads) hodnoty 1. Ostatní proměnné x_i jsou vysvětlující. Poměr dobrých a špatných účtů je cca 5 400 Goods : 2 600 Bads.

5.4.2 Splitting data

Soubor byl rozdělen na dva vzorky – testovací a trénovací. Trénovací obsahuje 6 200 klientů a testovací cca 1 800 klientů.

5.4.3 Intuitive Behaviour – Weight of Evidence

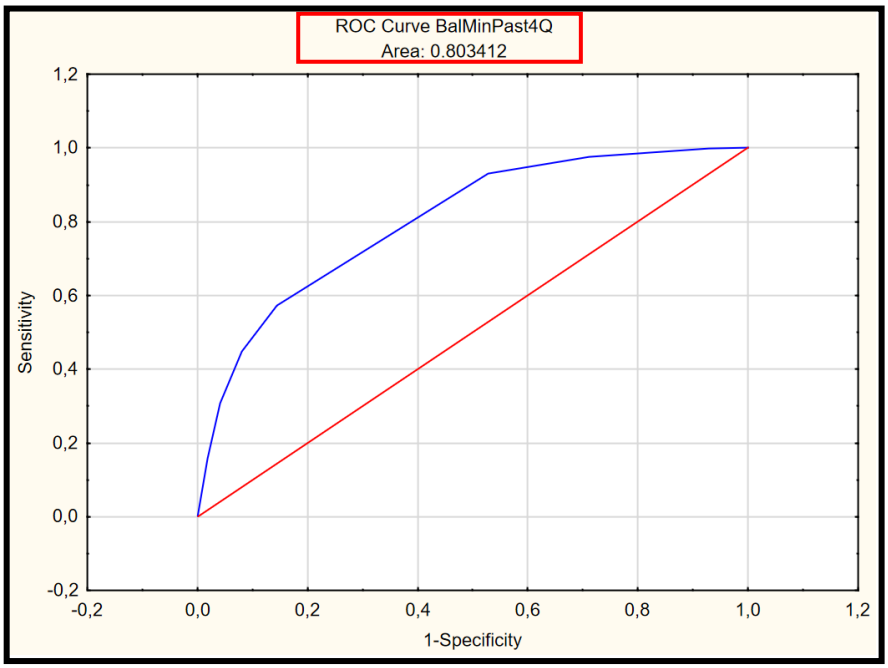
Proměnná s nejvyšší hodnotou Information Value z celého souboru je *BalMinPast4Q*. Její hodnoty a chování je vidět v následujícím grafu:



57) Graf – WOE graf pro proměnnou BalMinPast4Q. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.4.4 Predictive Power

Následující ROC křivka znázorňuje sílu proměnné *BalMinPast4Q*, která získala nejvyšší hodnotu Information Value v souboru behaviorálních dat:



58) Graf – ROC křivka BalMinPast4Q. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.4.5 Výsledky modelu

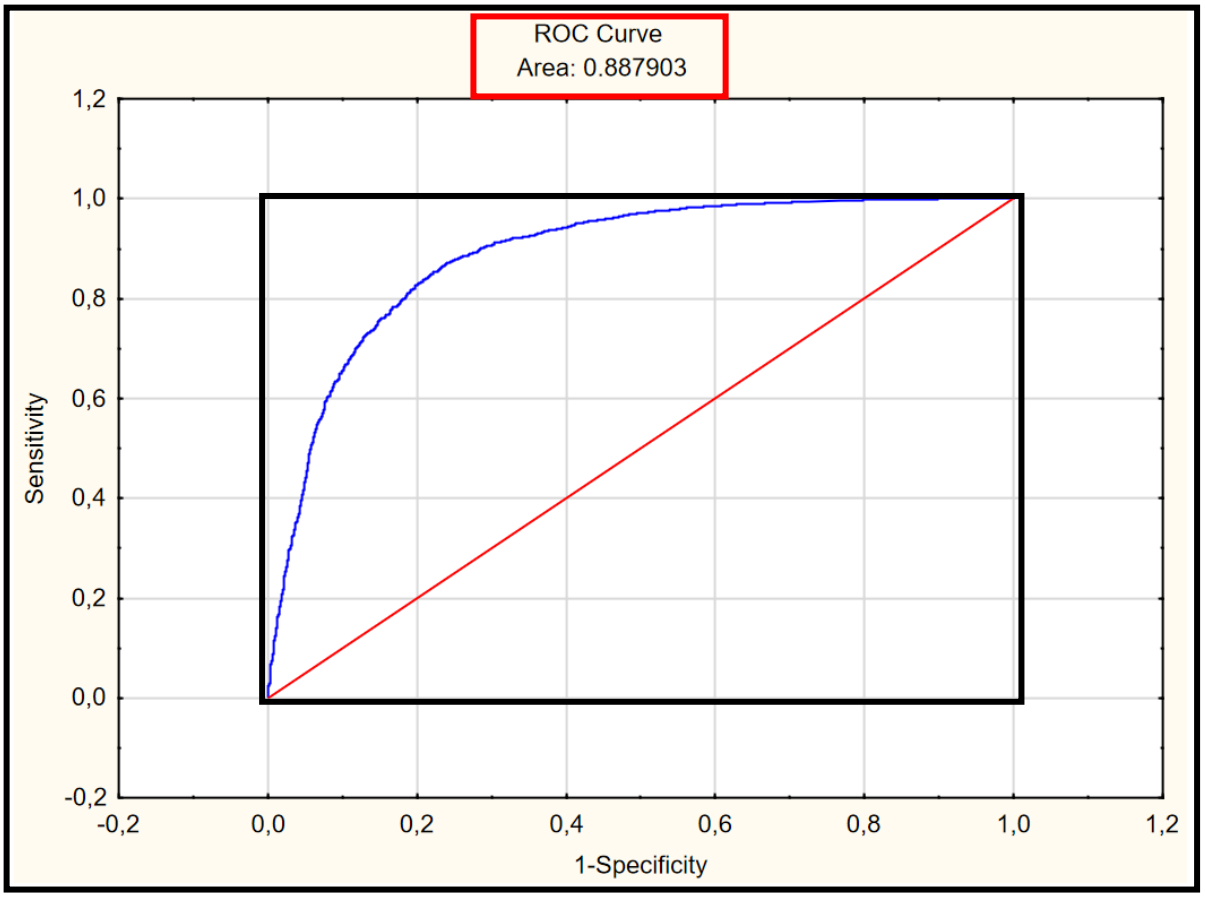
Na základě testování je zvolen finální model, u kterého byla nastavena hladina cut-off na 0,95. Model byl otestován na základě Cross-Validace. Výsledky jsou uvedeny v následující tabulce a histogramu:

	Summary Frequency Table BSDATA, CUT-OFF 0,5			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	**959**	96	1055
Total Percent		**53,67%**	5,37%	59,04%
Count	1	100	**632**	732
Total Percent		5,6%	**35,37%**	40,96%
Count	All Grps	1059	728	1787
Total Percent		59,26%	40,74%	

59) Tabulka – Výsledky testu Cross-Validace behaviorálních dat. Výstup z programu STATISTICA 13. Zdroj: vlastní zpracování

5.4.6 ROC křivka celého modelu

Následující graf zobrazuje sílu výsledného modelu pomocí ROC křivky. AUC dosáhla hodnoty 0,8879.



60) Graf – ROC křivka výsledného modelu behaviorálních dat. Výstup z programu STATISTICA13. Zdroj: vlastní zpracování

5.4.7 Výsledná scorekarta behaviorálního modelu, č. 61, zdroj: vlastní zpracování

BEHAVIORAL SCORECARD			
#	Variable	Cathegory	Estimate
0		Intercept	-6,9399
1	WOE_BalMinPast4Q	BalMinPast4Q <= -255	0
2		-255 < BalMinPast4Q <= -85,19	0,1114
3		-85,19 < BalMinPast4Q <= 0	0,3592
4		5415,81 < BalMinPast4Q <= 32195,2	0,4223
5		795,38 < BalMinPast4Q <= 5415,81	0,5719
6		0 < BalMinPast4Q <= 795,38	0,6000
7		BalMinPast4Q > 32195,2	0,6402
8	WOE_DebtMonthCntPast4Q	DebtMonthCntPast4Q > 6	0
9		2 < DebtMonthCntPast4Q <= 6	0,8615
10		0 < DebtMonthCntPast4Q <= 2	1,3409
11		DebtMonthCntPast4Q <= 0	1,6094
12	WOE_BalAct	BalAct <= -30	0
13		-30 < BalAct <= 0	0,3063
14		0 < BalAct <= 552,48	0,4296
15		6610,5 < BalAct <= 16810,3	0,5698
16		552,48 < BalAct <= 2097,18	0,6105
17		2097,18 < BalAct <= 6610,5	0,6287
18		16810,3 < BalAct <= 45201,8	0,8160
19		45201,8 < BalAct <= 171943	0,9169
20		BalAct > 171943	1,4551
21	WOE_ODDelqStatMaxPast4Q	ODDelqStatMaxPast4Q > 1	0
22		0 < ODDelqStatMaxPast4Q <= 1	1,5758
23		ODDelqStatMaxPast4Q <= 0	2,5122
24	WOE_OutsideDeposMin_6M	0 < OutsideDeposMin_6M <= 6000	0
25		OutsideDeposMin_6M <= 0	0,0197
26		6000 < OutsideDeposMin_6M <= 10509	0,2110
27		10509 < OutsideDeposMin_6M <= 13241	0,3428
28		16338 < OutsideDeposMin_6M <= 20304	0,3565
29		13241 < OutsideDeposMin_6M <= 16338	0,4018
30		OutsideDeposMin_6M > 37961	0,5382
31		20304 < OutsideDeposMin_6M <= 25561	0,8784
32		25561 < OutsideDeposMin_6M <= 37961	0,9851
33	WOE_StandOrderCnt_6M	StandOrderCnt_6M = 0	0
34		0 < StandOrderCnt_6M <= 6	0,0367
35		12 < StandOrderCnt_6M <= 21	0,2256
36		6 < StandOrderCnt_6M <= 12	0,3184
37		21 < StandOrderCnt_6M <= 32	0,7400

38		StandOrderCnt_6M > 32	0,9218
39	WOE_EBWithdCnt_6M	EBWithdCnt_6M = 0	0
40		EBWithdCnt_6M > 55	0,0258
41		39 < EBWithdCnt_6M <= 55	0,2819
42		0 < EBWithdCnt_6M <= 7	0,3115
43		17 < EBWithdCnt_6M <= 27	0,4254
44		27 < EBWithdCnt_6M <= 39	0,5543
45		7 < EBWithdCnt_6M <= 17	0,5801
46	WOE_ODDebtPast	ODDebtPast > 0	0
47		ODDebtPast <= 0	0,3753
48	WOE_OutsideDeposCnt_6M	OutsideDeposCnt_6M > 24	0
49		8 < OutsideDeposCnt_6M <= 11	0,1996
50		14 < OutsideDeposCnt_6M <= 18	0,3347
51		11 < OutsideDeposCnt_6M <= 14	0,3498
52		6 < OutsideDeposCnt_6M <= 8	0,3832
53		18 < OutsideDeposCnt_6M <= 24	0,3894
54		3 < OutsideDeposCnt_6M <= 6	0,4129
55		OutsideDeposCnt_6M <= 3	0,4196
56	WOE_SinceOpenLiveMax	SinceOpenLiveMax <= 38	0
57		204 < SinceOpenLiveMax <= 239	0,2463
58		181 < SinceOpenLiveMax <= 204	0,2693
59		38 < SinceOpenLiveMax <= 64	0,3435
60		128 < SinceOpenLiveMax <= 156	0,3721
61		SinceOpenLiveMax > 239	0,3983
62		156 < SinceOpenLiveMax <= 181	0,3984
63		99 < SinceOpenLiveMax <= 128	0,4020
64		64 < SinceOpenLiveMax <= 99	0,5303
65	WOE_PensionCnt_12M	PensionCnt_12M = 0	0
66		PensionCnt_12M > 0	0,1938
67	WOE_ODUtilSumPast4Q	29793,6 < ODUtilSumPast4Q <= 57672,1	0
68		0 < ODUtilSumPast4Q <= 8421,28	0,0157
69		8421,28 < ODUtilSumPast4Q <= 29793,6	0,0225
70		ODUtilSumPast4Q > 105614	0,0758
71		57672,1 < ODUtilSumPast4Q <= 105614	0,2822
72		ODUtilSumPast4Q <= 0	0,4344
73	WOE_DirectDebitCnt_6M	DirectDebitCnt_6M <= 0	0
74		0 < DirectDebitCnt_6M <= 6	0,2357
75		DirectDebitCnt_6M > 13	0,2974
76		6 < DirectDebitCnt_6M <= 13	0,4514
77	WOE_CardTotalCntAvg_6M	CardTotalCntAvg_6M <= 0,17	0

78		3,17 < CardTotalCntAvg_6M <= 5	0,3620
79		CardTotalCntAvg_6M > 21,5	0,4219
80		0,17 < CardTotalCntAvg_6M <= 1,67	0,4375
81		14,67 < CardTotalCntAvg_6M <= 21,5	0,4769
82		5 < CardTotalCntAvg_6M <= 7,33	0,5261
83		1,67 < CardTotalCntAvg_6M <= 3,17	0,5317
84		10,33 < CardTotalCntAvg_6M <= 14,67	0,6328
85		7,33 < CardTotalCntAvg_6M <= 10,33	0,7025
86	WOE_OutsideDeposFlg_6M	OutsideDeposFlg_6M <= 2	0
87		2 < OutsideDeposFlg_6M <= 5	0,1788
88		OutsideDeposFlg_6M > 5	0,6128

5.3.8 Výsledná rovnice logistické regrese behaviorálních dat:

P **=** **[**exp(-6,9399 + 0,1114 * WOE_BalMinPast4Q (-255 < BalMinPast4Q <= -85,19) + 0,3592 * WOE_BalMinPast4Q (-85,19 < BalMinPast4Q <= 0) + 0,4223 * WOE_BalMinPast4Q (BalMinPast4Q > 32195,2) + 0,5719 * WOE_BalMinPast4Q (795,38 < BalMinPast4Q <= 5415,81) + 0,6000 * WOE_BalMinPast4Q (0 < BalMinPast4Q <= 795,38) + 0,6402 * WOE_BalMinPast4Q (5415,81 < BalMinPast4Q <= 32195,2) + 0,8615 * WOE_DebtMonthCntPast4Q (2 < DebtMonthCntPast4Q <= 6) + 1,3409 * WOE_DebtMonthCntPast4Q (0 < DebtMonthCntPast4Q <= 2) + 1,6094 * WOE_DebtMonthCntPast4Q (DebtMonthCntPast4Q <= 0) + 0,3063 * WOE_BalAct (-30 < BalAct <= 0) + 0,4296 * WOE_BalAct (0 < BalAct <= 552,48) + 0,5698 * WOE_BalAct (6610,5 < BalAct <= 16810,3) + 0,6105 * WOE_BalAct (552,48 < BalAct <= 2097,18) + 0,6287 * WOE_BalAct (2097,18 < BalAct <= 6610,5) + 0,8160 * WOE_BalAct (16810,3 < BalAct <= 45201,8) + 0,9169 * WOE_BalAct (45201,8 < BalAct <= 171943) + 1,4551 * WOE_BalAct (BalAct > 171943) + 1,5758 * WOE_ODDelqStatMaxPast4Q (0 < ODDelqStatMaxPast4Q <= 1) + 2,5122 * WOE_ODDelqStatMaxPast4Q (ODDelqStatMaxPast4Q <= 0) + 0,0197 * WOE_OutsideDeposMin_6M (OutsideDeposMin_6M <= 0) + 0,2110 * WOE_OutsideDeposMin_6M (6000 < OutsideDeposMin_6M <= 10509) + 0,3428 * WOE_OutsideDeposMin_6M (16338 < OutsideDeposMin_6M <= 20304) + 0,4018 * WOE_OutsideDeposMin_6M (13241 < OutsideDeposMin_6M <= 16338) + 0,5382 * WOE_OutsideDeposMin_6M (OutsideDeposMin_6M > 37961) + 0,8784 *

* WOE_OutsideDeposMin_6M (20304 < OutsideDeposMin_6M <= 25561) + 0,9851 * WOE_OutsideDeposMin_6M (25561 < OutsideDeposMin_6M <= 37961) + 0,0367 * WOE_StandOrderCnt_6M (0 < StandOrderCnt_6M <= 6) + 0,2256 * WOE_StandOrderCnt_6M (12 < StandOrderCnt_6M <= 21) + 0,3184 * WOE_StandOrderCnt_6M (6 < StandOrderCnt_6M <= 12) + 0,7400 * WOE_StandOrderCnt_6M (21 < StandOrderCnt_6M <= 32) + 0,9218 * WOE_StandOrderCnt_6M (StandOrderCnt_6M > 32) + 0,0258 * WOE_EBWithdCnt_6M (EBWithdCnt_6M > 55) + 0,2819 * WOE_EBWithdCnt_6M (39 < EBWithdCnt_6M <= 55) + 0,3115 * WOE_EBWithdCnt_6M (0 < EBWithdCnt_6M <= 7) + 0,4254 * WOE_EBWithdCnt_6M (17 < EBWithdCnt_6M <= 27) + 0,5543 * WOE_EBWithdCnt_6M (27 < EBWithdCnt_6M <= 39) + 0,5801 * EBWithdCnt_6M (7 < EBWithdCnt_6M <= 17) + 0,3753 * WOE_ODDebtPast (ODDebtPast <= 0) + 0,1996 * WOE_OutsideDeposCnt_6M (8 < OutsideDeposCnt_6M <= 11) + 0,3347 * WOE_OutsideDeposCnt_6M (14 < OutsideDeposCnt_6M <= 18) + 0,3498 * WOE_OutsideDeposCnt_6M (11 < OutsideDeposCnt_6M <= 14) + 0,3832 * WOE_OutsideDeposCnt_6M (6 < OutsideDeposCnt_6M <= 8) + 0,3894 * WOE_OutsideDeposCnt_6M (18 < OutsideDeposCnt_6M <= 24) + 0,4129 * WOE_OutsideDeposCnt_6M (3 < OutsideDeposCnt_6M <= 6) + 0,4196 * WOE_OutsideDeposCnt_6M (OutsideDeposCnt_6M <= 3) + 0,2463 * WOE_SinceOpenLiveMax (204 < SinceOpenLiveMax <= 239) + 0,2693 * WOE_SinceOpenLiveMax (181 < SinceOpenLiveMax <= 204) + 0,3435 * WOE_SinceOpenLiveMax (38 < SinceOpenLiveMax <= 64) + 0,3721 * WOE_SinceOpenLiveMax (128 < SinceOpenLiveMax <= 156) + 0,3983 * WOE_SinceOpenLiveMax (SinceOpenLiveMax > 239) + 0,3984 * WOE_SinceOpenLiveMax (156 < SinceOpenLiveMax <= 181) + 0,4020 * WOE_SinceOpenLiveMax (99 < SinceOpenLiveMax <= 128) + 0,5303 * WOE_SinceOpenLiveMax (64 < SinceOpenLiveMax <= 99) + 0,1938 * WOE_PensionCnt_12M (PensionCnt_12M > 0) + 0,0157 * WOE_ODUtilSumPast4Q (0 < ODUtilSumPast4Q <= 8421,28) + 0,0225 * WOE_ODUtilSumPast4Q (8421,28 < ODUtilSumPast4Q <= 29793,6) + 0,0758 * WOE_ODUtilSumPast4Q (ODUtilSumPast4Q > 105614) + 0,2822 * WOE_ODUtilSumPast4Q (57672,1 < ODUtilSumPast4Q <= 105614) + 0,4344 * WOE_ODUtilSumPast4Q (ODUtilSumPast4Q <= 0) +

+ 0,3620 * WOE_CardTotalCntAvg_6M (3,17 < CardTotalCntAvg_6M <= 5) + 0,4219 * WOE_CardTotalCntAvg_6M (CardTotalCntAvg_6M > 21,5) + 0,4375 * WOE_CardTotalCntAvg_6M (0,17 < CardTotalCntAvg_6M <= 1,67) + 0,4769 * WOE_CardTotalCntAvg_6M (14,67 < CardTotalCntAvg_6M <= 21,5) + 0,5261 * WOE_CardTotalCntAvg_6M (5 < CardTotalCntAvg_6M <= 7,33) + 0,5317 * WOE_CardTotalCntAvg_6M (1,67 < CardTotalCntAvg_6M <= 3,17) + 0,6328 * WOE_CardTotalCntAvg_6M (10,33 < CardTotalCntAvg_6M <= 14,67) + 0,7025 * WOE_CardTotalCntAvg_6M (7,33 < CardTotalCntAvg_6M <= 10,33) + 0,1788 * WOE_OutsideDeposFlg_6M (2 < OutsideDeposFlg_6M <= 5) + 0,6128 * WOE_OutsideDeposFlg_6M (OutsideDeposFlg_6M > 5)**]** / **[**(1+ exp(-6,9399 + 0,1114 * WOE_BalMinPast4Q (-255 < BalMinPast4Q <= -85,19) + 0,3592 * WOE_BalMinPast4Q (-85,19 < BalMinPast4Q <= 0) + 0,4223 * WOE_BalMinPast4Q (BalMinPast4Q > 32195,2) + 0,5719 * WOE_BalMinPast4Q (795,38 < BalMinPast4Q <= 5415,81) + 0,6000 * WOE_BalMinPast4Q (0 < BalMinPast4Q <= 795,38) + 0,6402 * WOE_BalMinPast4Q (5415,81 < BalMinPast4Q <= 32195,2) + 0,8615 * WOE_DebtMonthCntPast4Q (2 < DebtMonthCntPast4Q <= 6) + 1,3409 * WOE_DebtMonthCntPast4Q (0 < DebtMonthCntPast4Q <= 2) + 1,6094 * WOE_DebtMonthCntPast4Q (DebtMonthCntPast4Q <= 0) + 0,3063 * WOE_BalAct (-30 < BalAct <= 0) + 0,4296 * WOE_BalAct (0 < BalAct <= 552,48) + 0,5698 * WOE_BalAct (6610,5 < BalAct <= 16810,3) + 0,6105 * WOE_BalAct (552,48 < BalAct <= 2097,18) + 0,6287 * WOE_BalAct (2097,18 < BalAct <= 6610,5) + 0,8160 * WOE_BalAct (16810,3 < BalAct <= 45201,8) + 0,9169 * WOE_BalAct (45201,8 < BalAct <= 171943) + 1,4551 * WOE_BalAct (BalAct > 171943) + 1,5758 * WOE_ODDelqStatMaxPast4Q (0 < ODDelqStatMaxPast4Q <= 1) + 2,5122 * WOE_ODDelqStatMaxPast4Q (ODDelqStatMaxPast4Q <= 0) + 0,0197 * WOE_OutsideDeposMin_6M (OutsideDeposMin_6M <= 0) + 0,2110 * WOE_OutsideDeposMin_6M (6000 < OutsideDeposMin_6M <= 10509) + 0,3428 * WOE_OutsideDeposMin_6M (16338 < OutsideDeposMin_6M <= 20304) + 0,4018 * WOE_OutsideDeposMin_6M (13241 < OutsideDeposMin_6M <= 16338) + 0,5382 * WOE_OutsideDeposMin_6M (OutsideDeposMin_6M > 37961) + 0,8784 *

* WOE_OutsideDeposMin_6M (20304 < OutsideDeposMin_6M <= 25561) + 0,9851 * WOE_OutsideDeposMin_6M (25561 < OutsideDeposMin_6M <= 37961) + 0,0367 * WOE_StandOrderCnt_6M (0 < StandOrderCnt_6M <= 6) + 0,2256 * WOE_StandOrderCnt_6M (12 < StandOrderCnt_6M <= 21) + 0,3184 * WOE_StandOrderCnt_6M (6 < StandOrderCnt_6M <= 12) + 0,7400 * WOE_StandOrderCnt_6M (21 < StandOrderCnt_6M <= 32) + 0,9218 * WOE_StandOrderCnt_6M (StandOrderCnt_6M > 32) + 0,0258 * WOE_EBWithdCnt_6M (EBWithdCnt_6M > 55) + 0,2819 * WOE_EBWithdCnt_6M (39 < EBWithdCnt_6M <= 55) + 0,3115 * WOE_EBWithdCnt_6M (0 < EBWithdCnt_6M <= 7) + 0,4254 * WOE_EBWithdCnt_6M (17 < EBWithdCnt_6M <= 27) + 0,5543 * WOE_EBWithdCnt_6M (27 < EBWithdCnt_6M <= 39) + 0,5801 * EBWithdCnt_6M (7 < EBWithdCnt_6M <= 17) + 0,3753 * WOE_ODDebtPast (ODDebtPast <= 0) + 0,1996 * WOE_OutsideDeposCnt_6M (8 < OutsideDeposCnt_6M <= 11) + 0,3347 * WOE_OutsideDeposCnt_6M (14 < OutsideDeposCnt_6M <= 18) + 0,3498 * WOE_OutsideDeposCnt_6M (11 < OutsideDeposCnt_6M <= 14) + 0,3832 * WOE_OutsideDeposCnt_6M (6 < OutsideDeposCnt_6M <= 8) + 0,3894 * WOE_OutsideDeposCnt_6M (18 < OutsideDeposCnt_6M <= 24) + 0,4129 * WOE_OutsideDeposCnt_6M (3 < OutsideDeposCnt_6M <= 6) + 0,4196 * WOE_OutsideDeposCnt_6M (OutsideDeposCnt_6M <= 3) + 0,2463 * WOE_SinceOpenLiveMax (204 < SinceOpenLiveMax <= 239) + 0,2693 * WOE_SinceOpenLiveMax (181 < SinceOpenLiveMax <= 204) + 0,3435 * WOE_SinceOpenLiveMax (38 < SinceOpenLiveMax <= 64) + 0,3721 * WOE_SinceOpenLiveMax (128 < SinceOpenLiveMax <= 156) + 0,3983 * WOE_SinceOpenLiveMax (SinceOpenLiveMax > 239) + 0,3984 * WOE_SinceOpenLiveMax (156 < SinceOpenLiveMax <= 181) + 0,4020 * WOE_SinceOpenLiveMax (99 < SinceOpenLiveMax <= 128) + 0,5303 * WOE_SinceOpenLiveMax (64 < SinceOpenLiveMax <= 99) + 0,1938 * WOE_PensionCnt_12M (PensionCnt_12M > 0) + 0,0157 * WOE_ODUtilSumPast4Q (0 < ODUtilSumPast4Q <= 8421,28) + 0,0225 * WOE_ODUtilSumPast4Q (8421,28 < ODUtilSumPast4Q <= 29793,6) + 0,0758 * WOE_ODUtilSumPast4Q (ODUtilSumPast4Q > 105614) + 0,2822 * WOE_ODUtilSumPast4Q (57672,1 < ODUtilSumPast4Q <= 105614) + 0,4344 * WOE_ODUtilSumPast4Q (ODUtilSumPast4Q <= 0) +

+ 0,3620 * WOE_CardTotalCntAvg_6M (3,17 < CardTotalCntAvg_6M <= 5) + 0,4219 * WOE_CardTotalCntAvg_6M (CardTotalCntAvg_6M > 21,5) + 0,4375 * WOE_CardTotalCntAvg_6M (0,17 < CardTotalCntAvg_6M <= 1,67) + 0,4769 * WOE_CardTotalCntAvg_6M (14,67 < CardTotalCntAvg_6M <= 21,5) + 0,5261 * WOE_CardTotalCntAvg_6M (5 < CardTotalCntAvg_6M <= 7,33) + 0,5317 * WOE_CardTotalCntAvg_6M (1,67 < CardTotalCntAvg_6M <= 3,17) + 0,6328 * WOE_CardTotalCntAvg_6M (10,33 < CardTotalCntAvg_6M <= 14,67) + 0,7025 * WOE_CardTotalCntAvg_6M (7,33 < CardTotalCntAvg_6M <= 10,33) + 0,1788 * WOE_OutsideDeposFlg_6M (2 < OutsideDeposFlg_6M <= 5) + 0,6128 * WOE_OutsideDeposFlg_6M (OutsideDeposFlg_6M > 5)**]**

6 Diskuze a výsledky

Po sestavení všech modelů a porovnání výsledků lze dospět k těmto závěrům:

6.1 Nejlepší typy modelu

V kapitole „Vlastní práce“ bylo testováno 12 variant modelů s různou kombinací vstupů a metod logistické regrese. Z výsledků Cross-Validace a pomocí změření síly modelu s využitím ROC křivek je patrné, že nejlépe predikoval výsledky MODEL10. Ten byl vyhotoven v kombinaci Long list of variables, WOE kategoriálních proměnných a logistická regrese byla provedena postupem Backward Stepwise. Celkem vyšlo 13 signifikantních proměnných na aplikačním datasetu, které nejlépe predikují pravděpodobnost splacení. Nejsilnější proměnnou aplikačních dat byla spojitá proměnná *Income (příjem),* která dosáhla hodnoty Information Value 0,2439. Model při cut-off hladině 0,87 predikoval správně výsledky s 87,95 % přesností s hodnotu AUC 0,7746.

Další model, který byl sestaven, byl model kreditní. Na základě výsledků testu Cross-Validace se podařilo sestavit model, který velice dobře predikuje výsledky, a to s přesností 91,97 %. Zde bylo nalezeno 12 signifikantních proměnných. Proměnná s nejvyšší hodnotou Information Value byla *Product_Type,* která nabyla hodnoty 0,4114. Hodnota AUC byla 0,8147.

Poslední model byl model behaviorální, který byl z výsledků ROC křivek prokázán jako nejsilnější. Hodnota AUC vyšla 0,8879. Behaviorální model predikoval výsledky s 89,03 % přesností. V behaviorálním modelu vyšlo 15 signifikantních proměnných. Všechny proměnné všech tří modelů jsou uvedeny v příloze D.

6.2 Výsledky modelů - Praktická zjištění a návrhy

Na základě výše uvedených výsledků a výsledků testování byly zjištěny tyto výstupy a doporučeny návrhy:

6.2.1 Porovnání hodnot Information Value

Pokud by se měly porovnat hodnoty Information Value nejsilnějších proměnných ze všech tří modelů, byly by to u aplikačního skóre proměnná *Income* s hodnotou 0,2439, u kreditního datasetu proměnná *Product_Type,* kde hodnota Information Value vyšla 0,4114 a u behaviráoních dat proměnná *Product_Type* s hodnotou 1,5999. Na základě výsledků všech ostatních hodnot vychází síla behaviorálních proměnných jako nejvyšší.

6.2.2 Power Indexy

Vzhledem k výsledkům přesnosti predikce bylo prokázáno, že přihlédnutí k hodnotám indexů Gini, Kolmogorov-Smirnov a Somer's D je pouze informativní.

6.2.3 Přístup k proměnným – WOE vs. nerozbinovaná proměnná

V kapitole 5.2.7 je názorná ukázka využití metody WOE, kde bylo prokázáno, že tato metoda je velice efektivní pro získání lepších informací a možnosti učinit přesnější rozhodnutí na základě dostupných informací. Výsledky rozbinovaných proměnných enormně zvednou sílu predikce.

6.2.4 Správné nastavení hladiny cut-off

Bylo zjištěno, že čím vyšší cut-off hladina byla nastavena, tím lépe model detekoval špatné chování klientů. Zároveň se ale zvyšovala chybovost detekce správně splácejících klientů. Banka při nastavení vyšší hladiny cut-off přicházela o dobré klienty. Jestliže nesprávně detekuje špatné chování klientů, zvýší se počet nesplacených úvěrů. Záleží tedy na strategii banky, kterou variantu by nakonec zvolila, vzhledem ke správnému nastavení vymáhání úvěrů nebo nastavení individuálního schvalování zamítnutých žádostí

6.2.5 Multikolinearita

Při práci s datasetem, který obsahuje cca 300 proměnných a více, je lepším přístupem provedení testu multikolinearity ještě před samotným binováním proměnných. Sníží se tím seznam proměnných a s datasetem se dále lépe pracuje.

6.2.6 Výběr výsledného listu proměnných

V modelu vyšel nejlépe Long list of variables, který vycházel pouze z hodnot Information Value. V praxi by banka přihlédla ještě k otázce jednoduchosti manipulace s listem proměnných. Vzhledem ke stejným výsledkům predikce a téměř totožným výsledkům síly ROC křivky, které se lišily o tisíciny, by banka přihlédla k faktu, že manipulace s velkým počtem proměnných je pracnější, proto by volila v tomto případě Short list of variables.

6.2.7 Výběr výsledného listu proměnných – Cross-Validační chyba

Pokud model zahrne všechny signifikantní proměnné, dochází k tzv. overfitted modelu (přefitování). Na první pohled se může zdát, že čím více proměnných bude model obsahovat, tím lépe bude predikovat. Přefitovaný model se pozná dle cross-validační chyby, kde už nedokáže na základě množství proměnných učinit správné rozhodnutí.

6.2.8 Interakce

V této práci s interakcemi pracováno nebylo a banky s nimi také většinou nepracují. V praxi můžou přinášet velice pozitivní výsledky, co se odhadu nesplácení týče. Jsou to různé kombinace binů proměnných, které se chovají v kombinaci jinak, než separátní proměnné. Banky s interakcemi většinou nepracují z důvodu náročné interpretace a složitého sledování velkého počtu kombinací binů.

6.3 Získávání dat – praktická zjištění a návrhy

Co se týká získávání dat bankou, měla by, dle zjištěných informací, postupovat takto:

6.3.1 Aplikační data vs. behaviorální data

Pokud by měla banka porovnat informace, které získá z behaviorálních dat a aplikačních dat, vyšly by jednoznačně prediktivnější a silnější data behaviorální. Proto, pokud nebankovní klient požádá o úvěr u jiné banky, než je jeho domovská, předkládá bance výpisy z účtu. Banka by s nimi měla umět následně pracovat. Možností by bylo elektronické čtení výpisů z účtu z jiné banky, který klient donese jako dokumentaci k úvěru. Pokud by se bance podařilo takto nasimulovat data, získala by lepší výsledky predikce než pouze u aplikačních dat a měla by možnost se lépe u klientů rozhodnout.

6.3.2 Zatajené informace

Dalším důvodem pro sběr dat touto formou může být zatajení některých informací ze strany klienta. Pokud by například hrál hazardní hry nebo sázel, banky se na takový fakt koukají velice negativně. Často z tohoto důvodu klienta rovnou zamítnou. To samé platí pro fakt, že klient může zatajit počet dětí, které má, popř. zatajit úplně, že děti má. Z občanského průkazu už není patrné, zda klient děti má. Každá osoba bez příjmu navyšuje existenční výdaje a snižuje potencionální výši limitu úvěru. Z účtu se dá ale vyčíst, jestli platí výdaje za stravné, školu nebo například alimenty. Pokud by klient informace zatajil, úvěr banka zamítne pro možný pokus o úvěrový podvod.

6.3.3 Další zdroje informací

Dále by banky měly zvážit získávání informací i z jiných dostupných zdrojů, jako jsou například sociální sítě. Jednalo by se ale zatím spíš o individuální schvalování, kde by se potvrdily informace sdělené klientem.

6.4 Slabiny úvěrového procesu a jejich možná vylepšení a praktické návrhy

Celá práce se také zabývá slabinami úvěrového procesu jak z pohledu banky, tak z pohledu klienta.

6.4.1 Doporučení pro klienta

Velkou hrozbou pro klienty je počet otevřených žádostí. Často, když se klient rozhodne zažádat o úvěr, obejde hned několik bank, kde podepíše žádost o úvěr.

Z výsledků je patrné, že proměnné *NumOfRejProd_off_us* a *NumOfProd_2y* (*počet zamítnutých žádostí a počet žádostí celkem*), jsou signifikantní a ovlivňují finální stanovisko banky. Proto si klient s každou další žádostí o úvěr škodí a snižuje tím svůj rating. Proto není dobré obcházení více bank a otvírání velkého množství žádostí.

Další hrozbou na stejné téma jsou tzv. zprostředkovatelé (makléři), kteří pro získání nejnižšího úroku obcházejí s klientem různé společnosti a snaží se tlačit cenu úvěru co nejníž. Jednak tímto chováním ubližují samotnému klientovi z důvodu počtu otevřených žádostí, ale také banky nutí klesnout s výší úroku na minimum, které pak neodráží skutečnou rizikovost klienta. Banky proto musí být přísnější, co se rozhodnutí o poskytnutí úvěru týče. Vzhledem k tomu, že u kreditního modelu vyšly signifikantní hned dvě proměnné, které v sobě zahrnují informaci ohledně žádostí o úvěr, tak takové jednání klientům ublíží. Pokud makléř takto projde například deset bank, klienta tím může velice poškodit tak, že se banky nakonec rozhodnou, vzhledem k velkému počtu otevřených žádostí, další žádost zamítnout. Každá zamítnutá žádost snižuje klientovi pravděpodobnost splacení a to významně. Takové makléřské jednání není proklientské, což si samotný klient neuvědomuje.

Dále z celého postupu vyplývá, že pokud je klient v kategorii New To Market, půjčí mu banka většinou nižší částky úvěru (v řádech desítek tisíc). Klient by tomuto mohl předejít tak, že by si otevřel kreditní kartu nebo kontokorent s nižším limitem. Na základě toho by sbíral kladnou historii v databázi CBCB. Banka by mohla při žádosti o úvěr přihlédnout k dobrému úvěrovému chování a půjčit klientům vyšší limit úvěru. Takový dotaz do externí databáze může pozitivně ovlivnit klientův profil z pohledu důvěryhodnosti a platební morálky.

Klienti by také měli mít po ruce záložní zdroj financí, pokud by se dostali do finančních potíží. Proto je otevření kreditní karty nebo kontokorentu, který budou mít klienti po ruce, dobrou variantou. Pokud by řešili úvěr ve finanční tísni, nemusela by banka žádost posoudit kladně.

6.4.2 Doporučení pro banku

U klientů NTB by se z uvedených výsledků mohlo zdát, že má takový klient vyšší pravděpodobnost na schválení úvěru, než klient bankovní. Proto, pokud by se klient nechoval dobře na účtech u své banky, měl by vyšší pravděpodobnost získání úvěru v bance jiné. I z tohoto důvodu by banky měly umět elektronicky přečíst informace z výpisů z účtu, které klient přinese k dokumentaci.

Další slabinou bylo zmíněno nahrávání dat bankami do externích registrů. Vzhledem k frekvenci nahrávání dat, která je jednou měsíčně, banky mohou dostávat již stará data. Například již zmíněná proměnná: počet otevřených žádostí, se může u klienta změnit ze dne na den z 0 například na 5. Toto časové zpoždění může negativně ovlivnit výsledné rozhodnutí o poskytnutí úvěru.

6.5 Ideální klient pro banku

Z dostupných testování ideální klient pro banku vypadá takto:

Na základě **aplikačních dat**:

Jednalo by se typicky o klienta ve věku od 29 do 34 let, s vlastním bydlením, vysokoškolským vzděláním, v manželství a se společným jměním manželů, ve vedoucí pracovní pozici v oboru financí, s pracovním poměrem trvajícím více, než 180 měsíců, který si platí pojištění vyšší částky, než 500Kč, s trvalým pobytem např. v Královehradeckém kraji kde zároveň i bydlí, s více než jednou osobou v domácnosti bez příjmu, který žádá o konsolidaci nebo o úvěr na bydlení.

Na základě **kreditních dat**:

Jednalo by se typicky o klienta, který měl v minulosti více, jak 4 správně splacené úvěry, z toho víc, jak 3 předčasně splacené, měl průměrnou výši kontokorentů větší než 20 000 Kč, 4 nebo 5 existujících produktů, které správně splácí, alespoň jednu existující hypotéku, kterou správně splácí, 0 zamítnutých žádostí, 0 otevřených žádostí, alespoň jeden uzavřený produkt za poslední dva roky, součet poskytnutých úvěrů za poslední 3 měsíce roven nule, nejlépe nula otevřených produktů v dané společnosti, kde žádá o úvěr, a který žádá o konsolidaci nebo o úvěr na bydlení.

Na základě **dat behaviorálních**:

Jednalo by se typicky o klienta, který má na účtu zůstatky za poslední rok vyšší, než 32 195 Kč, za poslední rok nebyl na účtu v dluhu, zůstatek na účtu v aktuálním měsíci má vyšší, než 171 943 Kč, minimální měsíční součet částek všech vkladových transakcí od ostatních než běžných účtů má vyšší, než 37 961 Kč za posledních 6 měsíců. Dále počet trvalých příkazů za posledních 6 měsíců má více, jak 32 kusů, minulý měsíc nebyl v debetu, maximální počet měsíců od skutečného otevření účtu by byl v intervalu od 64 do 99 měsíců, počet plateb na penzijní fond za poslední 1 rok měl vyšší, než jednu, skutečné čerpání úvěru nebylo provedenou ani jednou, měl průměrný počet plateb kartou za 1 měsíc za posledních 6 měsíců mezi 8 a 10 a počet transakcí inkasa za posledních 6 měsíců měl od 7 do 13.

6.6 Praktická aplikace modelu na náhodně vybraných klientech

Pokud bude brán v úvahu klient **aplikační modelu**, který bude vypadat následovně:

Věk 27, svobodný, s maturitou příjmem 20 424 Kč, který bude žádat o konsolidaci, bydlí u rodičů ve Středočeském kraji, kde se i narodil, pracuje jako duševně pracující zaměstnanec u soukromé společnosti, kde pracuje 81 měsíců a neplatí si pojištění.

U takového klienta vyjde pravděpodobnost splacení 88,82 %, takže by mu úvěr byl poskytnut. Ze známých výsledků je patrné, že tento klient úvěr opravdu splatil, tudíž model predikuje správně. Vzhledem k výši platu bude možné klientovi poskytnout půjčku s měsíční splátkou v maximální výši 12 254 Kč.

Pokud bude brán v úvahu klient **kreditního modelu**, který bude vypadat následovně:

V minulosti více, než 3 splacené produkty, z toho dva předčasně, kontokorent ve výši 30 000 Kč, který nečerpal poslední rok, vlastní hypotéku, kterou řádně splácí a zároveň má úvěr a kreditní kartu, kterou také nečerpá. V posledním roce nežádal o žádný úvěr a nemá žádnou zamítnutou žádost o úvěr. Před rokem a půl splatil řádně jeden úvěr a žádá o konsolidaci.

U takového klienta by model predikoval splacení na 99,52 %, takže by mu úvěr byl poskytnut. Ze známých výsledků je patrné, že tento klient úvěr opravdu splatil, tudíž model predikuje správně.

Pokud bude brán v úvahu klient **behaviorálního modelu**, který bude vypadat následovně:

Za poslední rok žádná platba na penzijní produkty, dvakrát v debetu na účtu, zůstatek na účtu v průměru za poslední rok 32 Kč, aktuálně má na účtu 48 Kč, nevkládá peníze na svůj účet, počet trvalých příkazů za posledních 6 měsíců je roven šesti, účet otevřený před dvěma lety, čerpání kreditní karty minimálně jednou měsíčně, počet transakcí inkasa za posledních 6 měsíců byl nula a kartou neplatí.

U takového klienta by model predikoval splacení na 17,91 %, takže by mu úvěr nebyl poskytnut. Ze známých výsledků je patrné, že tento klient úvěr opravdu nesplatil, tudíž model predikuje správně.

7 Závěr

Cílem této práce bylo sestavit na základě statistické analýzy credit scoringový model s nejlepšími prediktivními vlastnostmi. Dílčím cílem bylo nalezení nejlepšího modelu a nejlepšího přístupu k analýze. Na základě rozsáhlého testování byly sestaveny modely a vybrán takový, který nejlépe odhadoval chování klientů v budoucnu.

Z uvedených výsledků se u všech tří modelů osvědčil přístup v kombinaci s Long listem proměnných, WOE kategoriálních proměnných a logistické regrese pomocí metody Backward Stepwise. Dále na základě kalibrace pomocí hodnoty cut-off byly nastaveny hladiny pravděpodobností, kdy by byla banka ještě ochotná klientovi prostředky půjčit. U nastavování hladiny cut-off bylo bráno v potaz, kolika procentní úspěšnost měření lze dosáhnout za současné minimalizace rizika a ušlé obchodní příležitosti. U takto sestavených modelů se podařilo predikovat u aplikačního modelu 87,95 % správně identifikovaných klientů, u kreditního modelu 91,97 % správně odhadnutých klientů a u modelu behaviorálního 89,03 % správně identifikovaných klientů. Dále byla zkoumána síla jednotlivých modelů pomocí ROC křivek kde byly naměřeny hodnoty AUC u aplikačního skóre 0,7746, u kreditního skóre 0,8147 a u behaviorálního skóre 0,8879.

Vzhledem k rozsáhlé analýze byly sestaveny tři modely. Aplikační, kreditní a behaviorální. Na základě testování byly nalezeny signifikantní proměnné. Jednalo se o 13 proměnných z aplikačního datasetu, 12 proměnných z kreditního datasetu a 15 proměnných z behaviorálního datasetu, které jsou uvedeny v příloze D.

Tento seznam proměnných v každém modelu nejlépe predikoval budoucí chování klienta, proto byl jedním z hlavních výstupů této práce. Každý model byl zároveň vždy po sestavení validován pomocí Cross-Validace, jestli je použitelný v praxi a zda správně predikuje výsledky. Praktické využití modelů bylo potvrzeno na náhodně vybraných klientech.

Výsledkem celého modelu byly tři sestavené scorkarty, dle kterých bude banka rozhodovat o poskytnutí úvěrů. Všechny tři modely jsou použitelné v praxi a mají vysokou predikční schopnost. Z těchto výsledků dále plynuly slabiny celého úvěrového procesu a zároveň byla navržena doporučení jak pro klienta, tak i pro banku. Jak s těmito slabinami pracovat je uvedeno v diskuzi. Využití těchto navržených řešení má vysoký potenciál pro zvýšení konkurenční výhody, k úspěšnější prediktabilitě klientského chování za kratší rozhodovací čas a k maximalizaci zisku a minimalizaci ztrát.

V závěrečné části byla v rámci diskuze navržena sada zlepšení celého procesu. V neposlední řadě mají doporučení i lidský rozměr. Neposkytnutí úvěru nebonitním klientům banka předchází možnému vzniku krizových životních situací typu exekucí. Současný trend regulací České národní banky se také snaží těmto krizovým situacím předejít.

8 Seznam zdrojů a použité literatury:

Seznam použité literatury:

1) ABBOTT, Dean. *Applied Predictive Analytics.* USA: John Wiley & Sons, Inc., 2014. ISBN 978-1-118-72796-6

2) BESSIS, Joël. *Risk management in banking*. USA: John Wiley & Sons, Inc., 2015. ISBN 978-1-118-66021-8

3) CIPRA, Tomáš. *Riziko ve financích a pojišťovnictví: Basel III a Solvency II*. Praha: Ekopress, s. r. o., 2015. ISBN 978-80-87865-24-8.

4) CIPRA, Tomáš. *Praktický průvodce finanční a pojistnou matematikou*. Praha: Ekopress, s. r. o., 2015. ISBN 978-80-87865-18-7

5) ČECHURA, L., P. HÁLOVÁ, Z. MALÁ, M. MALÝ, J. PETEROVÁ a L. RUMÁNKOVÁ. *Cvičení z Ekonometrie.* Praha: Česká zemědělská univerzita v Praze, 2013. ISBN 978-80-213-2405-3

6) FINLAY, Steven. *Credit Scoring, Response Modelling and Insurance Rating.* USA: Palgrave Macmillan, 2010. ISBN 978-0-230-57704-6

7) HINDLS, R., S. HRONOVÁ, J. SEGER a J. FISCHER. *Statistika pro ekonomy.* Praha: Professional Publishing, 2007. ISBN 978-80-86946-43-6

8) KÁBA, Bohumil. a L. SVATOŠOVÁ, *Statistika.* Praha, Česká zemědělská univerzita v Praze, 2013. ISBN 978-80-213-0746-9.

9) MEJSTŘÍK, Michal, Magda PEČENÁ a Petr TEPLÝ. *Banking in Theory and Prectice.* Praha: Karolinum, 2014. ISBN 978-80-246-2870-7

10) PEČENÁ, Magda a P. TEPLÝ. *Credit Risk and Financial Crises.* Praha: Karolinum, 2010. ISBN 978-80-246-1872-2

11) SIDDIQI, Naeem. *Intelligent Credi Scoring.* New Jersey: John & Sons, Inc., 2017. ISBN 978-1-119-272915-0.

12) TUFFÉRY, Stéphane. *Data Mining and Statistics for Decision Making*. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

Seznam použitých internetových zdrojů:

13) *ČESKÝ STATISTICKÝ ÚŘAD,* 2018. *Mzdy 3. čtvrtletí 2018*. [online]. Praha, Česká republika. [cit: 2019-02-22]. Dostupné z: https://www.czso.cz/csu/czso/cri/prumerne-mzdy-3-ctvrtleti-2018

14) *ČESKÝ STATISTICKÝ ÚŘAD,* 2014. *Rozvody dlouholetých manželství, průměrný věk při rozvodu v hl. m. Praze 2001-2010.* [online]. Praha, Česká republika. [cit: 2019-02-22]. Dostupné z: https://www.czso.cz/csu/czso/104007-11-n_2011-11__rozvody_dlouholetych_manzelstvi-_prumerny_vek_pri_rozvodu_v_hl__m__praze_2001_2010

15) *DEEPANSHU BHALLA,* 2018. *Weight of Evidence and Information Value explained*. [online]. [cit: 2019-02-07]. Dostupné z: https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

16) *CZECH BANKING CREDIT BUREAU,* 2019. *Bankovní registr klientských informací*. [online]. Praha, Česká republika. [cit: 2019-01-03]. Dostupné z: https://www.cbcb.cz/

17) *CZECH NON-BANKING CREDIT BUREAU,* 2019. *Nebankovní registr klientských informací*. [online]. Praha, Česká republika. [cit: 2019-01-03]. Dostupné z: https://www.cncb.cz/

18) *SOLUS*, 2019, *Sdružení na Ochranu Leasingu a Úvěrů Spotřebitelům* [online]. Praha, Česká republika. [cit: 2019-01-03]. Dostupné z: http://www.soulus.cz/

19) *HONGRI, Jia*. *WECLOUDDATA,* 2018. *Scorecard: Credit Scoring with Machine Learning* [online]. Toronto, Kanada. [cit: 2019-01-12]. Dostupné z: https://weclouddata.com/credit-scoring-with-machine-learning/

20) *ČESKO.* Zákon č. 257/2016 Sb. ze dne: 14. července 2016 o spotřebitelském úvěru. In: *Sbírká zákonů České republiky.* 2016, částka:100. Dostupné také z: https://www.zakonyprolidi.cz/cs/2016-257

21) *KALER, I.* 2017. *How to Add Value to Your Clusters*. [online]. USA. [cit: 2019-02-01]. Dostupné z: https://medium.com/square-corner-blog/so-you-have-some-clusters-now-what-abfd297a575b

22) *MĚŠĚC*, 2019, *DTI*. [online]. Praha, Česká republika [cit: 2019-03-03]. Dostupné z: https://www.mesec.cz/slovnicek/dti/ . ISSN 1213-4414

23) *BANKY.CZ*, 2019, *DSTI*. [online]. Praha, Česká republika [cit: 2019-03-03]. Dostupné z: https://www.banky.cz/hypotecni-slovnik/dsti/ . ISSN 2464-4579

24) *HYPOTEČNÍ BANKA, a.s*, 2019. *Zodpovědné úvěrování*. [online]. Praha, Česká republika [cit: 2019-03-12]. Dostupné z: https://www.hypotecnibanka.cz/o-bance/odpovedne-financovani1/principy-zodpovedneho-financovani/

25) *THE MARKET SEGMENTATION STUDY GUIDE*, 2019. *Market segmentation example for banking*, 2019. [online]. [cit: 2019-01-12]. Dostupné z: https://www.segmentationstudyguide.com/understanding-market-segmentation/market-segmentation-examples/market-segmentation-example-banking/

26) *WHAT DO YOU MEAN BY NTB IN BANKING TERMS?*, 2018. [online]. [cit: 2019-01-12]. Dostupné z: https://www.quora.com/What-do-you-mean-by-ETB-and-NTB-in-banking-term

27) *ČESKÁ SPOŘITELNA*, 2019. *Půjčky.* [online]. Praha, Česká republika [cit: 2019-01-12]. Dostupné z: https://www.csas.cz/cs/osobni-finance/pujcky/pujcka

28) *WIKIPEDIA: the free encyclopedia*, 2019 [online]. *Úvěr.* St. Petersburg, Florida. [cit: 2018-12-12]. Dostupné z: https://cs.wikipedia.org/wiki/%C3%9Av%C4%9Br

29) *BISKUP, ROMAN, 2019: Statistika. Regresní a korelační analýza Úvod do problému.* [online]. *Úvěr.* St. Petersburg, Florida. [cit: 2018-2-13]. Dostupné z: https://docplayer.cz/3373309-Statistika-regresni-a-korelacni-analyza-uvod-do-problemu-roman-biskup.html

Seznam příloh:

Příloha A: Výsledky χ^2 testů na případný výskyt multikolinearity

Příloha B: Výsledky Cross-Validace, CUT-OFF hladina 0,95, Aplikační data

Příloha C: Výsledky Cross-Validace, MODEL 10, různé hladiny cut-off , Aplikační data

Příloha D: Výsledné proměnné z modelu aplikačního, kreditního a behaviorálního

9 Přílohy:

Příloha A: Výsledky χ^2testů; Zdroj: vlastní zpracování

WOE_ProductType	Chi-square	p-value
WOE_Education	209,4268	0,000000
WOE_WorkingPosition	194,8856	0,000000
WOE_DepartmentEmpl	166,4765	0,000000
WOE_CountryCont	165,1513	0,000000
WOE_EmployerType	114,9786	0,000000
WOE_Title	82,8871	0,000000
WOE_RegionCont	54,0193	0,000000
WOE_DifferentAdress	42,4547	0,000000
WOE_RegionPerm	31,5357	0,000113
WOE_CommonEquity	23,7255	0,000091
WOE_MaritalStatus	20,0875	0,000480
WOE_HousingStatus	9,0076	0,060909

WOE_WorkingPosition	Chi-square	p-value
WOE_Education	1226,584	0,000000
WOE_DepartmentEmpl	591,547	0,000000
WOE_CountryCont	532,333	0,000000
WOE_EmployerType	313,628	0,000000
WOE_HousingStatus	305,461	0,000000
WOE_Title	280,677	0,000000
WOE_RegionCont	196,837	0,000000
WOE_ProdectType	194,886	0,000000
WOE_MaritalStatus	99,825	0,000000
WOE_CommonEquity	80,679	0,000000
WOE_RegionPerm	18,113	0,001173
WOE_DifferentAdress	9,519	0,008572

WOE_HousingStatus	Chi-square	p-value
WOE_MaritalStatus	1728,137	0,000000
WOE_CommonEquity	1531,622	0,000000
WOE_CountryCont	711,864	0,000000
WOE_WorkingPosition	305,461	0,000000
WOE_RegionCont	212,990	0,000000
WOE_DifferentAdress	168,839	0,000000
WOE_Title	62,207	0,000000
WOE_RegionPerm	42,591	0,000000
WOE_Education	22,366	0,000002
WOE_ProdectType	9,008	0,060909
WOE_EmployerType	5,315	0,150115
WOE_DepartmentEmpl	2,034	0,361597

WOE_Education	Chi-square	p-value
WOE_Title	2013,219	0,000000
WOE_WorkingPosition	1226,584	0,000000
WOE_DepartmentEmpl	313,471	0,000000
WOE_EmployerType	256,043	0,000000
WOE_ProdectType	209,427	0,000000
WOE_DifferentAdress	49,380	0,000000
WOE_RegionCont	35,268	0,000000
WOE_CountryCont	26,831	0,000000
WOE_HousingStatus	22,366	0,000002
WOE_RegionPerm	16,401	0,000275
WOE_MaritalStatus	9,261	0,002341
WOE_CommonEquity	5,018	0,025088

WOE_MaritalStatus	Chi-square	p-value
WOE_CommonEquity	9461,981	0,000000
WOE_HousingStatus	1728,137	0,000000
WOE_CountryCont	349,882	0,000000
WOE_RegionCont	115,537	0,000000
WOE_WorkingPosition	99,825	0,000000
WOE_DifferentAdress	81,894	0,000000
WOE_Title	39,558	0,000000
WOE_ProdectType	20,087	0,000480
WOE_RegionPerm	16,062	0,000325
WOE_EmployerType	10,224	0,016751
WOE_Education	9,261	0,002341
WOE_DepartmentEmpl	5,191	0,074591

WOE_CommonEquity	Chi-square	p-value
WOE_MaritalStatus	9461,981	0,000000
WOE_HousingStatus	1531,622	0,000000
WOE_CountryCont	297,575	0,000000
WOE_RegionCont	103,932	0,000000
WOE_DifferentAdress	85,134	0,000000
WOE_WorkingPosition	80,679	0,000000
WOE_Title	37,184	0,000000
WOE_ProdectType	23,726	0,000091
WOE_RegionPerm	14,063	0,000883
WOE_EmployerType	8,493	0,036845
WOE_Education	5,018	0,025088
WOE_DepartmentEmpl	3,008	0,222269

WOE_RegionCont	Chi-square	p-value
WOE_RegionPerm	5922,055	0,000000
WOE_CountryCont	3864,593	0,000000
WOE_HousingStatus	212,990	0,000000
WOE_WorkingPosition	196,837	0,000000
WOE_MaritalStatus	115,537	0,000000
WOE_CommonEquity	103,932	0,000000
WOE_DifferentAdress	61,286	0,000000
WOE_ProdectType	54,019	0,000000
WOE_Education	35,268	0,000000
WOE_EmployerType	9,365	0,024810
WOE_DepartmentEmpl	7,835	0,019890
WOE_Title	7,636	0,005723

WOE_RegionPerm	Chi-square	p-value
WOE_RegionCont	5922,055	0,000000
WOE_CountryCont	223,910	0,000000
WOE_EmployerType	56,131	0,000000
WOE_HousingStatus	42,591	0,000000
WOE_ProdectType	31,536	0,000113
WOE_WorkingPosition	18,113	0,001173
WOE_Education	16,401	0,000275
WOE_DifferentAdress	16,282	0,000291
WOE_MaritalStatus	16,062	0,000325
WOE_CommonEquity	14,063	0,000883
WOE_DepartmentEmpl	10,877	0,027982
WOE_Title	1,230	0,540530

WOE_DifferentAddress	Chi-square	p-value
WOE_HousingStatus	168,8386	0,000000
WOE_CommonEquity	85,1341	0,000000
WOE_MaritalStatus	81,8940	0,000000
WOE_CountryCont	72,7341	0,000000
WOE_RegionCont	61,2860	0,000000
WOE_Education	49,3803	0,000000
WOE_ProdectType	42,4547	0,000000
WOE_DepartmentEmpl	32,2245	0,000000
WOE_EmployerType	21,8837	0,000069
WOE_RegionPerm	16,2817	0,000291
WOE_WorkingPosition	9,5186	0,008572
WOE_Title	4,5102	0,033693

WOE_EmployerType	Chi-square	p-value
WOE_DepartmentEmpl	4316,085	0,000000
WOE_WorkingPosition	313,628	0,000000
WOE_Education	256,043	0,000000
WOE_ProdectType	114,979	0,000000
WOE_CountryCont	65,974	0,000000
WOE_Title	58,771	0,000000
WOE_RegionPerm	56,131	0,000000
WOE_DifferentAdress	21,884	0,000069
WOE_MaritalStatus	10,224	0,016751
WOE_RegionCont	9,365	0,024810
WOE_CommonEquity	8,493	0,036845
WOE_HousingStatus	5,315	0,150115

WOE_CountryCont	Chi-square	p-value
WOE_RegionCont	3864,593	0,000000
WOE_HousingStatus	711,864	0,000000
WOE_WorkingPosition	532,333	0,000000
WOE_MaritalStatus	349,882	0,000000
WOE_CommonEquity	297,575	0,000000
WOE_RegionPerm	223,910	0,000000
WOE_ProdectType	165,151	0,000000
WOE_DifferentAdress	72,734	0,000000
WOE_EmployerType	65,974	0,000000
WOE_DepartmentEmpl	29,192	0,000000
WOE_Education	26,831	0,000000
WOE_Title	16,028	0,000062

WOE_DepartmentEmpl	Chi-square	p-value
WOE_EmployerType	4316,085	0,000000
WOE_WorkingPosition	591,547	0,000000
WOE_Education	313,471	0,000000
WOE_ProdectType	166,477	0,000000
WOE_Title	123,454	0,000000
WOE_DifferentAdress	32,225	0,000000
WOE_CountryCont	29,192	0,000000
WOE_RegionPerm	10,877	0,027982
WOE_RegionCont	7,835	0,019890
WOE_MaritalStatus	5,191	0,074591
WOE_CommonEquity	3,008	0,222269
WOE_HousingStatus	2,034	0,361597

WOE_Title	Chi-square	p-value
WOE_Education	2013,219	0,000000
WOE_WorkingPosition	280,677	0,000000
WOE_DepartmentEmpl	123,454	0,000000
WOE_ProdectType	82,887	0,000000
WOE_HousingStatus	62,207	0,000000
WOE_EmployerType	58,771	0,000000
WOE_MaritalStatus	39,558	0,000000
WOE_CommonEquity	37,184	0,000000
WOE_CountryCont	16,028	0,000062
WOE_RegionCont	7,636	0,005723
WOE_DifferentAdress	4,510	0,033693
WOE_RegionPerm	1,230	0,540530

Příloha B: Výsledky Cross-Validace, cut-off 0,95, Aplikační data; Zdroj: vlastní zpracování

	Summary Frequency Table MODEL 01			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	2870	528	3398
Total Percent		78,95%	14,53%	93,48%
Count	1	129	108	237
Total Percent		3,55%	2,97%	6,52%
Count	All Grps	2999	636	3635
Total Percent		82,50%	17,50%	

	Summary Frequency Table MODEL 02			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	2870	528	3398
Total Percent		78,95%	14,53%	93,48%
Count	1	129	108	237
Total Percent		3,55%	2,97%	6,52%
Count	All Grps	2999	636	3635
Total Percent		82,50%	17,50%	

	Summary Frequency Table MODEL 05		
	Real	Predicted 0	Row Totals
Count	0	3398	3398
Total Percent		93,45%	93,48%
Count	1	237	237
Total Percent		6,52%	6,52%
Count	All Grps	3635	3635
Total Percent		100,00%	

	Summary Frequency Table MODEL 06		
	Real	Predicted 0	Row Totals
Count	0	2870	3398
Total Percent		78,95%	93,48%
Count	1	129	237
Total Percent		3,55%	6,52%
Count	All Grps	2999	3635
Total Percent		82,50%	

	Summary Frequency Table MODEL 07		
	Real	Predicted 0	Row Totals
Count	0	2870	3398
Total Percent		78,95%	93,48%
Count	1	129	237
Total Percent		3,55%	6,52%
Count	All Grps	2999	3635
Total Percent		82,50%	

	Summary Frequency Table MODEL 08		
	Real	Predicted 0	Row Totals
Count	0	2870	3398
Total Percent		78,95%	93,48%
Count	1	129	237
Total Percent		3,55%	6,52%
Count	All Grps	2999	3635
Total Percent		82,50%	

	Summary Frequency Table MODEL 11			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	2256	1142	3398
Total Percent		62,06%	31,42%	93,48%
Count	1	46	191	237
Total Percent		1,27%	5,25%	6,52%
Count	All Grps	2302	1333	3635
Total Percent		63,33%	36,67%	

	Summary Frequency Table MODEL 12			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	2256	1142	3398
Total Percent		62,06%	31,42%	93,48%
Count	1	46	191	237
Total Percent		1,27%	5,25%	6,52%
Count	All Grps	2302	1333	3635
Total Percent		63,33%	36,67%	

Příloha C: Výsledky Cross-Validace, MODEL 10, různé hladiny cut-off , Aplikační data; Zdroj: vlastní zpracování

	Summary Frequency Table MODEL 10; CUT-OFF 0,85			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	3208	190	3398
Total Percent		88,25%	5,23%	93,48%
Count	1	186	51	237
Total Percent		5,12%	1,40%	6,52%
Count	All Grps	3394	241	3635
Total Percent		93,37%	6,63%	

	Summary Frequency Table MODEL 10; CUT-OFF 0,86			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	3171	227	3398
Total Percent		87,24%	6,24%	93,48%
Count	1	176	61	237
Total Percent		4,84%	1,68%	6,52%
Count	All Grps	3347	288	3635
Total Percent		92,08%	7,92%	

	Summary Frequency Table MODEL 10; CUT-OFF 0,87			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	3121	277	3398
Total Percent		85,86%	7,62%	93,48%
Count	1	161	76	237
Total Percent		4,43%	2,09%	6,52%
Count	All Grps	3282	353	3635
Total Percent		90,29%	9,71%	

	Summary Frequency Table MODEL 10; CUT-OFF 0,88			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	3053	345	3398
Total Percent		83,99%	9,49%	93,48%
Count	1	144	93	237
Total Percent		3,96%	2,56%	6,52%
Count	All Grps	3197	438	3635
Total Percent		87,95%	12,05%	

	Summary Frequency Table MODEL 10; CUT-OFF 0,89			
	Real	Predicted 0	Predicted 1	Row Totals
Count	0	2991	407	3398
Total Percent		82,28%	11,20%	93,48%
Count	1	133	104	237
Total Percent		3,66%	2,86%	6,52%
Count	All Grps	3124	511	3635
Total Percent		85,94%	14,06%	

Příloha D: výsledné proměnné z modelu aplikačního, kreditního a behaviorálního

Aplikační model:

- WOE_ProdectType
- WOE_HousingStatus
- WOE_WorkingPosition
- WOE_Education
- WOE_MaritalStatus
- WOE_CommonEquity
- WOE_InsurancePay
- WOE_Age
- WOE_RegionPerm
- WOE_DifferentAdress
- WOE_EmployerType
- WOE_PersonWithNoInc
- WOE_CurrentEmpSince

Kreditní model:

- WOE_Product_Type
- WOE_NumOfClosProd
- WOE_NumOfPreMatProd
- WOE_Avg_KU

- WOE_NumOfExProd
- WOE_NumOfExHU
- WOE_NumOfRejProd_off_us
- WOE_NumOfClosCL_2y
- WOE_NumOfProd_2y
- WOE_NumOfCL_3M
- WOE_NumOfExProd_on_us
- WOE_NumOfProd_3M_off_us

Behaviorální model:

- WOE_BalMinPast4Q
- WOE_DebtMonthCntPast4Q
- WOE_BalAct
- WOE_ODDelqStatMaxPast4Q
- WOE_OutsideDeposMin_6M
- WOE_StandOrderCnt_6M
- WOE_EBWithdCnt_6M
- WOE_ODDebtPast
- WOE_OutsideDeposCnt_6M
- WOE_SinceOpenLiveMax
- WOE_PensionCnt_12M

- WOE_ODUtilSumPast4Q
- WOE_DirectDebitCnt_6M
- WOE_CardTotalCntAvg_6M
- WOE_OutsideDeposFlg_6M